Using Business Intelligence to Predict Student Behaviour

Richard Scott Wilson

A thesis submitted in partial fulfilment of the requirements of
Sheffield Hallam University
for the degree of Doctor of Philosophy

June 2013

ABSTRACT

In light of reduced Higher Education funding, increased student contributions and competition between institutions, finding ways to understand student progression and improve the student experience are integral to the student, institution and state (York and Longden 2008).

This research uses Business Intelligence, specifically Data Warehousing and Data Mining, to build models that can be used to predict student behaviour. These models relate to final award classification, progression onto postgraduate studies at Sheffield Hallam University and employment type post undergraduate degree completion. This work builds upon the recommendations of Burley (2007) where the Department of Computing, at Sheffield Hallam University, was used to prove the applicability of such techniques.

It is fair to state that the field of student progression has been well documented over the years. Numerous authors (Tinto 1993, Yorke 1999, McGivney 2003) have all developed strategies and intervention techniques to help aid student progression. The evolving field of Educational Data Mining has focused, in the main, upon student interactions with web-based learning environments (Romero and Ventura 2006). Few studies have tackled the subject of using Business Intelligence as a method of understanding student progression (Dekker *et al* 2009, Herzog 2006).

The data was collected from the universities information systems and through the process of Data Warehousing and Data Mining a number of predictive models were constructed. This resulted in the identification of some interesting rules and variables, such as course and ethnicity, which are also fundamental in the more traditional student progression literature, such as Yoke and Longden (2008).

Overall, this research has further proved the applicability of Data Mining in Higher Education. The major institutional findings that have been established are: added value students are more likely to take postgraduate studies at Sheffield Hallam University, and a student's ethnicity can influence progression onto postgraduate studies and obtaining a graduate job.

CONTENTS

PREFACE

This thesis is submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Doctor of Philosophy and outlines the process followed and the results obtained from carrying out research into using Business Intelligence to predict student behaviour.

The first three chapters outline the area under investigation and previous research that has been conducted into understanding student progression in Higher Education, from both a traditional and Educational Data Mining perspective. Chapter 4 provides an overview of Business Intelligence, Data Warehousing and Data Mining. The approach that was followed when carrying out the research is discussed in Chapter 5. Chapters 6 and 7 explore the process of understanding and mining the student data. The findings of the research are then presented in Chapter 8 and recommendations for future research are made in Chapter 9. The whole process is then reviewed in the reflective summary in Chapter 10. Finally, Chapter 11 reiterates the main findings and recommendations of the research.

Part of this work has been presented in the following conference paper:

> BURLEY, Keith M and WILSON, Richard S (2012), *Understanding Student Progression for Data Mining Analysis*, HEIR, Presented at the Fifth Annual Conference of the Higher Education Institutional Research Network for the United Kingdom and Ireland.

This sparked a healthy debate about the quality of data within Higher Education institutions.

ACKNOWLEDGEMENTS

I would like to dedicate this work to my wife Vanessa Grace Wilson, whose tireless patience, endurance and encouragement over the last few years has allowed me to continue when at times completion seemed so far away. I would also like to thank my little boy, Reuben John Harry Wilson, who has sat, so patiently, at the bottom of the stairs waiting for his daddy to come and play with him.

A special thank you also needs to be made to my parents, Ronald and Linda Wilson, and my grandparents, Harry and Mable Ingham, for continuing to support me on my journey through Higher Education. Without their constant love, support and encouragement I would not be where I am today.

My supervisors Dr Keith M Burley and Dr Paul Crowther who have been an inspiration to me. Without their encouragement, knowledge and research experience I doubt that I would have completed this work. Additionally, I would also like to thank Dr Teresa Brunsdon for her guidance through the Data Mining phase of this research.

I would also like to acknowledge the staff, at Sheffield Hallam University, who have enriched my understanding of the university and its processes. In particularly, I would like to thank Dr Helen Kay and Caroline Heaton as without their help I would not have been able to obtain the data for this research.

I would also like to thank SAS® for allowing me to use their software and providing financial support to help pay for materials, tuition fees and conferences.

Finally, since starting my Higher Education journey, in 1999, there have been some major changes in my life: marriage, the birth of my son and sadly the loss of my grandparents - Raymond and Alice Wilson. Therefore, this acknowledgement wouldn't be complete without mentioning them. I am sorry you weren't there at the end of my journey.

GLOSSARY

Added Value

Or value-added relates to “Student achievement, which is inextricably connected to institutional success, must be measured by institutions on a ‘value-added’ basis that takes into account students academic baseline when assessing their results.” (Spellings 2006, p04).

Adults with Higher Education Qualifications by Postcode (QAHE)

Is a measure, on a scale of 1 to 5, which forms part of the HEFCE work into POLAR2. It is used to rate the number of adults, in a region of the United Kingdom, who have obtained Higher Education qualifications (where 1 is low and 5 is high) (HEFCE 2012).

Business

Is concerned with any particular employment or occupation that is engaged in for gain, livelihood or a profession; can also refer to financial dealings such as the buying or selling of an item(s) (Richardson and Richardson 1992).

Business Dimensional Lifecycle Diagram (BDLD)

“A methodology for planning, designing, implementing, and maintaining data warehouses [...].” (Kimball and Ross 2002, p393).

Business Intelligence (BI)

“A generic term to describe leveraging the organization’s internal and external information assets for making better business decisions.” (Kimball and Ross 2002, p393).

Categorical Variable Consolidation

Using a decision tree to group the levels of a categorical exploratory variable based on its associations with target variable to create a new model input. (Georges *et al.* 2010).

Categorical Variables

“A variable whose values are not numerical. Examples include gender (male, female), paint colour (red, white, blue), [...].” (Upton and Cook 2002).

Classification

The process “[...] assigning a newly presented object to one of a set of predefined classes.” (Berry and Linoff 2011, p86).

Continuous Variables

“A variable whose set of possible values is a continuous interval of real numbers *x*, such that a < x > b, in which *a* can be –[infin] and *b* can be [infin].” (Upton and Cook 2002).

Cube

“Name for a dimensional structure on a multidimensional or online analytical processing [...] database platform, originally referring to the simple three-dimension case of product, market and time.” (Kimball and Ross 2002, p395).

Data Cleansing

Is the act of detecting and removing/correcting data in a database that is deemed to be dirty (English 1999).

Data Mart

In top-down Data Warehousing:

A Data Mart is “a collection of subject areas organized for decision support based on the needs of a given department”. In top-down Data Warehousing the Data Marts extract the data from the Enterprise Data Warehouse, they are dependent on the data stored within the Enterprise Data Warehouse (Inmon 1999).

In bottom-up Data Warehousing:

A Data Mart is “a flexible set of data, ideally based on the most atomic (granular) data possible to extract from an operational source, and presented in a symmetric (dimensional) model.” In bottom-up Data Warehousing the Data Marts are independent of the Data Warehouse, the Data Marts are consolidated to form the Data Warehouse (Kimball and Ross 2002, p396).

Data Mining (DM)

“[...I]s the automated analysis of large data sets to identify previously unknown patterns or trends of information in the data that may be used to make valid predictions. It uses standard statistical analysis and modelling techniques to discover patterns that typically would go undetected using ordinary statistical methods”. (Samli *et al.* 2002, p219)

Data Sparcity

A poorly designed multidimensional database (cube) can have a larger physical size then the information it retains, it is sparse. This results in a cube that is larger than necessary and can lead to problems with usability and performance (Kimball and Ross, 2002).

Data Warehouse

“The conglomeration of an organizations data warehouse staging and presentation areas, where operational data is specifically structured for query and analysis performance and ease-of use.” (Kimball and Ross 2002, p397).

Data Warehousing (DW)

“[...] what you need to do in order to create a data warehouse, and what you do with it. It is the process of creating, populating, and then querying a data warehouse and can involve a number of discrete technologies [...].”(Reed no date).

Educational Data Mining (EDM)

“Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using these methods to better understand students, and the settings which they learn in.” (Baker and Yacef no date, p02).

Enterprise Data Model

Defines all the data that is common to a business, from a high-level business view to a generic logical data design, including links to the physical data designs of individual applications (Singh 1998).

Enterprise Data Warehouse (EDW)

Is a centralised, normalised and atomic data store that is used to populate a number of dependent Data Marts. An Enterprise Data Warehouse is arranged around the enterprise subject areas found in the enterprise data model (Inmon 2002).

Entity-Relationship (ER) Modelling

“[...I]s a way of graphically representing the logical relationships of entities (or objects) in order to create a database. The ER model was first proposed by Peter Pin-Shan Chen of Massachusetts Institute of Technology (MIT) in the 1970s.” (Rouse 2005).

Epistemology

“[W]hat is (or should be) regarded as acceptable knowledge in a discipline.” (Bryman 2012, p27)

Estimation

Whilst “[c]lassicfaction deals with discrete outcomes: yes or no; [...]. Estimation deals with continuously valued outcomes. Given some input data, estimation comes up with a value for some unknown continuous variable such as income, order size, or credit card balance.” (Berry and Linoff 2011, p86).

Ethnomethodology

"[I]s a family of related approaches concerned with describing and portraying how people construct their own definitions of social situations or, more broadly, with the social construction of knowledge." (Schwandt 1997, p44).

Explanatory Variable(s)

Otherwise known as the predictor variable(s) or independent variable(s) and refers to the inputs or predictors of a model that is used to derive an equation or rules to predict an output variable - target variable (Two Crows no date)

Higher Education Funding Council for England (HEFCE)

"HEFCE distributes public money for higher education to universities and colleges in England, and ensures that this money is used to deliver the greatest benefit to students and the wider public." (HEFCE 2012a)

Higher Education Statistics Agency (HESA)

"[...I]s the official agency for the collection, analysis and dissemination of quantitative information about higher education." (HESA no datea)

Higher Education (HE)

"[... Is] a diverse range of courses and qualifications, such as first degrees, higher national diplomas and foundation degrees. Many courses take place in universities, but plenty are also taught at higher education colleges, specialist art institutions and agricultural colleges." (UCAS no date).

Joint Academic Coding System (JACS)

"[...] is owned and maintained by the Universities and Colleges Admissions Service (UCAS) and the Higher Education Statistics Agency (HESA) and is used for subject coding of provision across higher education in the UK. [...] JACS is currently used to code the subjects of both higher education courses and the individual modules within them across the full range of higher education provision." (HESA no date)

Key Information Set (KIS)

"[...A]re comparable sets of information about full or part time undergraduate courses and are designed to meet the information needs of prospective students." (HEFCE 2012b).

Layer

"Nodes in a [Neural Network] are usually grouped into layers, with each layer described as input, output or hidden. There are as many input nodes as there are input (independent) variables and as many output nodes as there are output (dependent) variables. Typically, there are one or two hidden layers." (Two Crows no date).

Lift

"The most common way to compare the performance of classification models is to use a ratio called lift. [...] What lift actually measures is the change in concentration of a particular class when the model is used to select a group from the general population" (Berry and Linoff 2004, p81).

Node

"A decision point in a [... Neural Network] that combines input from other nodes and produces an output through application of an activation function." (Two Crows no date).

Nominal Variable

A variable that has no numerical values, such as gender or marital status (Hand *et al.* 2001)

On-Line Analytical Processing (OLAP) System(s)

Is concerned with extracting the data in the Data Warehouse and presenting it to the users. A On-Line Analytical Processing System(s) can be thought of as the front-end to a Data Warehouse. Increasingly software vendors are using the term to refer to their front-end analytical software. (Kimball and Ross 2002).

On-Line Transactional Processing (OLTP) System(s)

Is concerned with loading an organisations day-to-day transactional data into a relational database, in this sense an On-Line Transactional Processing System(s) can be thought of as a front-end to a relational database (Kimball and Ross 2002).

Ontology

Research ontology is concerned with investigating the nature or essence of social phenomena (Bryman 2012).

Operational Data Store (ODS)

“A physical set of tables sitting between the operational systems and the data warehouse [...]. The main reason for the ODS is to provide immediate reporting of operational results if neither the operational system nor the regular data warehouse can provide satisfactory access. Because an ODS is necessarily an extract of the operational data, it also may play a role of source for the data warehouse” (Kimball and Ross 2002, p408).

Optimisation

“The minimization or maximization of some function, usually subject to restrictions (which are often on the values of the variables over which the optimization takes place).” (Upton and Cook 2002).

Ordinal Variable

A variable that has values that have a natural order, such as months of the year or education status (Hand *et al.* 2001).

Organisation

“[Is a group] of people who co-ordinate their activities in pursuit of a common purpose.” (Richardson and Richardson 1992, p03)

Oversampling

Is the increasing of the classification of a rare event, so that a higher proportion of the rare event exists in the population. (SAS 2013)

Participation of Young People in Higher Education by Postcode (QYPR)

Is a measure, on a scale of 1 to 5, which forms part of the HEFCE work into POLAR2. It is used to look at the number of young people who entered into Higher Education based on where they live in the UK (where 1 is low and 5 is high). (HEFCE 2012)

Phenomenology

“A philosophy that is concerned with the question of how individuals make sense of the world around them and how in particular the philosopher should bracket out preconceptions concerning his or her grasp of that world.” (Bryman 2012, p714).

Participation of Local Areas (POLAR2)

Is a classification of areas in the United Kingdom used to analyse the participation of young people in Higher Education based on where they live. (HEFCE 2012)

Prediction

“[...I]s the same as classification or estimation, except that the records are classified according to some predicted future behaviour or estimated future value.” (Berry and Linoff 2004, p10).

Quality Assurance Agency (QAA)

Carryout out assessments of institutions and “offer advice, guidance and support to help UK universities, colleges and other institutions provide the best possible student experience of higher education.” (The Quality Assurance Agency for Higher Education 2012)

Regression

Is a data analysis technique that is used to build predictive models. Regression is used to determine the expected value of the target variable from the actual values of related explanatory variables that tend towards a straight line (Wetherill 1986).

Relational Model

Is a way of storing and processing data in a Data Warehouse, in this model the data is stored in the form of a Relational Database Management system, this model is therefore similar to a transactional system (Schwatz 1996).

Research Territory Map

A high level conceptual map of the area under investigation, which helps to identify links between related topics and provides a way to classify and sort the research material obtained (Dawson 2000).

Sheffield Hallam University (SHU)

Is "[o]ne of the UK's most progressive and innovative universities, Sheffield Hallam is a multicultural institution with a vibrant and diverse student population [...]" (Sheffield Hallam University no date). Located in Sheffield, South Yorkshire, SHU is a former polytechnic that was given university status by the government in 1992 - all of the establishments granted this status are today collectively known as 'post 1992' universities.

Star Schema

“The generic representation of a dimensional model in a relational database in which a fact table with a composite key is joined to a number of dimension tables, each with a single primary key.” (Kimball and Ross 2002, p414).

Structured Query Language (SQL)

"First developed in the early 1970s at IBM by Raymond Boyce and Donald Chamberlin, SQL was commercially released by Relational Software Inc. (now known as Oracle Corporation) in 1979. [... SQL] is a standard computer language for relational database management and data manipulation. SQL is used to query, insert, update and modify data." (Janalta Interactive Inc. 2013).

Symbolic Interactionism

“A theoretical perspective in sociology and social psychology that views social interaction as taking place in terms of the meanings actors attach to action and things.” (Bryman 2012, p716).

Target Variable

Otherwise referred to as the outcome variable, dependent variable or response variable, is determined through the rules or equations of a model from a number of explanatory variable(s) (Two Crows no date).

Topology

Topology is used in relation to Neural Networks and therefore is defined as the number of layers and nodes in each layer of a Neural Network (Two Crows no date).

1 INTRODUCTION

> *"The importance of student success in higher education is incontestable, whether one's standpoint is that of a student, a programme team, a department, an institution, or a higher education system" (Yorke and Longden 2008, p04)*

This thesis seeks to introduce Business Intelligence (BI) tools and techniques to the problem of student progression in Higher Education (HE). It will attempt to create a number of intelligent user profiles of Sheffield Hallam University (SHU) undergraduate students, to answer the following research question: *How can Business Intelligence be used to predict student behaviour as an aid to improving student progression?* The research also builds upon the recommendations of Burley (2006). It is perhaps important to state that this research was conducted in the climate of reduced HEFCE (Higher Education Funding Council for England) funding and the consequential increase in student fees. However, it pre-dates the withdrawal of HEFCE funding since the data is taken from 2006.

The word *'student progression'* will be used throughout this document to refer to the maintenance of students on their *original course* and *final completion*. The word progression implies a more optimistic approach to tackling the problem. This is in sharp contrast to the word retention, which has a more managerial feel. Indeed, Yorke and Longden (2004) suggest that retention implies the measurement of efficiency and effectiveness of a system or institution. They go on to suggest that the rationale for retention and completion as an indicator of success is weak. It is perhaps important to define the word progression in the context of this study. Student progression is associated with much more than dealing with the academic issues. Indeed, it is about helping the student to overcome the issues associated with entering HE and dealing with the personal issues that they face as part of the process (Moxley *et al.* 2001).

The *student experience* is fundamental to improving student progression as it helps the student to overcome the personal issues that they face as part of the HE process. The importance of this is visible, at SHU, in the appointment of an Assistant Dean for Student Experience in the faculty of ACES (Art, Computing, Engineering and Sciences). The quality of the student experience is something that institutions in the UK (United Kingdom) have had a high reputation for delivering. However, a reduction in funding has led to unhappiness in the sector over worries of the decline in the student experience (Yorke and Longden 2008). Therefore, the term student experience will be associated throughout this research with quality and understanding how the institution has managed the expectations of its students.

Yorke and Longden (2008) suggest that student progression needs to be considered from three perspectives – the *state*, the *institution* and the *student*. Therefore, where possible this research will consider the interests of these three stakeholders in relation to student progression in HE. Through carrying out a comparative analysis of HE systems (in England, Australia, South Africa and the United States of America), they found that widening participation, increasing access and student funding were common reoccurring themes across all countries. In addition to this, they also found that within each of the countries there were differing rates of student progression and success in different groups and different institutions in HE.

In 2007, the National Audit Office found that 21,504 first year full time degree students, who enrolled in 2004-05, failed to progress into their second year. Whilst this is a slight improvement to 1999-2000, this still represents a significant financial loss to all of the HE stakeholders, introduced above. Indeed, Yorke and Longden (2008) estimate that cost of non-progression is £110 million per annum. All stakeholders have an active part to play in improving student progression and there is an assumption that the student wishes to progress (Burley 2006). However, Peelo *et al*. (2002) suggests that failure to progress should be accepted as a normal part of the learning process and as a result students should not be protected from failure.

2 Research Aims and Objectives

> *"General aims must then lead to a statement of specific aims, and these should be turned into operationalized aims; that is, a specified set of practical issues or hypotheses to be investigated."* (Oppenheim 1992, p07)

The intention of this chapter is to outline the research question, aim and objectives and provide a rationale as to the importance of the work.

2.1 Research Question and Aim

Burley (2006) carried out research into exploring the issues that affected the progression of computing students at SHU through the use of DM techniques. The main focus of the research was to test the value of DM in understanding student progression. The results of his research highlighted a number of recommendations, one of which was to extend the research to include all faculties at SHU. Therefore, the question that this research intends to answer is:

> *How can Business Intelligence be used to predict student behaviour as an aid to improving student progression?*

With this in mind the aim of the research is to:

> *Explore, through the application of BI tools, the issues that affect the progression of all undergraduate students at SHU. It is intended that a number of predictive models will also be constructed to predict student behaviour.*

2.2 Objectives of the Research

At this stage, it is important to breakdown the research aim into a number of manageable objectives and associated measures. These will be used to help plan and assess the success of the research. Further discussions around how these objectives will be achieved can be found in Chapter 5. The following six objectives were identified along with associated measures for success, as figure 2.1 below shows.

No.	Objective	Measure
1.	Review, compare and contrast existing knowledge to develop a theoretical framework on which to base the rest of the study.	Completed literature review.
2.	Develop knowledge of the relevant SHU information systems and DM software to form an understanding of the underlying data structures and mining software.	Understanding of the student data and SAS® software through speaking to experts.
3.	Explore existing data sets, inductively, to build inferences and determine patterns in the data.	Reduced variables in the data set and the introduction of new variables through the iterative use of DM.
4.	Apply suitable DM techniques to build a number of predictive models.	Final models built and assessed.
5.	Validate the findings of study by comparing the results of the quantitative analysis to the current body of knowledge.	Completed findings.
6.	Compile a list of recommendations for the future uses of DM in this area based on the findings of the study.	Completed list of recommendations.

Figure 2.1 - Research Objectives and Measures.

2.3 Research Rationale

There has been increased pressure placed upon institutions to widen participation and increase access to HE (Yorke and Longden 2008). It is perhaps important to point out that this landscape is now changing with the reduced number of university places post 2011. However, the widening participation agenda (defined in section 3.2.1) is pertinent to the time (2006) from which the data used in this research was taken. In the current economic climate the number of university places available, financing and employability are major concerns for the institution, student and government. Arguably, in providing a rationale for the research it is necessary to discuss the importance of the research, who will be interested in the work, how the research will inform institutional policies and the tools used for data collection and analysis.

2.3.1 The Importance of the Research

This research is important for a number of reasons, the principle ones being educational and financial. Indeed, Yorke (1999, p01) attests that:

> *"Governments around the world are increasingly calling higher education to account for the money that is invested in institutions, as is evidence by the rise of national quality assurance systems during the 1990s and the interest shown in performance indicators of various kinds. [...] There is a general international perception that economies are best served by maximizing the level of education in the populace."*

Over the last ten to fifteen years, there have been a number of significant changes in the way that HE has been funded in England. Indeed, the burden of financing education has moved from the local education authorities to the student. Prior to the mid-1990s, HE was funded by local education authorities. However, during the mid-1990s student loans were introduced, which paved the way for the introduction of a student contribution to tuition fees in 1998. A further top-up fee was later introduced in 2006 but this was at the discretion of the individual institution. Today the top-up fee has now become mandatory across HE institutions, which has reduced the dependency on the local authority but increased the burden on the student - the average student debt is expected to rise to over £20,000 (Garner 2008). The student contribution is collected after the student graduates and is earning past a certain threshold. This has provided non-traditional students with access to a university education. However, if the student fails to obtain a graduate salary, through non-progression or poor employment opportunities, then they will find it harder to repay the debt.

2.3.2 Educational Importance

It is well documented that there is a high risk of first year undergraduate students failing to progress, the reasons for this can be grouped into *educational* and *behavioural issues* (Yorke 1999).

One of the biggest educational issues is that the student is moving from a relatively protected environment of school or college, where they are encouraged, monitored and guided to complete their work, to the much more relaxed environment of academia (Burley 2006). Furthermore, some students will also be moving away from the protected environment of home to study. However, the number of students moving away to study appears to have decreased in recent years. Indeed, faced with increasing debts, it is believed that more students are electing to attend local universities and stay at home (Coughlan 2009). Moxley *et al.* (2001) suggest that the majority of problems affecting the student are outside the educational process.

> *"Most institutions recognise that undergraduate education is much more than formal instruction and encompasses opportunities to develop socially, culturally, physically, spiritually and ethically."* (Moxley *et al.* 2001, p58)

However, there are few mechanisms in place for identifying those students who fail to engage with the university. Indeed, such students are rarely detected until several weeks into the first semester, by this time the student has developed certain behavioural issues that are very difficult to change. From an educational perspective it would be beneficial to both the student and the institution if these students could be identified through predictive modelling. This could help the institution to facilitate the student's progression into academia, through for example offering help in terms of bridging courses, group events and the like. This has the potential to improve both student progression and the student experience.

In 1997 the Labour government committed itself to increasing student participation to 50% by 2010 (Slack and Casey 2002). The number of applicants accepted, through UCAS (Universities and Colleges Admissions Service), in the UK rose from 332,000 in 2002-03 to 346,000 in 2006-07, this shows an increase in accepted places of 4.1%. During this time, there have been changes in the subjects that students have applied to study. Indeed, applications for subjects allied to medicine increased whilst there were large reductions in applications for computer science, mathematics and engineering courses. The government has also incentivised institutions to widen participation. This has resulted in an increase in the proportion of students entering HE from black and minority ethnic groups, disabled students, and students with non-traditional HE backgrounds (National Audit Office 2007).

However, as pointed out by Yorke (1999) the risk of increasing and widening participation is the potential decrease in student progression. In 2007, the National Audit Office undertook research into student retention in HE. They found that 8.4% of first year full time degree students, who enrolled in 2004-05, failed to progress into their second year (National Audit Office 2007, p05). Indeed, this study highlights that the "retention of full-time, first degree students has improved slightly since 1999-2000" (National Audit Office 2007, p05). Nationally this could be looked upon as a failure to educate the population to its full potential. At the institutional level this will affect progression rates, funding, future availability of places and ranking within published league tables. From the student perspective the issues are psychological and the incurring of debts.

2.3.3 Financial Importance

At this stage, it is perhaps prudent to discuss the assertion that, given the level of student contribution, some may argue that the student is a customer of an institution. However, students are not merely purchasing a degree, they are purchasing access to a product that they have to successfully interface with to achieve their aim – the opportunity to participate. As a result the student will still incur debts from student loans and bank overdrafts, regardless of whether the student progresses to subsequent years of study or completion, which will need to be repaid. Arguably, students attaining a graduate level salary are much more likely to pay off these debts quicker as non-graduate earning potential is significantly lower – on average graduates earn over £100,000 more than non-graduates over there lifetime (National Audit Office 2007).

Progression is also important for the institution as they only receive funding from the HEFCE and\or the student for the number of completed years. This has a detrimental effect upon university rankings and the amount of future funding received, as the HEFCE will reduce its funding for the proceeding years. The university budget will also be impacted upon as there is a large effort, in terms of cost, associated with marketing and recruitment of students (Yorke and Longden 2008).

It is perhaps important to clarify that at the time that the data, used in this research, was recorded the institution received around £5,000 per student per year from HEFCE. On top of this the institution also received £3,500 from the student. However if the student failed to progress or failed to engage with the course (fails to submit any work for any module) then the institution was fined the following year (HEFCE 2009). In 2011, the HEFCE withdraw funding for undergraduate courses and most institutions now charge £9,000 per student per annum.

In turn, poor progression rates also reflect badly upon the government as the rates are often published in the media. Yorke (1999, p02) highlights that:

> *"Non-completion and delayed completion rates can be constructed as inefficiencies in the use of public finances, and hence they become political issues."*

Indeed, Yorke and Longden (2008) estimate that non-completion costs the state around £110 million per annum.

2.3.4 Interested Parties

The management of SHU will be interested in the results of this research. It is also expected that the models developed, as part of this research, will generate interest and debate, from institutions and subject matter experts, regarding the use of modelling to predict student behaviour. Vendors of BI applications may also be interested in the results of this research, as HE is a relatively untapped market (Luan 2002). The models generated as part of the research could also be of use to university admissions and marketing staff, tutors and student support staff. Indeed, an understanding of student behaviour is fundamental in helping those who have direct contact with the students (Moxley *et al.* 2001).

2.3.5 Informing Institutional Policies

Given that the outcomes of this research will help to foster an improved understanding of student opinion and behaviour, admissions and marketing staff, tutors and student support staff may find the results interesting so that selection of students and intervention can be improved. This will also be useful to students as identifying that they may require intervention, may provide them with the skills to complete their degree and earn a graduate salary. Furthermore, every effort will be made to try to attract SHU staff into using the models.

Whilst the models will be built at SHU, it is expected that they could be of use to other post 1992 universities. The research will also add a further dimension to the HE knowledge domain and has the potential to cause a debate as to the future use of modelling of student behaviour.

The application of BI tools and techniques in the field of HE is relatively new. There has been little research into the problem of student progression using BI and few have tried to model student behaviour. Arguably, the results of this research will contribute to the body of knowledge that already exists through publications in journals. This research will add to the current knowledge, stimulating a healthy debate amongst the subject matter experts and at least be of use to students and institutional staff.

2.4 Summary

This chapter provides an overview of the research question aim and objectives along with a rationale as to why the work is important. The research rationale is broken down into five key areas. These are the importance of the work from an educational and financial perspective, who will be interested in the research and how the research will inform institutional policies. The chapter highlights that an improved understanding of student progression could have a positive impact upon all of the stakeholders who have a vested interest in HE success.

3 Previous Research

> *"Around 28,000 full-time and 87,000 part-time students who started a first-degree course in 2004-05 were no longer in higher education a year later." (Parliamentary 2008).*

The subject of student progression in HE is a high profile issue, this is reflected by the fact that there has been a significant amount of research carried out within this area (Yorke and Longden 2008). Since 2005, the subject of the application of BI in HE has grown substantially. Indeed, the growth of interest in the area of Educational Data Mining (EDM) can be seen by the recent creation of the International Conference for EDM and the increased publications of journals and books in this area (Romero *et al.* 2011).

This review will initially focus upon the findings of Burley (2007) which will then be expanded to include, in the first instance, research carried out by Yorke (1999), Moxley *et al.* (2001), McGivney (2003), and Yorke and Longden (2008). This will then be extended to include literature from the evolving field of EDM and its application in the problem of student progression.

3.1 Data Mining Techniques in Higher Education Research

In this section the findings of Burley (2007), titled "*Data Mining Techniques in Higher Education Research - The Example of Student Retention*", will be reviewed. The section will provide a brief overview of the study, including any relevant recommendations, and highlight the main strengths and weaknesses of the work.

Burley's research seeks:

> *"[...] to explore interrelationships between factors that contribute to student attrition and hence establish the demographics of at risk students" (Burley 2007, p01)*

Burley's research is concerned with establishing the issues associated with student progression to demonstrate the suitability of DM in the field of HE. This is a reasonably unique approach of looking at the problem as very few authors have tried to understand the problem in this way – see section 3.3. In the main, his research focuses upon students belonging to the Department of Computing at SHU. Through an extensive review of the literature, Burley identifies three key themes (Casual Problems, Modelling and Intervention). These are drawn from the work of McGiveny (1996), Yorke (1999) and Moxley *et al.* (2001). These are then considered from both the institutional and individual (student) perspective.

DM is then introduced to the problem of student progression, with an extensive discussion around supervised and unsupervised DM techniques. The research takes a mixed methods approach to help understand the issues. Having gained an understanding of the problems, Burley carries out a number of exploratory face-to-face interviews to gain further insights into problems pertinent to SHU computing students. These insights are then used to develop an online questionnaire that is targeted, in the main, at students within the Department of Computing at SHU (15.5% of respondents were from other similar universities).

In his evaluation of the process, Burley notes that there were two problems with the design of his questionnaire. These related to pigeon holing students into predefined age groups and rating responses on a five point likert scale, all of which hindered the analysis process. Burley collects his data over a nine month period, which is then categorised as student demographics and response to attitude issue. Given the period of data collection, it is questionable whether there were enough responses (587) to provide a representative sample of the population as DM is more effective with larger data sets. According to Berry and Linoff (2011, p167)

> *"Data Mining is most useful when sheer volume of data obscures patterns that might be detectable in smaller databases [...] We generally start with tens of thousands if not millions of pre-classified records so that the model set contains many thousands of records."*

However, the research remains a good example of how effective DM could be in understanding the problem as the result corresponded to previous findings, such as York (1999).

After preparing the data, the results are then mined using a combination of both unsupervised (Clustering and Rule Association) and supervised (Decision Tree Analysis) DM techniques – see section 4.3. These techniques identify five key problems that effect student progression. These are: Course, Stress, Distraction, Examinations and Leave. All of these problems are related and whilst the findings agree with previous research, the transferability of the results to other departments and universities is questionable. Indeed, McGivney (2003, p102) points out that "[t]he evidence indicates that the reasons for withdrawal vary according to student group, the nature of the institution, the support available and the subject studied."

Through considering a number of key demographic features, Burley constructs two profiles that can be used to identify vulnerable and less vulnerable groups of students. These findings are then used to inform a focus group meeting, at SHU. From this thirteen recommendations are developed, which take into consideration such things as student service intervention and the quality of teaching received. In his reflective summary, Burley discusses some of the issues associated with his research, such as the sample size of the students interviewed. The research concludes with a number of recommendations for future research. The one that is pertinent to this study is:

> *The analysis of historical student data to build a DM model that can be used to predict student classifications.*

It is this recommendation that the current research seeks to take forward.

3.2 Locating the Research

> *"The field of student progression has been well documented over the years."* (Burley 2007, p01).

Numerous studies have been conducted into understanding the problems within HE - the majority of which focuses upon the American HE system (Yorke 1999). Whilst there is a plethora of material in this domain, it is important to consider research that is pertinent to the current HE environment. Indeed,

> *"[s]ince the beginning of the 1990s, changes in economic patterns have combined with changes in education policy and structure to create a new landscape for adult learners." (McGivney 2003, p03)*

There have been a number of notable changes that have affected the current situation. These include:

- the introduction of student loans during the 1990s;
- students contributions to tuition fees (in 1998 and in 2006);
- attempts to increase and widen HE participation; and
- the recent HE funding reforms and the consequent reduction in student numbers.

Therefore, this review will, in the main, consider research taken from the mid-1990s onwards. As stated previously, Burley (2006) identifies three texts that are useful in understanding the subject area of student progression. These are:

1. Staying or Leaving the Course by Veronica McGivney (2003);
2. Keeping Students in Higher Education by David Moxley *et al.* (2001); and
3. Leaving Early by Mantz Yorke (1999).

Since Burley's research, the following study was also identified:

4. Retention and Student Success in Higher Education by Mantz Yorke and Bernard Longden (2008).

In understanding the problem of student progression in HE, Yorke and Longden (2008) suggest that the problem should be considered from three perspectives, the student, the institution and the state. They point out that a student's failure to progress will have an impact on all of these stakeholders, the most obvious being financial loss. "The institution may not receive its full public funding entitlement if the student does not complete a period of study [...]" (Yorke and Longden 2008, p10), which will vary depending on the course and its funding structure. They state that the pressures to improve progression in undergraduate HE courses became increasingly important during the mid-1990s. The estimated cost of non-progression to English institutions during 1999, for full-time first degree student, was in the region of £74 million.

Yorke and Longden (2008) provide some useful background information of four different HE systems - Australia, South Africa, the UK and the United States of America. From here they go on to dissect the HE systems of Australia, South Africa and the UK – all of which prove to be useful in the proceeding sections. They then go on to discuss how best to interpret institutional data in relation to the numerous performance indicators. They introduce some of the key theoretical ideas that have been formed over the years and consider them from three perspectives: psychological, sociological and other. They argue that current theory isn't extensive enough to understand the problems associated with student progression.

Through looking at the result of two large scale qualitative surveys, conducted into full-time and sandwich students, Yorke and Longden highlight some of the more important reasons behind why students fail to progress. They argue that quantitative results do not go far enough to understanding the problems and that a better understanding is gained through considering quantitative data in conjunction with qualitative results – a mixed methods approach. It identifies four general categories as to how HE stakeholders can aid progression. These are: facilitating the student's decision-making about courses; improving the student's experience of the course and institution; helping students to cope with the demands of the course; and understanding that events impact on students' lives outside the institution. Their research has a large qualitative element that is aimed at students who already withdrew. In their concluding chapter, Yorke and Longden look at ways in which institutions, students and the HE systems can improve student progression.

McGivney (2003) offers an institutional perspective to understanding the problems associated with student progression. She provides some background to the problem by explaining how the FE and HE landscapes have changed over the last ten to fifteen years, brought about "[...] by increasing flexibility in entry requirements, course structures, learning modes and assessment methods." (McGivney 2003, p03). Her research was conducted with mature students, those aged twenty one and over, in further or HE during 1995. The findings are made up from mail surveys, previous research (such as Kember

1995), consultations with institutional representatives, Access Validating Agencies and subject matter experts.

McGivney highlights that before the mid-1990s very little data was collected, regarding student progression and attrition patterns, due to data collection problems and reluctance on the institutions part to do so. She examines non-completion rates from an institutional perspective and notes that national figures provide inadequate measures. She highlights that comparisons between institutions are difficult due to differences in how institutions measure progression and collect data. She then goes on to examine the variables that affect completion rates and highlights the issues associated with measuring these. She suggests that results from such studies are only meaningful within the context of each individual institution or subject area. Indeed, McGivney suggests that progression will vary depending upon the institution, student cohort, subject area, type of course and the mode of learning.

According to McGivney, there are six ways in which students can exit from a course, these are:

1. Non-starter;
2. Informal withdrawal;
3. Transfer to other programmes;
4. Academic Failure;
5. Formal withdrawal; and
6. Non continuer.

She goes on to examine some of the more common variations and findings associated with progression. From these she concludes that both the institution and the individual have a responsibility for ensuring student progression. McGivney suggests that institutions can make improvements by improving the information that students receive. The research concludes with a look at the various support mechanisms available to students.

Moxley *et al.* (2001) adopt a qualitative approach to understanding the problems associated with student progression. They are strong advocates of the previous national drive to widen participation. Indeed, they attest that:

> *"higher education should not be closed to those individuals who wanted to improve themselves" (Moxley et al. 2001, pix).*

The authors recognise that there is no individual panacea to improving student progression, as retention methods should be individualised to each institution. They argue that student progression is the responsibility of both the institution and individual and they observe that student progression is about more than achieving academic standards – the student experience.

From their findings Moxley *et al.* (2001) develop a 'Pathway to Retention Model', which provides a number of objectives and supportive practices to help to facilitate student progression within institutions, see below:

OBJECTIVES

Objective 1: The institution perceives a need for retention

Objective 2: The institution establishes retention as an institutional aim

Objective 3: The institution expands involvement in retention and creates partnerships that support and contribute to the success of students

Objective 4: The institution builds a retention capacity and establishes a formal programme for keeping students in higher education

Objective 5: The institution keeps students enrolled and persisting towards the fulfilment of their educational aspirations and aims

SUPPORT PRACTICES

Support Practice 1:	Emotional support and sustenance
Support Practice 2:	Informational Support
Support Practice 3:	Instrumental Support
Support Practice 4:	Material Support
Support Practice 5:	Identity Support

These will be discussed in further detail in section 3.2.4. According to Moxley *et al.* (2001) student progression can be improved through what they call *proactive retention*. This is the art of informing and teaching students how to become students. They suggest that institutions can facilitate this through providing relevant student support systems and they highlight that academics and student support services are vital to achieving this.

Yorke (1999) uses both a mail and telephone survey to investigate the reasons behind student attrition – mixed methods. The focus of his research is on students who had already failed to progress with their full time sandwich degree courses in 1994-95, at six institutions situated in the North West of England. Yorke identifies some potential areas of where bias could be introduced to his research and he attempts to reduce this through a telephone survey. Yorke suggests this type of research is important for political reasons, as governments are holding HE institutions to account for their expenditure. From his research Yorke (1999, p39) identifies six factors as to why students fail to progress, these are:

Six Factor Solution

Factor 1:	Poor quality of the student experience
Factor 2:	Inability to cope with the demands of the programme
Factor 3:	Unhappiness with the social environment
Factor 4:	Wrong choice of programme
Factor 5:	Matters related to financial need
Factor 6:	Dissatisfaction with aspects of institutional provision

Yorke also reviews a number of models that have been developed to help facilitate an understanding of the student progression problem – see section 3.2.4. Arguably, the most famous of which is Tintos (1993) model of departure. Yorke appears to be critical of this model citing that it is too general in its approach and that the HE in the UK is funded differently. Overall, Yorke appears to advocate the widening of participation but attests that this cannot be achieved without a risk to student progression.

3.2.1 Widening Participation

> *"[...] participation in higher education had widened considerably over the preceding two decades, but there was still under-representation of young people from poor backgrounds and from some specific ethnic minority groups." (Yorke and Longden 2008, p50).*

Arguably, for the stability of the British Economy, it is imperative that Britain maintains a diverse and well educated workforce. Indeed, Martinez (1996) warns that Britain is falling behind many of its major competitors. These concerns were further reflected in the previous Labour Governments' target of 40% of adults in England to have received a university education by 2020 (Geoghegan 2009).

Widening participation is about much more than increasing the numbers of students entering HE (Kennedy 1997). Indeed, widening participation is ultimately concerned with allowing non-traditional students access to a HE qualification, particularly those from poorer backgrounds and ethnic minorities (Yorke and Longden 2008). Archer (2002) notes that these types of non-traditional students are being catered for by the post 1992 universities. SHU is one of the post 1992 universities, which has managed to increase participation whilst also improving student progression from 91.2% in 2001-02 to 92.3% in 2004-05 (National Audit Office 2007). However, numerous studies warn that increasing access to HE cannot be achieved without the risk of non-completion (Yorke 1999, Peelo and Wareham 2002).

Indeed, numerous other studies have noted that the opening up of HE in this way, to non-traditional students, has the potential to increase inequalities as opposed to tackling them (Yorke 1999, Peelo and Wareham 2002). Archer (2002) points out for example that students from poor backgrounds are likely to take on increased work, during term time, to reduce the financial burden. These students are also more than likely to graduate from university with larger debts due to a lack financial support from parents (Callender 2001).

In addition to the financial considerations, the drive to widen participation has also resulted in an increase in the number of local students (Archer 2002). In 2006-07 around 20% of students were local to the institution, this is an increase of around 12% since 1984 (Coughlan 2009). Slack and Casey (2002) warn that home students (who traditionally are catered for by the post 1992 universities) are likely to develop different relationships, to that of non-local students, with the institution and their colleagues. They go on to state that local students may have other pressures and commitments outside of university that inhibits them from taking part in the extra-curricular activities of a conventional student. This suggests that the potential lack of local student integration into the institutional society could increase the risks of non-student progression.

According to a report by the National Audit Office, the cost of widening participation to non-traditional students is on average around £900 per student, this was addressed in 1999-2000 with the introduction of a new funding scheme called the 'widening participation element' (National Audit Office 2007). In 2003-04 the Funding Council added a retention element to this, the reasons for this were "[...] to remove a disincentive to recruit students who may be more likely to leave early." (National Audit Office 2007, p30). In 2006-07, the total expenditure allocated through this funding stream was £345 million. However, due to recent changes in the funding of HE, the widening participation programme has been discontinued and student bursaries are the responsibility of the institution (Crown 2011).

3.2.2 Improving Student Progression

> *"In 2000 the UK government indicated that its commitment to expanding and widening participation in higher education should not be accompanied by lower levels of programme completion." (Yorke and Longden 2008, p50)*

It would be fair to say that the problem of student progression came to light during the mid-1990s. Indeed, during the 1990s there was considerable growth in the number of students entering full and part-time education, which placed significant pressure upon public finances to fund the extra places. Pressure was placed upon institutions to widen participation, particularly those from poorer backgrounds and minority groups, and to ensure that students already within the system progressed (Yorke and Longden 2008). The current economic climate could have a positive impact on student progression. Indeed, according to a 2009 BBC News article, it is estimated that the government will only fund an extra 10,000 new places. Arguably, as institutions tighten their admission processes, this will have a negative impact on the effort to increase student access to HE whilst potentially improving student progression (Geoghegan 2009). However, the dynamics of HE are set to change again, post 2011, due to increase in fees and government control over student numbers (Crown 2011).

The number of targets (indicators) that an institution has to meet is a good indication of how important an issue student progression has become, as this provides a way for the state to measure its expenditure. According to Yorke and Longden (2008), there are many indicators for measuring the performance of HE institutions in the UK. During the mid-1990s, a number of student progression indicators were developed (along with measures for monitoring access) to measure institutional performances, these included:

- *"Rates of non-completion following the first year of full-time undergraduate study;*
- *Projected completion rates for full-time undergraduates;*
- *Demographic data relating to participation (such as the proportion of entrants from 'working class' backgrounds, and of 'mature' entrants) and;*
- *Employment following graduation." (Yorke and Longden, 2008:64).*

Further indicators include institutional league tables and rankings (published in both the Guardian and Times newspapers), and data published by the HEFCE. Yorke and Longden (2008) warn that indicators don't provide a full enough picture to understand the problems of student progression. They suggest that a differentiation needs to be made between those who fail to progress for institutional reasons and those for personal reasons (outside the institutions control), Tinto (1975) refers to this as academic dismissal and voluntary withdrawal.

Through undertaking an extensive review of the literature in this area, a number of reoccurring themes were identified. Burley (2006) refers to these as *Casual Problems*, *Modelling* and *Intervention*. Arguably, these problems are stakeholder specific thus it is suggested that these reoccurring themes can be grouped into the following three categories:

a. Stakeholder Influences;
b. Theoretical Perspectives; and
c. Methods for Intervention.

The proceeding sections will consider each of these categories separately.

3.2.3 Stakeholder Influences on Progression

> *"The negative aspects of withdrawal, however, represent a waste of resources and of opportunity for students and universities alike, and for the broader society."* (Pitkethly and Prosser 2001, p186)

Student progression is influenced by three main stakeholders, the institution, the student and the state (Yorke and Longden 2008). Arguably, all of these stakeholders have a responsibility for student progression and the student experience. In their 2008 book, Yorke and Longden compiled a comprehensive list of suggestions as to how institutions, students and the state can help facilitate progression. They suggest that progression could be improved through focusing stakeholder efforts in four areas:

- Facilitating the students decision-making about courses;
- Improving the students experience of the course and institution;
- Helping students to cope with the demands of the course; and
- Understanding that events impact on students lives outside the institution.

Therefore, what follows is an in depth review of these areas in relation to each stakeholder.

3.2.3.1 The Institution Dimension

> *"It is important to understand that universities and colleges do not simply react to student expectations. They shape them as well."* (Ramsden no date, p03)

The institutional dimension is dominated by a number of reoccurring themes. Since the mid-1990's institutions have been placed under significant pressure to widen participation and improve student progression along with the student experience (McGivney 2003). This has resulted in: institutions having to make improvements in the way that they record and measure student progression, increases in financial pressures and the adopting of more flexible approaches to studying and part-time employment (Yorke 1999, Moxley *et al.* 2001, McGivney 2003, Yorke and Longden 2008). With the advent of capped student numbers and increased tuition fees it is becoming questionable whether the goal of widening participation is still being pursued with as much importance. This section will mainly focus on the four areas introduced above, and given the quantitative nature of this research, consider potential difficulties that may be faced in measuring progression.

It is widely acknowledged that institutions need to provide more information to help students to select their programme of study to make a more informed choice (McGivney, 2003). Ramsden (no date, p12) states that "[...] students are often poorly informed about what they can expect. Martinez (2001) suggests that students who feel well informed, about their programme, are more likely to progress. He points out that the evidence indicates that students fail to progress due to:

> *"insufficient understanding [...] of the demands of their course (eg the balance of practical and classroom work, assessment requirements and the balance of different components of the course)"* Martinez (2001, p04).

This is also highlighted by Yorke and Longden (2008) and McGivney (2003) who suggest that student expectations could be better managed by providing additional information on:

- *"course content;*
- *methods of assessment;*
- *work placements;*
- *expected time-commitment;*
- *ancillary costs;*
- *success rates of past students; [...]*
- *employment; [... and]*
- *the quality of the student experience.*" (Yorke and Longden 2008, p134).

However, institutions are trying to address this with the introduction of Key Information Set (KIS) Statements (see glossary page viii) for every course as from 2013. In addition to this, students should also be able to:

- attend organised open days;
- access specific programme information;
- be given an opportunity to visit individual departments; and
- obtain answers to question such as "what can the course offer me?" and "Is this course right for me?".

Further to this, institutional literature tends to be compiled in a manner that can alienate students on the basis of their age, gender, disability and ethnicity (Yorke and Longden 2008). Furthermore, Institutions could improve student progression by thoroughly assessing the students suitability, to the institution, and making students aware of the practices and expectations of HE (Yorke and Longden 2008, Thomas 2002). It is also recognised that accepting students on programmes without the key entry qualifications, to complete the course, has a negative impact on progression rates (Yorke and Longden 2008, Moxley *et al.* 2001).

> *"[...] institutions are likely to maximise their students' chances of success if they pay particular attention to the first year experience" (Yorke and Longden 2008, p136)*

It is fair to say that there has been quite a large burden placed upon the institution to improve the student experience (Thomas 2002). The literature, in this area, tends to focus on the institution engaging with students, at two levels General (social interactions) and Academic, before and after they have entered into HE.

> *"There is no one way to address readiness. In the United States, undergraduate courses are increasingly using the first two terms as periods in which to socialize students into a culture of post-secondary or higher education."* (Moxley *et al.* 2001, p114).

Yorke and Longden (2008), McGivney (2003), Thomas, (2002), Martinez (2001) and Ramsden (no date) all note the importance of providing opportunities to encourage social interactions between students and academics, and build good initial impressions. They suggest that institutions should minimise the number of unsystematic and bureaucratic arrangements and provide a welcoming and effective induction. The facilitation of exchanges between peers and tutors are believed to be vital in fostering the opinion that an institution is offering a good social experience (Thomas 2002, Ramsden no date). In addition to this, the centralisation of institutional support services are seen as being key to resolving student problems, in the most efficient and effective manner, and facilitating the student experience (Yorke and Longden 2008). It is widely recognised that students are more likely to progress if they feel like they belong at the institution (Tinto 1993). This can be aided by the institution:

- promoting a sense of community amongst the student population; and
- preparing information about the local area, where the university is situated (Yorke and Longden 2008).

Additionally, institutions should be prepared to help students to become familiar with their environment and promote a sense of academic and social wellbeing (Yorke and Longden 2008). This should be done whilst also supporting academic staff to develop their teaching expertise. This will ultimately enable the students to engage better with the HE process (Yorke and Thomas 2003, Ramsden no date).

Research indicates that academic efforts and resources should be focused on improving the first year student experience, in that students are more likely to complete their degrees if they progress beyond the first year (Yorke and Longden 2008, McGivney 2003, Martinez 2001, Ramsden no date).

> *"[R]etention efforts that focus on performance need to identify students who struggle academically, assess their situations and develop individualized plans that advance their skills, competencies and proficiencies"* (Moxley *et al.* 2001, p83).

Institutions need to support the student's transition into HE by building a culture of support and learning, ensuring that teaching approaches and programme structures are conducive to student success and through making good use of formative assessments (Ertl and Wright 2008, McGivney 2003).

> *"Students felt disadvantaged by a lack of background knowledge because courses were sometimes pitched at a level which assumed some prior knowledge."* (McGivney 2003, p125).

It is acknowledged that Institutions can improve progression by understanding the students pre-existing level of knowledge. Student's current level of knowledge should be assessed, before entry, so that suitable learning experiences and materials can be provided to help bring them up to speed (McGiveny 2003). Students can be brought up to speed through pre-entry workshops or as part of induction sessions at the start of the first semester. However, Ramsden (no date, p12) suggests that: "induction should be seen as a lengthy process rather than an event". Expectations must therefore be clearly defined, from the start of the course, and exercises ought to be undertaken to assess that the students approach to HE is suitable. Institutions need to make study support and student mentors available to aid the student's transition into HE. The provision of formative feedback is recognised as being fundamental to helping the student's transition into HE. Indeed, academics should make constructive criticisms and help students to improve future assignments. Early academic failure needs to be seen as an opportunity to succeed further on in the process. Institutions should adopt, what Yorke and Longden (2008) term as, a 'not yet competent' perspective to failure, which encourages the student to understand gaps in their learning and take the appropriate actions to eventually become successful. Institutions need to be aware that there are many factors that influence non-progression, these can be individual or a combination of issues and include:

- problems grasping subject matter;
- misunderstanding of what was expected;
- problems with exam nerves;
- incorrect choice of course; and
- lifestyle unsuited to learning (Yorke and Longden 2008).

Institutions should promote early failure as an interim problem that can be worked on so that the student progresses.

Institutions need to work with students to help them deal with conflicting external pressures. An example of which is part-time working, previously institutions had a no tolerance policy to part-time working. However, institutions have accepted the students need to undertake part-time work and there are examples of where the institutions have employed students on a part-time basis (Yorke and Thomas 2003).

> *"It is not so long ago that term-time working by students was a breach of the rules. However it is now looked upon by many as a necessity in order to help fund study [...]."* (Burley 2006, p13)

There are also other events, in addition to part-time work, such as illness and criminal attacks that can have an adverse effect on progression. It is suggested that Institutions should make allowances for this and offer help and support and, if possible, resist from making the student restart the year. Yorke and Longden (2008) suggest that institutions should invite potential non-progressors to an exit interview as this might help them to understand that withdrawal is not the only option available to them.

> *"Institutions are now required to monitor retention rates and collect and record student data more carefully and in more detail than in the past. However, concerns about funding and reputation have made non-completion a sensitive issue and institutions are not always keen to publicise their rates."* (McGivney 2003, p03).

It is recognised that the collecting of student data varies between institutions and there are problems when comparing such data on a national level. The main reasons for this are because of variation in data collection methods and in the definitions of, and way of calculating, non-progression. Indeed, some institutional data includes all types of non-progression including academic failure and transfers, others exclude transfers and forms of non-progression (McGivney 2003).

> *"Although the data published by HEFCE are undoubtedly of high quality, they do not fully illuminate the retention/completion picture. They do not differentiate between student departures that could (at least in part) be attributed to institutionally-related causes and those that arise from the students' own life style choices or from extraneous events"* (Yorke and Longden 2008, p71)

As a result there is a wide variety of data available, in terms of quality and quantity, and institutional methods and time-scales employed, in calculating non-completion rates after the first year of study, tend to vary from institution to institution. It has been noted that without any central direction, on the collection and recording of information, the accuracy of existing data will be questionable (McGivney 2003).

Further to this, Yorke and Longden (2008) highlight that quantitative research in HE appears to take two forms. The first looks at the analysis of datasets to identify correlations in student behaviour and the second form attempts to test theoretical models by combining results with demographic data. Finally it is perhaps important to point out that:

> *"It is obvious that not all types of withdrawal can be influenced by the university"* (Pitkethly and Prosser 2001, p186).

3.2.3.2 The Student Dimension

> *"In all post-compulsory education sectors, some degree of student loss is inevitable. [...] The fundamental question is why some leave and others do not."* (McGivney 2003, p85).

The majority of state commissioned reports place an emphasis on the Institutions responsibilities to improving student progression through improving the student experience (Parliamentary 2008). However, research studies suggest that the Institution shouldn't be held solely responsible for poor progression rates. Therefore, this section will discuss how students can help improve progression in relation to the four areas introduced above.

The importance of selecting the right programme of study is widely recognised as being one of the main influences on student progression (Yorke and Longden 2008, McGivney 2003, Martinez 2001). The literature suggests that students are less likely to progress when they make rushed and ill-informed decisions about their programme of study. It is noted that those students who take time to consider their reasons for entering HE and what they want to achieve in life are more likely to progress (Yorke and Longden 2004). However, the feasibility of this is questionable, given that the majority of students are selecting their programmes of study, whilst studying their A-levels or other level three programmes, with very little practical experiences (Davies and Elias 2002). Students can make a more informed decision if they spend some time researching, beyond what is provided by the Institutions their applying to, about their course and Institution (McGivney 2003). This includes speaking to careers advisors or friends, taking up paid or voluntary work, or taking sometime out to travel. The UK clearing is one example of where students are forced into making decisions with limited time to research their options (Richardson 2011). It is argued that students who are committed to their course are likely to cope better with the academic, social and financial pressures that they will face at some time during their studies (Yorke and Longden 2008).

Students have an important role to play in facilitating their own experience. Indeed, it is noted that a well-motivated student who is willing to work (not just transpose the work of others) and act on the feedback about their performance are more likely to progress (Yorke and Longden 2008). It is widely acknowledged that HE expects students to develop themselves to become autonomous learners and that the transition from level three study, where there is a high level of supervision, to that of HE can catch some students unawares (Yorke and Longden 2008). First year students are more likely to be caught unawares by assignment deadlines and instead of working constantly towards completing assignments they end up finishing them in a frantic rush. Students are expected to read deeper into their subjects, offer their own opinions (supported with appropriate literature) and manage their time so that they can plan their workload. One of the most prominent reoccurring themes in the literature, about the student experience, is the importance of students acting on the feedback received from the work they have submitted. In relation to this, Yorke and Longden (2008, p143) offer the following view:

> *"Some students - perhaps those more committed to performance goals than to learning goals - may merely note the grade (with or without satisfaction) and move on. The opportunity to maximize the learning potential is forfeited in such circumstances."*

Indeed, this point is also raised by Ertl and Wright (2008, p202):

> *"One common finding is that assessment can dictate to a considerable extent how students approach their learning, and that students focus on what is assessed."*

Not all students have the skills to cope with the demands of HE and their programme and progress (Moxley *et al.* 2001). It is suggested that first year students need to be prepared for the possibility of obtaining low grades and be mature enough to use this as tool to stimulate their learning (Ertl and Wright 2008). Not all students will have the right skills, at the start of their programme, to prepare assignments that meet the expectations of the institution and/or programme (Moxley *et al.* 2001). According to Yorke and Longden (2008), first year students need to develop their skills so that they are able to identify weaknesses in their work and take the necessary action, such as asking for help. Most HE institutions make some allowance for the student to develop, this is reflected by the fact that first year grades don't have a significant impact on the student's final degree classification (Yorke and Longden, 2008).

The external influences effecting student progression are well documented, these relate to managing finances, living arrangements, personal attacks and inappropriate behaviour (Yorke and Longden, 2008). It is widely acknowledged that the managing of finances can be difficult for students, especially for those who are leaving home for the first time. Students living arrangements have also been shown to have a significant influence on progression; these problems arise from failing to get on with their house mates to burglaries. The effects of irresponsible exposure to alcohol and drugs are also well documented as having an adverse effect on academic work and ultimately progression (Burley, 2006).

3.2.3.3 The State Dimension

> *"Governments around the world are increasingly calling higher education to account for the money that is invested in institutions [...]. The failure of undergraduate students to complete their studies is a cost to a government which funds higher education institutions [...]. A government's concern to keep public spending as low as possible means that the overt aspect of its economic agenda is best served by minimizing non completion [...]."* (Yorke 1999, p01)

It is the government's responsibility to ensure that there are adequate funding and quality systems that will help to foster engagement and partnership between students and institutions.

> *"Governments and agencies should be ready to introduce funding models and quality systems that will realise a vision of higher education as an engaged partnership between students and providers"* (Ramsden no date)

It is worth noting that from 2011 the government's priorities have changed from directly funding institutions to assessing quality, as measured by the Quality Assurance Agency (QAA) - see glossary page ix. The funding of universities is now through student fees which are set by the individual universities. The four areas introduced previously will be considered in relation to how the state can help improve the HE system to support progression.

Governments should help facilitate course selection by operating a more flexible post-qualification entry system as the inflexibility of the current systems works to the disadvantage of the student (Davies and Elias 2002). The majority of applicants who enter HE are accepted on the basis of predicted exam grades and conditional acceptances narrow down the students options, if they fail to meet the expected grades (Davies and Elias 2002). In addition to this, if applicants achieve a better or worse grade (in the subjects they have studied prior to entering HE) their choice of institution or programme may also change (Davies and Elias 2002). Indeed, applicants are more likely to make ill informed decisions when they are forced into making a rushed decision (Davies and Elias 2002).

> *"The student experience is currently high on the political and policy agenda"* (Ertl and Wright 2008, p195).

Governments should ensure procedures are in place that recognise and reward teaching (Ramsden no date.). In some non UK institutions teaching expertise are seen as vital to obtaining promotion within the institution. Government policies should not distract the institutions attention from the student experience. In that institutions should not be encouraged to seek other funding, for example research performance, at the expense of learning and teaching (Yorke and Longden 2008).

Governments can help students and institutions with programme demands by ensuring that there are guidelines on what institutions should deliver to ensure that the students have an experience that is of a reasonable quality and perceived as being value for money (Yorke and Longden 2008). However, these guidelines should merely be used to inform best practice as the institution is best placed, at the local level, to determine what qualifies as a quality experience (Pitkethly and Prosser 2001).

The perception of value for money will not be realised until the student gets the opportunity to reflect on the educational experience and the realisation of the economic rewards of obtaining the qualification (Yorke and Longden 2008). In addition to this, time-scales for collecting completion results may discriminate against institutions whose students come from less well-off backgrounds and take longer to complete their studies as they have to deal with external influences beyond their control (Yorke and Longden 2008, McGivney 2003). Finally as student contributions increase (and they exert more of a consumer like role as regards participation), the less significant the completion of progression statistics become at a national level (Yorke and Longden 2008).

Governments should minimise external influences by ensuring that funding systems for students are as straightforward as possible (Yorke and Longden 2008). The more complex the HE funding system is the less likely the students will be to take full advantage of the support that is available and may be less likely to progress (Yorke and Longden 2008). It is also important that the initiative designed to support certain student groups are supported by other initiatives. It is suggested that students from poorer backgrounds are at greater risk of unlinked initiatives (Yorke and Longden 2008).

3.2.4 Development of Models

> *"The theoretical literature on retention has drawn inspiration from a range of disciplines – psychology [...], sociology [...] and organizational behaviour [...] – though in no case can it be convincingly argued that the theoretical formulations that have been produced are monodisciplinary in character."* (Yorke and Longden 2008, p76).

A number of models have been developed over the years that have attempted to model student progression, the most famous of which is Tinto (1975)'s model of departure. Since the conception of Tinto's model authors have either made enhancements to his work or constructed completely new models. Yorke and Longden (2008) suggest that these models can be considered from three perspectives – physiological, sociological and other. This section will therefore review a number of the more prominent progression models and will identify the main factors that affect progression.

> *"Tinto's work, developed over a considerable time, has been very influential in studies of retention and attrition." (Yorke and Longden 2008, p76).*

Arguably Tinto's longitudinal model of institutional departure is one of the most prominent models of student progression (Yorke and Longden 2008). Tinto has developed his model from the mid-seventies to the early nineteen nineties. Central to Tinto's approach is the transition from one culture to another. The development of his model draws inspiration initially from Durkheim's (1951) theory of suicide and, later on, van Genneps (1908) study of rites of passage. The model itself considers progression from the student perspective but is relatively weak at addressing the external factors that influence student's perceptions, reactions and commitments (Yorke 1999). Yorke (1999) also goes on to criticise the model for its lack of emphasis on the institutions contribution to non-progression.

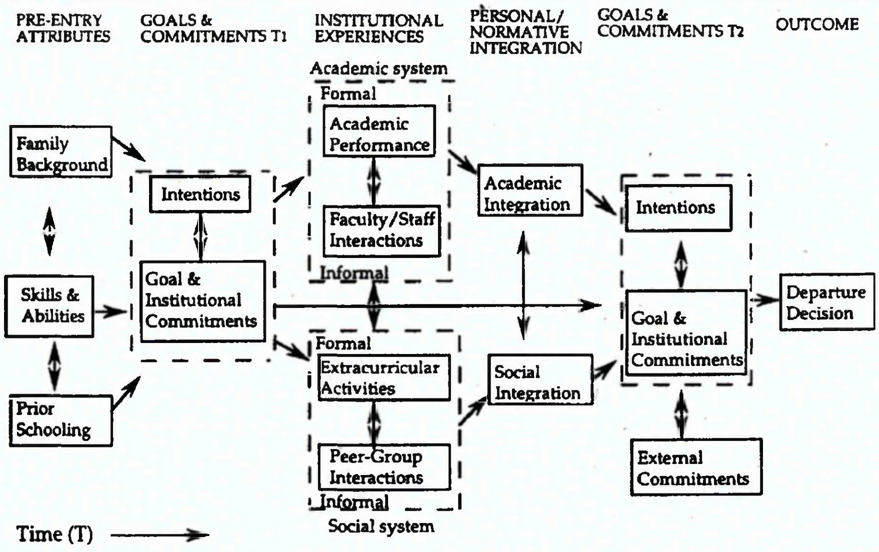


Figure 3.1 – A longitudinal model of institutional departure (Tinto 1993, p114).

Tinto (1993) argues that student progression can be anticipated by the student's level of academic and social integration. According to Tinto (1993), the key areas that influence progression, in relation to academic integration, include:

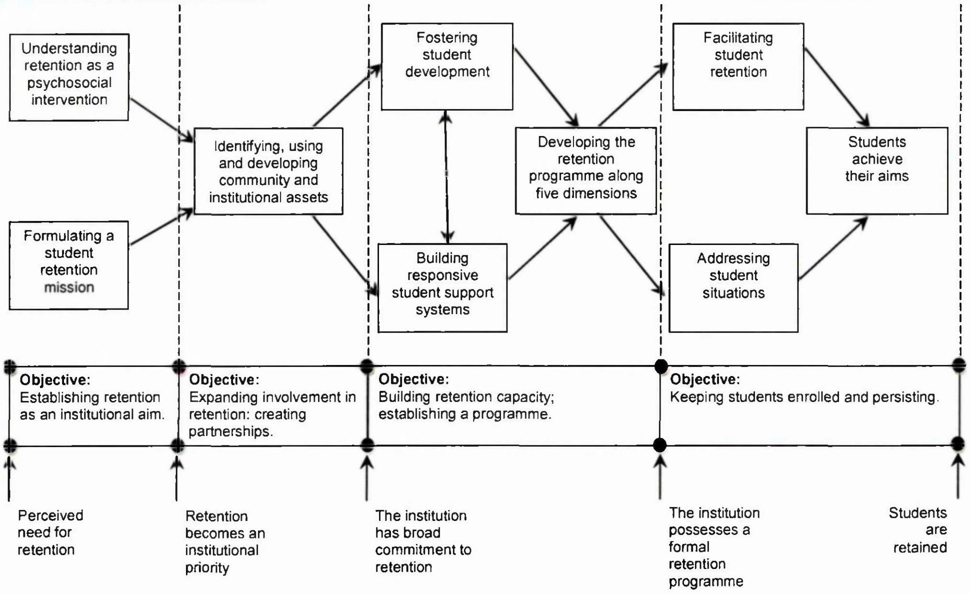
- assessment performance;
- personnel development;
- academic self-esteem (students perception of progress);
- enjoyment of studying the subjects;
- identification with academic norms and values; and
- the students identification of their role as a student.

The areas of social integration that Tinto identifies has being fundamental to student progression pertain to:

- number of friends;
- personal contact with academic staff; and
- enjoyment of their time at university.

Tinto (1993) points out that these factors develop over time, as integration and commitment interact, and he suggests that progression is dependent on the student's commitment at the time of decision.

The Pathway to Retention model developed by Moxley *et al.* (2001) considers progression from the institutional perspective.



Understanding retention as a psychosocial intervention
Formulating a student retention mission
Identifying, using and developing community and institutional assets
Fostering student development
Building responsive student support systems
Developing the retention programme along five dimensions
Facilitating student retention
Addressing student situations
Students achieve their aims
Objective: Establishing retention as an institutional aim.
Objective: Expanding involvement in retention: creating partnerships.
Objective: Building retention capacity; establishing a programme.
Objective: Keeping students enrolled and persisting.
Perceived need for retention
Retention becomes an institutional priority
The institution has broad commitment to retention
The institution possesses a formal retention programme
Students are retained

*Figure 3.2 – Pathway to retention (*Moxley *et al. 2001, p20).*

They argue that student progression can be facilitated, in higher or post-secondary education, by achieving five objectives and providing five support practices. These relate to:

Objectives

Objective 1:	The institution perceives a need for retention
Objective 2:	The institution establishes retention as an institutional aim
Objective 3:	The institution expands involvement in retention and creates partnerships that support and contribute to the success of students
Objective 4:	The institution builds a retention capacity and establishes a formal programme for keeping students in higher education
Objective 5:	The institution keeps students enrolled and persisting towards the fulfilment of their educational aspirations and aims

Support Practices

Support Practice 1:	Emotional support and sustenance
Support Practice 2:	Informational Support
Support Practice 3:	Instrumental Support
Support Practice 4:	Material Support
Support Practice 5:	Identity Support

Moxley *et al.* (2001) point out that their model does not provide direction as to how the student experiences progression, or how the progression effort is organised for each individual student.

> *"[...] we remain unconvinced that a single theoretical formulation – a 'grand theory' – can be constructed to include all of the possible influences that bear, via the student's psychological state, on retention and success, whilst being practicable in terms of research and institutional practice."* (Yorke and Longden 2008, p84).

Yorke and Longden (2008) provide a schematisation of the influences that affect student progression.

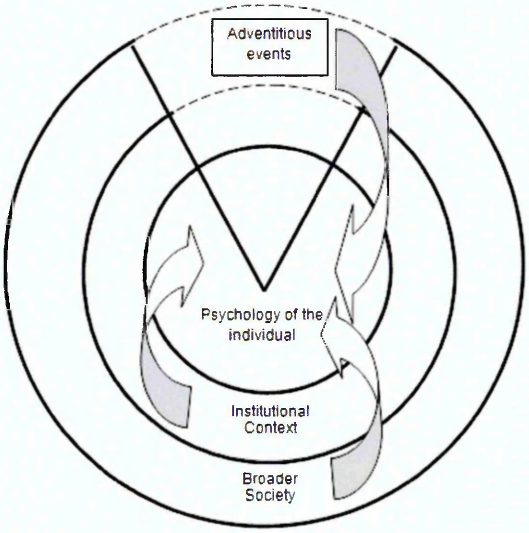


Figure 3.3 – A schematization of the influences on students psychological state (Yorke and Longden 2008, p85).

They suggest that this can be used to inform thinking about the problem of student progression. Central to this model is the student's psychological state, when deciding whether to progress. They acknowledge that that the context of the institution and broader social environment exert an influence on the student's decision, but they go on to point out that there are many influences (outside of the control of the institution) that will impact the student's decision to progress. They argue that the chances of a student failing to progress are greater when the influences effect the students experience. Yorke and Longden (2008), point out that there is no single panacea to understanding student progression and that there are a number of causes (individually or collectively) that may excrete an influence on a student's decision to progress.

> *"[... R]etention and student success are influenced by a complex set of considerations which are primarily psychological and sociological, but which are in some cases influenced by matters that might be located under other disciplinary banners such as that of economics."* (Yorke and Longden 2008, p77)

3.2.5 Main Influences on Progression

> *"[W]e know that it [non-progression] is multi causal, that it is complex and highly context specific, but we also know that it is significantly caused by things which colleges and education centres can do something about."* (Martinez 1995, p23).

The main influences on student progression are well documented in the literature. Indeed, the research carried out by Martinez (1996), on progression in post-16 colleges, provides a good place to start. Through his research he identifies a number of broad issues that are thought to exert an influence on progression. These relate to:

Motivation: Issues might include little or no career and/or progression objectives, no real reason for choosing the institution, having to re-sit the course or transferring (after non-progression) to another institution.

Social: Problems may occur due to a lack of friends on the same course and/or a lack of support from family. Other issues in this area could include an imbalance in the age or gender of people on the same course.

Time Pressures: Difficulties in this area might relate to caring for sick relatives, being a single parent or having to work, in a part-time job, whilst also studying.

Financial: Here problems tend to relate to the loss of income support, delays in obtaining grants/loans, daily travel costs or examination fees.

Qualifications: When starting the course some students will only have obtained the minimum academic qualification and may lack studying experience.

Any other difficulties: These relate to unhappiness with certain aspects of the course, health problems, domestic circumstances, immaturity and travel difficulties.

Figure 3.4 – Issues that have a bearing on student progression (Martinez 1996, p16).

A 2007 National Audit Office survey into student progression issues highlighted that the most common reasons for non-progression were:

- *"personal reasons;*
- *lack of integration;*
- *dissatisfaction with course/institution;*
- *lack of preparedness;*
- *wrong choice of course*
- *financial reasons; and*
- *to take up a more attractive opportunity." (National Audit Office 2007, p25).*

In addition to this, the report also goes on to state that "[...] some students fail their assessments, are excluded or take an intermediate qualification rather than proceed with their original course." (National Audit Office 2007, p25).

In 2008, Yorke and Longden (2008b) carried out research for the HE Academy into the first year experience of HE in the UK. The research was carried out in two phases:

Phase 1: In 2006, a survey of first-year full-time undergraduate students from contrasting institutions was carried out. This assessed the student's perception of their experience of being a student.

Phase 2: In 2007, a follow up postal questionnaire of all those students who failed to re-enrol on the second year of their course in their original institution was conducted.

The results of phase 2 of the research identified a number of reasons as to why the students failed to progress. The most pertinent ones being:

- *"Poor quality learning experience;*
- *Not coping with academic demand;*
- *Wrong choice of field of study;*
- *Unhappy with location and environment;*
- *Dissatisfied with institutional resourcing;*
- *Problems with finance and employment; and*
- *Problems with social integration." (Yorke and Longden, 2008b, p06)*

From his 1995 survey, Martinez identifies five common factors that affect student progression. These relate to the students:

- justification for coming to college;
- source of information about the college;
- level of support received from the college;
- having problems with the programme; and
- personal and financial situation.

Similarly, Moore (1995) identifies a number of factors as to why students fail to progress and weights these in order of importance:

Course unsuitable/dislike	41%
Personal reasons	17%
Academic problems	11%
Financial problems	11%
Accommodation	6%
Other	14%

Moore (1995) identified that first term first year students were the ones who were most at risk of non-progression. It was determined that the main reasons for this were due to course related issues. However, student isolation and loneliness were also identified as being contributory factors. In response to this, Moore (1995, p23) attests that "students may find it easier to say they left for course related reasons than, for example, acknowledge personal problems." He goes on to point out that over half of the students surveyed indicated that the reality of HE was different from their expectations.

3.3 Educational Data Mining and Student Progression

> *"Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using these methods to better understand students, and the settings which they learn in." (Baker and Yacef no date, p02).*

This section will provide a brief overview of EDM (Educational Data Mining) before discussing EDM in relation to student progression. It will conclude by summing up the main point about EDM and student progression.

3.3.1 Educational Data Mining

It is fair to state that the majority of research undertaken in this domain has focused on student interaction with web-based learning environments (Romero and Ventura 2006, Baker and Yacef no date). However, a recent book by Romero *et al.* (2011) indicates that EDM can be divided into six main applications or tasks.

> The first is *using EDM techniques to help administrators and academics in analysing course activities and course usage information.* The techniques used in this type of area are exploratory data analysis through statistical analysis, visualisations, reports and process mining.
>
> The second relates to *the maintenance and improvement of courses.* The objective is to determine the best way of improving certain aspects of courses, by helping administrators and academics to use information about student learning and usage. Association, clustering and classification analysis have been frequently used in this area.

The third is in the *generation of recommendations in terms of which material would be most appropriate given the students current level of knowledge*. EDM techniques widely used in this area are clustering, association, sequencing and classification.

The fourth relates to *predicting learning outcomes and student grades*. Here the objective is to predict student classifications or other learning outcomes, such as non-progression, based on the data recorded by the institution. Clustering, classification and association are the most frequently used DM techniques.

The fifth relates to *the modelling of students*. There are numerous applications of user modelling in the HE domain, some of which relate to detecting student states and characteristics such as motivation, satisfaction, progress and problems. Commonly a student model is created from usage information. There have been numerous techniques applied in this area, of which the main ones are clustering, classification, association, statistical analysis, Bayes networks, psychometric models and reinforcement learning.

The sixth and final area relates to *the structural analysis of the domain*. This is concerned with using the ability to predict student performance as a measure of quality of a domain structure model. The most frequent EDM techniques applied here are space searching algorithms, association rules and clustering methods.

Romero and Venture (2006, p137) point out that “Data Mining can be applied to data coming from two types of educational systems: traditional classroom and distance learning”. Romero and Venture (2006), go on to highlight the most prominent DM applications that have been used in web-based learning environments, these include:

- Statistics and visualisation;
- Web mining;
- Clustering , classification and outlier detection;
- Association rule mining and sequential pattern mining; and
- Text mining.

However, a more recent study by Baker and Yacef (no date), comparing early EDM to EDM in 2008/09, suggests that relationship mining is becoming rare and that prediction modelling is increasing in popularity. Romero *et al.* (2011) indicate there are three avenues of EDM research, theses relate to:

- identifying which DM techniques are best suited to interrogating large HE data sets;
- determining the most appropriate questions to ask the data; and
- targeting EDM reports at the right HE stakeholders.

Previous research (Romero and Ventura 2006, Baker and Yacef no date, Dekker *et al.* 2009) indicates that little research has been undertaken in the area of EDM and student progression.

3.3.2 Student Progression Focused Educational Data Mining

> *"There is almost a complete absence of literature concerning Data Mining applied to Higher Education" (Burley 2006, p122).*

Indeed, whilst numerous authors have tackled the subject of student progression, very few have considered the use of BI as a method of understanding the problem (Dekker *et al.* 2009, Herzog 2006). What follows is a review of those authors who have used BI in understanding the problem of student progression.

Luan appears to be the leading expert in this field (Romero and Ventura 2006, Herzog 2006). Baker and Yacef (no date) also highlight that other key papers in this area include Superby *et al.* (2006), Romero *et al.* (2008) and Dekker *et al.* (2009). Burley (2007) has already been discussed in section 3.1.

Luan (2001) and Luan (2002) investigates the potential applications of DM techniques in HE. Both these papers attempt to address the question:

> *"What are the transferable techniques in data mining that are readily applicable in higher education?"* (Luan 2002, p04).

Luan's research is at a single college and is based on data taken in the autumn of 2000. This data is used to help illustrate the potential of both unsupervised (cluster analysis (Kohonen and K-means)) and supervised (Neural Networks) DM techniques, see section 4.3, to predict student persistence. The actual results appear to be of little significance as the data is used simply to demonstrate the applications of DM techniques. This research is then reworked and published in conjunction with SPSS® (Luan 2004). This work is similar to the research carried out by Burley (2007), using data taken from the Department of Computing at SHU and applies both supervised and unsupervised DM techniques to demonstrate the applicability of DM in HE.

In his research published with SPSS®, Luan (2004) introduces three, HE, case studies where such applications have proven useful and attempts to draw parallels between HE and the private sector. Luan substantiates his claim by showing the similarities between the business questions answered by DM in both the private sector and in HE. This is also evident in the research carried out by Burley (2007), who likens a HE institution to a supermarket where the students are shoppers and the student problems are the products purchased in the supermarket.

In a more recent study, Luan (2006) looked at predicting success rates of students across a number of courses at Cabrillo College. He assess data taken from a five year period and compares the success rates of students who received intervention (supervised tutoring) to those that didn't. The results are then analysed against a number of demographics, such as gender, ethnic groups and age. The results showed that students that underwent intervention were 10% more likely to be successful than those that didn't.

Superby *et al.* (2006) attempt to identify the factors that influence the achievements of first year university students, using DM techniques, at three dissimilar Belgian universities. Adapting the work of Parmentier (1994), they use a combination of questionnaire and assessment results to gather data about students:

- demographics;
- attitudes towards studying;
- perceptions/experiences of the institution; and
- average marks received in January 2004.

The student's average marks were then used to create three groups of students:

- low-risk of non-progression
- medium-risk of non-progression; and
- high-risk of non-progression.

Superby *et al.* (2006) attest that this information could then be used to identify those students who are in need of intervention during the first year of their degrees. Data was collected from 533 first year students from across all three universities during November of the 2003/04 academic year. This data was first used to attempt to identify the most correlated variables in predicting success and ultimately progression. This identified a number of factors associated with success, the main ones being course attendance and previous academic experience. Interestingly, gender, parental education level/occupation, number of brothers/sisters older/younger and whether or not they were already in HE were not identified as factors significantly correlated with success. Arguably, this is dependent on the type of institution and course studied (Dekker *et al.* 2009). Indeed, Dekker *et al.* (2009, p43) point out that "[a]ll studies show that academic success is dependent on many factors [...]".

Superby *et al.* (2006) split their data 70/30 into training and validation respectively, they use SAS®, Enterprise Miner and R Software to compare the predictability of four DM techniques (Decision Trees, Random Forests, Neural Networks and Linear Discriminate Analysis) in predicting students at low, medium and high risk of non-progression. Through comparing the results of the four models they find that Linear Discriminate Analysis was better at classifying students into low, medium and high risk groups. However, the total rate of classification for this method is 57.35% which the authors themselves note is not remarkable. They highlight disparities between the students at the three universities from where the data was collected from as some justification for the poor classification percentage of their best model.

Romero *et al.* (2008) compare a number of different DM techniques for predicting/classifying student's final marks. The research uses 438 records taken from seven different e-learning courses at Cordoba University using Moodle – this keeps detailed numerical logs of the activities performed by student in e-learning environments. From this student results are grouped into four categories (Excellent, Good, Pass and Fail), this results in an imbalanced data set as 59.81% of students failed and only 3.89% of students obtained excellent. In order to boost the minority variables they use random over-sampling and measure the quality of the induced classifier by the geometric mean. The original numerical data is also converted into categorical data, which results in three data sets:

- original numerical;
- categorical; and
- rebalanced.

A number of different DM techniques are then assessed against each data set. The results of which highlight that:

- Decision Trees, Rule Induction, Fuzzy Rule Learning and Neural Networks all perform well (more than 65% accuracy) with the numerical data;
- Decision Trees provided the best accuracy, more than 65%, with the categorical data; and
- Rule Induction and Fuzzy Rule Learning provided the best results, more than 60% accurate) with the rebalanced data.

Whilst the accuracy of these results are poor they do provide some indication as to the performance of different classification methods with different types of data.

Dekker *et al.* (2009) attempt to predict the non-progression of first year, Electrical Engineering, students at the Eindhoven University of Technology in the Netherlands. They obtain three data sets from the university, which they group into three categories:

- pre university data;
- university grades only; and
- number of attempts taken and grades achieved at each attempt.

Data is extracted for students who studied Electrical Engineering between 2000 and 2009, through this process 648 records of first year students were obtained. Dekker et al then use WEKA to compare two decision tree algorithms (CART and C4.5), a Bayesian classifier, a logistic model, a rule-based learner and a

Random Forest. In addition to this, they also used a OneR classifier to assess the predictive power of individual attributes. The results of their research indicate that the classifiers produce accuracy between 75% and 80%. In their concluding remarks Dekker *et al.* (2009) attest that the strongest predictors of success are the grades achieved in certain modules (mainly Linear Algebra) and that the most relevant information is collected by the university itself. In summing up Dekker *et al.* (2009, p49) point out that "[...] it is not easy to find an objective way of classifying students."

A number of research papers refer to the research carried out by Herzog (2006), which focuses on comparing the prediction accuracy of Decision Trees (CART, CHAID and C5.0) and Neural Networks (simple topology, multi-topology and three hidden layer pruned) with that of multi-nominal logistic regression in predicting student progression and degree completion time. He points out that:

> *"[... H]igher education research provides little insight into which specific data-mining method to use when predicting key outcomes such as retention or degree completion"* (Herzog 2006, p20)

Herzog (2006) gathers two types of data from Carnegie Doctoral Degree and Research University. He obtains 8,018 records of full time second year students, who started in the autumn semesters of 2000 through 2003, to predict student retention. In addition to this, he also collects 15,457 records of forth year undergraduate students from spring 1995 through to summer 2005 to predict time to degree completion. Herzog collects three types of data that relates to student:

- demographics
- academic, financial aid and residential situation; and
- parental income and transfers out of university.

Herzog (2006) draws upon previous studies and advocates that boosting algorithms can help to improve prediction accuracy of models by up to 80% when compared to standard algorithms of a neural network. Herzog (2006) uses SPSS® Clementine software and randomly splits his data 50-50 into training and validation. In doing so he determines that the results from the pruned neural network indicated that credit hour related predictors, student age, residency and stop-out time were the most influential on retention and degree completion times. Herzog (2006, p26) concludes that “On average the decision tree and neural network performed at least as good as the regression model.”

An example of where BI applications have been used in HE to identify students at risk of non-progression is at the University of Alabama. In the light of increased competition and ever decreasing budgets, the University of Alabama in conjunction with SAS® developed a student progression model. This was used to help improve the selection of first year students, at enrolment, and the early identification of students that require intervention. The model, developed by Professor Michael Hardin, looks at enrolment records and ‘freshman’ surveys to identify the variables that affect student progression. He found that American College Test scores, high school grade point average, college majors and parent’s education level to be key in predicting non-progression. This has had a positive effect on the Universities progression rates and hence university rankings. Indeed, the university predicts that the success rate of those students identified as requiring intervention will increase from 50% to 83%, which will have a positive effect upon university finances (University Business 2004).

3.3.3 Main Points on EDM and Progression

Since 2005, interest in the area of EDM has grown rapidly. The majority of research carried out in this domain has focused on the data collected from data gathered from computer systems on distance learning courses (Romero and Venture 2006). Applications of DM in HE are:

- Helping academics and administrators in analysing course information;
- Improving and maintaining courses;
- Generating recommendation as to the most appropriate course material;
- Predicting learning outcomes and student grades;
- Modelling students; and
- Analysing the structure of the domain (Romero *et al.* 2011).

Whilst numerous DM and statistical analysis techniques have been applied in these areas, the application of cluster analysis, classification analysis and association analysis seem to be more common. However, this is likely to change if research in prediction modelling is, as suggested by Baker and Yacef (no date), increasing in popularity in EDM.

Whilst the general area of EDM has grown rapidly few studies have been undertaken into using DM to help improve student progression. A number of studies in this area (Luan 2001, Luan 2002, Burley 2007) have assessed the applicability of DM in HE and, in doing so, have likened HE to that of a business where student are shoppers. In addition to this, a number of studies have been carried out into trying to find the best DM technique to predict progression (Superby *et al.* 2006, Romero *et al.* 2008, Dekker *et al.* 2009).

It is evident, from all of the student progression literature reviewed in this chapter, that the research carried out into student progression is very focused in terms of specific institutions and/or specific courses. In such cases data is either collected via a questionnaire and/or extracted from university systems.

> *"[a]ll studies show that academic success is dependent on many factors [...]".*(Dekker *et al.* 2009, p43).

In general the data gathered from such research tends to be less than a 1,000 records with the exception of a few studies that take data from a number of years directly from the university. However, taking large amounts of data over a long time period could skew the results due to policy changes in the academic landscape. Data gathered from such studies tends to be:

- Student demographics;
- Attitude /behavioural related;
- Perceptions/experiences related; and
- Results/number of attempts.

Such data has then been used to create a predictor variable that highlights the student's likelihood of non-progression. For example, Superby *et al.* (2006) used student's average marks to create three groups (high, medium and low) that highlighted the student's risk of non-progression.

> *"[... H]igher education research provides little insight into which specific data-mining method to use when predicting key outcomes such as retention or degree completion"* (Herzog 2006, p20)

Indeed, researchers in this area have applied a number of unsupervised and supervised (see section 4.3) DM and statistical analysis tools to the problem of student progression, these include:

- Classification
- Clustering;
- Neural Networks;
- Decision Trees;
- Rule Induction;
- Linear Discriminate Analysis.

> *"[...] it is not easy to find an objective way of classifying students."* (Dekker *et al.* 2009, p49)

The results from such research vary, as the results are dependent on the type of student, course and institution. However, in general results from such studies indicated that: attendance; experience of academia; grades achieved; student's age; living arrangements; and stop out time where fundamental to predicting non-progression. The majority of the factors effecting non-progression are also visible in the non EDM literature reviewed in section 3.2.

3.4 Summary

This chapter provides an overview of the literature on student progression, EDM and EDM in relation to predicting student progression. The review of student progression highlights that progression is the responsibility of the student, institution and the state. Efforts to improve progression are the responsibility of all three and should focus on the decision making process, the student experience, coping with course demands, and dealing with external events. The area of EDM and student progression has tended to focus on assessing the applicability of DM in the context of HE and trying to find the best DM technique to predict progression. The findings from EDM research tend to agree with the recommendation made in the literature about student progression. In that, course attendance, previous experience of academia, grades achieved, student's age, living arrangements and stop out time are all fundamental, in both the student progression and EDM literature, to improving non-progression.

4 Data Warehousing and Data Mining

This chapter discusses some of the Data Warehousing (DW) and DM concepts introduced previously and in subsequent chapters. It has been specifically created to present an overview of BI, DW approaches and methodologies, Data Quality (DQ) issues and DM techniques and methodologies. The chapter concludes by outlining what is new in the context of DW and DM in HE.

4.1 Background to Business Intelligence

According to Kimball *et al.* (2008) BI is all of the processes and systems used by an enterprise to gather, process, access and analyse data. Through a better understanding of its data an organisation can acquire a better understanding of itself. Kimball *et al.* (2008) go on to suggest that a Data Warehouse is the platform for all BI within an organisation.

4.2 Data Warehousing

A Data Warehouse is defined as "a collection of integrated, subject-oriented databases designed to support the DSS [(Decision-Support Systems)] function, where each unit of data is relevant to some moment in time. The data warehouse contains atomic data and lightly summarized data." (Inmon 2005, p495). A data mart is defined as "a departmentalized structure of data feeding from the data warehouse where data is de-normalized based on the department's need for information." (Inmon 2005, p494).

English (1999, p04) highlights that "Data warehousing projects fail for a number of reasons, which can be traced to a single cause: *nonquality."* According to English this nonquality is a result of a number of issues associated with the quality of data - data defects. In relation to data quality Greenfield (2004) identifies an informal taxonomy of DW errors, Greenfield summaries data errors into four categories these are:

a. Incomplete errors.
b. Incorrect errors.
c. Incomprehensible errors.
d. Inconsistent errors.

As this research will be creating a number of DM marts using the DW process, it is imperative to provide a definition of DW. Reed (no date) defines DW as:

> *"[...] what you need to do in order to create a data warehouse, and what you do with it. It is the process of creating, populating, and then querying a data warehouse and can involve a number of discrete technologies [...]."(Reed no date).*

Hence, DW can be thought of as the process of turning data into knowledge, a flow of information. Figure 4.1, below, shows this information flow from data through to knowledge Kimball and Ross (2002) refer to this as the four stages.

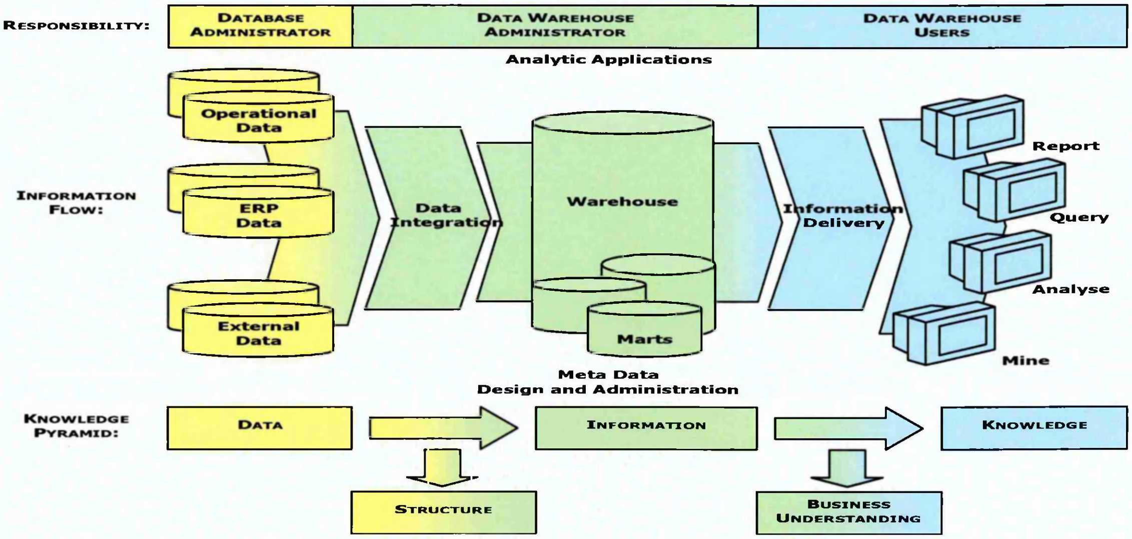


Figure 4.1 – Schematic of Information Flow adapted from (Oracle no date, p09, Marco 2003, Kimball and Ross 2002).

A DM mart is a clean, merged and reduced copy of the transactional data taken from the OLTP (On-Line Transactional Processing) systems, see glossary page ix. This would be located in the information flow, in figure 4.1, between the OLTP systems and the Data Warehouse. Kimball and Ross (2002) refer to this as the Operational Data Store (ODS), see glossary page x.

4.2.1 Top-Down Data Warehousing

The top-down or Enterprise approach to DW was founded by William. H. Inmon in the early 1990s and the DW terminology that is used today was, in the main, defined during this time. Inmon is known as the father of DW and is a keen advocate of the Entity-Relationship (ER) modelling approach to DW. The top-down approach begins with the building of the data warehouse. This is achieved through extracting the transactional data from one or more OLTP systems and then integrating this data within a normalised enterprise data model in a relational database. The data within the Enterprise Data Warehouse (EDW) is then summarised, dimensionalised and distributed to one or more dependent data marts as cubes. Each data mart derives data directly from the EDW (Eckerson 2004). According to Inmon (1999) a data mart is "a collection of subject areas organized for decision support based on the needs of a given department".

4.2.2 Bottom-Up Data Warehousing

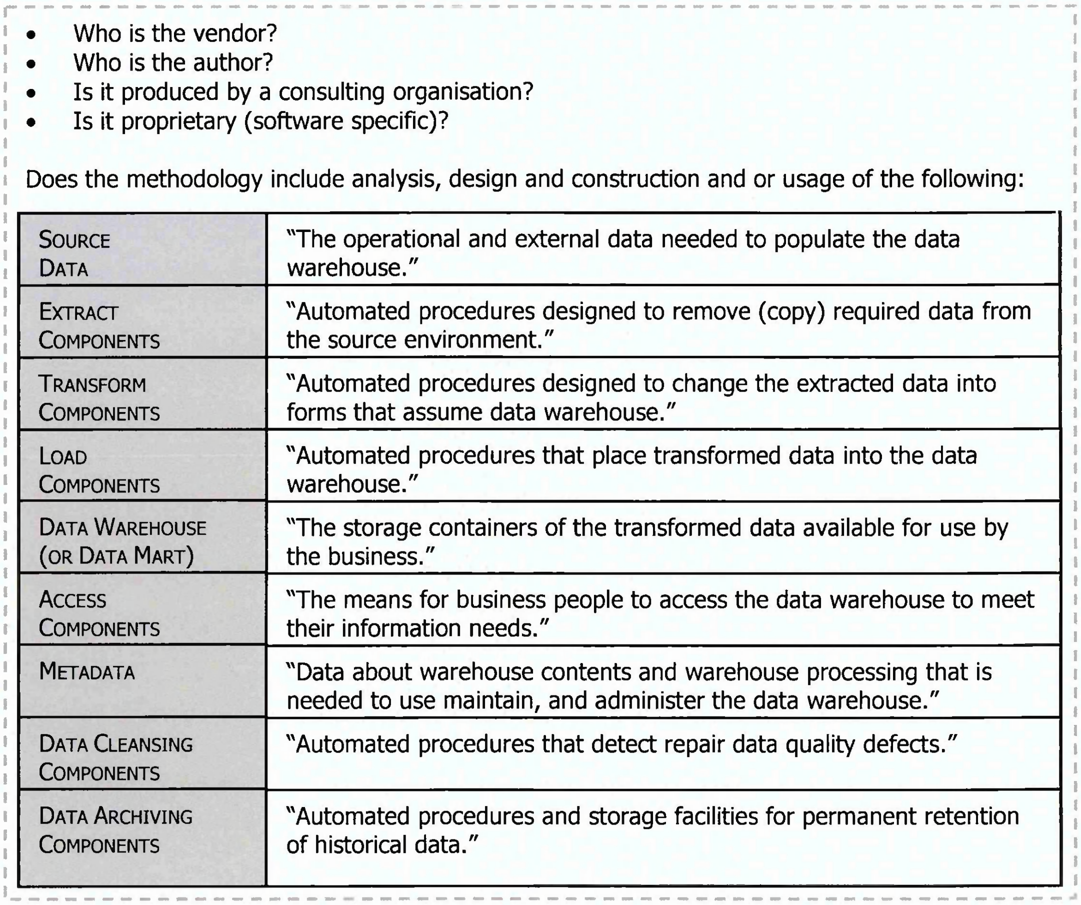
The bottom-up approach was developed by the DW consultant Ralph Kimball. Kimball is a noted expert in the field DW and he argues that a DW should be developed through dimensional modelling as opposed to ER modelling (Kimball 1997). The Kimball method is where individual data marts are constructed, which contain aggregated transactional data that has been modelled into star schemas for optimised usability and query performance. In this approach data is extracted from the OLTP systems, transformed and then loaded into data marts. The data warehouse therefore, consists of a number of consolidated independent data marts that have been built one on top of the other to allow users to query across them. Some argue that this approach negates the need for a data warehouse as it is seen as a development of end-to-end data marts (Eckerson 2004).

4.2.3 Data Warehousing Methodologies

Thomann and Wells state that a methodology

> *"is a detailed set of steps or procedures to accomplish a defined goal. [...] The primary purpose of a methodology is to achieve a predictable result. A secondary goal of methodology is to provide a process that is repeatable, trainable and consistent." (Thomann and Wells no date, p01).*

Thus a DW methodology is concerned with the stages of the DW process and what needs to be completed within those stages. There is no 'one-size-fits-all' methodology for implementing a Data Warehouse; numerous methodologies exist that are both formal and informal. The selection of a DW methodology is dependent on a number of factors, such as the needs of the individual organisation, process maturity and the like (Thomann and Wells no date). Indeed, the selected DW approach, discussed above, will ultimately determine the DW methodology to be used. Thomann and Wells have developed a method for selecting a methodology, this is summarised in figure 4.2.



- Who is the vendor?
- Who is the author?
- Is it produced by a consulting organisation?
- Is it proprietary (software specific)?

Does the methodology include analysis, design and construction and or usage of the following:

Source Data	"The operational and external data needed to populate the data warehouse."
Extract Components	"Automated procedures designed to remove (copy) required data from the source environment."
Transform Components	"Automated procedures designed to change the extracted data into forms that assume data warehouse."
Load Components	"Automated procedures that place transformed data into the data warehouse."
Data Warehouse (or Data Mart)	"The storage containers of the transformed data available for use by the business."
Access Components	"The means for business people to access the data warehouse to meet their information needs."
Metadata	"Data about warehouse contents and warehouse processing that is needed to use maintain, and administer the data warehouse."
Data Cleansing Components	"Automated procedures that detect repair data quality defects."
Data Archiving Components	"Automated procedures and storage facilities for permanent retention of historical data."

Figure 4.2 – Data Warehouse Methodologies Checklist

(Thomann and Wells no date, p03).

The methodologies associated with the approaches highlighted in sections 4.2.1 and 4.2.2 will be assessed against the criteria identified above in figure 4.2. Arguably, there is little value in assessing a federated methodology as the approach aims to integrate multiple heterogeneous data warehouses, data marts and packaged applications that already exist within an organisation. This is in contrast with the aim and objectives of this project (Eckerson 2004).

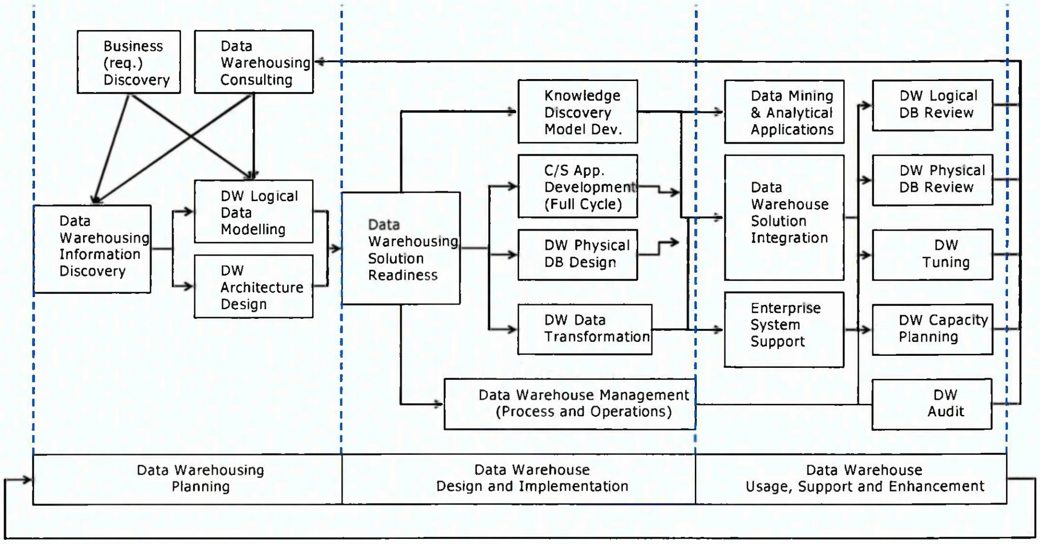
Therefore, this sub-section will assess the most prominent methodologies associated with the approaches introduced above The National Cash Register Company (NCR) Data Warehousing Method, Kimball's Business Dimensional Lifecycle Diagram (BDLD) and the SAS Rapid Data Warehousing Methodology (O'Donnell *et al*. 2002).

4.2.3.1 THE NCR DATA WAREHOUSING METHOD

The NCR method (figure 4.3) is a top-down approach to developing a data warehouse. The developers of the method, NCR consultants (along with William H Inmon), are strong advocates of ER modelling and the Inmon approach to DW. Hence, the method advocates the developing/deploying of an entire data warehouse as opposed to building individual data marts. There are three main phases to the NCR method:

- Data Warehouse Planning;
- Data Warehouse Design Implementation;
- Data Warehouse Usage, Support and Enhancement.

The Data Warehouse Planning phase comprises of five activities, which account for 60% of the data warehouse development effort. These activities are concerned with the identification of users, data requirements, data sources, tools to access and load the data, hardware, software, the business problems that will be addressed and the criteria to measure the success of the data warehouse.



Business (req.) Discovery
Data Warehousing Consulting
Data Warehousing Information Discovery
DW Logical Data Modelling
DW Architecture Design
Data Warehousing Solution Readiness
Knowledge Discovery Model Dev.
C/S App. Development (Full Cycle)
DW Physical DB Design
DW Data Transformation
Data Warehouse Management (Process and Operations)
Data Mining & Analytical Applications
Data Warehouse Solution Integration
Enterprise System Support
DW Logical DB Review
DW Physical DB Review
DW Tuning
DW Capacity Planning
DW Audit
Data Warehousing Planning
Data Warehouse Design and Implementation
Data Warehouse Usage, Support and Enhancement

Figure 4.3 – The NCR Data Warehousing Method (O'Donnell et al 2002, p04).

The Data Warehouse Design and Implementation phase consists of six activities, which examine potential risks associated with the project, such as change management. These activities are also concerned with the creation, construction and testing of a physical database design and the data extraction and load processes. The activities also focus on the design/development of applications to exploit the data and end-user training. The Data Warehouse Usage, Support and Enhancement phase consists of eight activities, which are concerned with maintenance and support of the data warehouse and reviewing aspects of the data warehouse, such as Return on Investment (ROI), the review activities provide information that may be useful in the next iteration. (O'Donnell *et al.* 2002, p03).

Figure 4.4 assesses the NCR Data Warehousing Method against the criteria cited earlier.

- Who is the vendor? **NCR**
- Who is the author? **NCR consultants and William H Inmon**
- Is it produced by a consulting organisation? **NCR**
- Is it proprietary (software specific)? **NCR products (Teradata Warehouse)**

Does the methodology include analysis, design and construction and or usage of the following:

Source Data	Makes provisions for all three areas
Extract Components	Makes provisions for all three areas
Transform Components	Makes provisions for all three areas
Load Components	Makes provisions for all three areas
Data Warehouse (or Data Mart)	Makes provisions for all three areas
Access Components	Makes provisions for all three areas
Metadata	Makes provisions for all three areas
Data Cleansing Components	Makes provisions for all three areas
Data Archiving Components	Makes provisions for all three areas

Figure 4.4 – An Assessment of the NCR Data Warehousing Method (O'Donnell et al. 2002).

4.2.3.2 Business Dimensional Lifecycle Diagram

The Kimball Method, figure 4.5, is a non-software specific method that was developed by the DW consultant Ralph Kimball. Kimball is an advocate of independent data marts, dimensional modelling and the use of bottom-up techniques in the development of a data warehouse. The Kimball Method has two initial activities, Planning & Growth and Business Requirements Definition. The outcomes of these stages feed directly into three different phases, which can be categorised as Architectural, Data Modelling and Analytic Application. The Architectural phase has two activities that are concerned with determining the physical architecture of the data warehouse and the selecting/installing of software and hardware (O'Donnell *et al.* 2002).

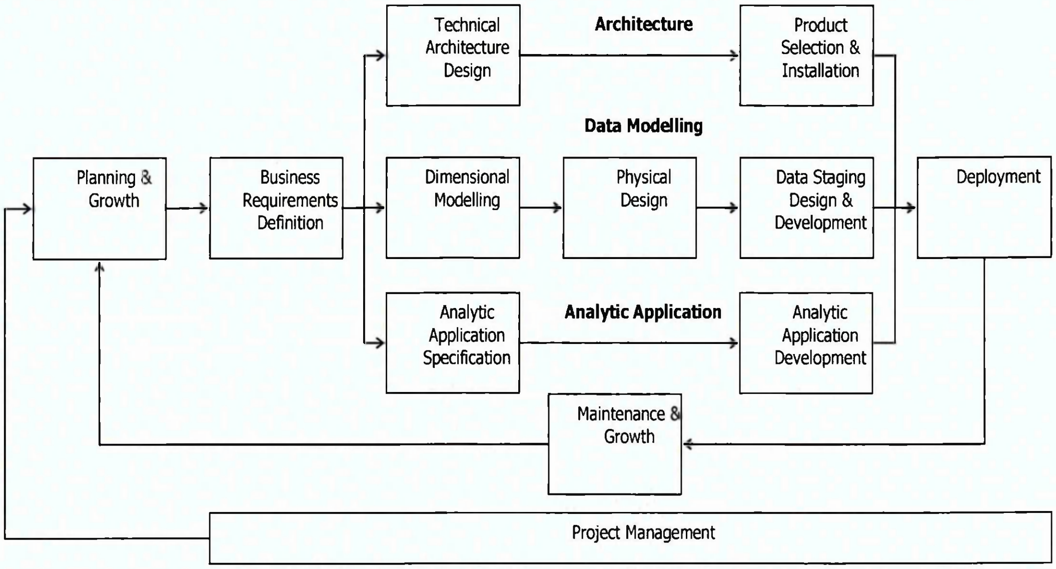
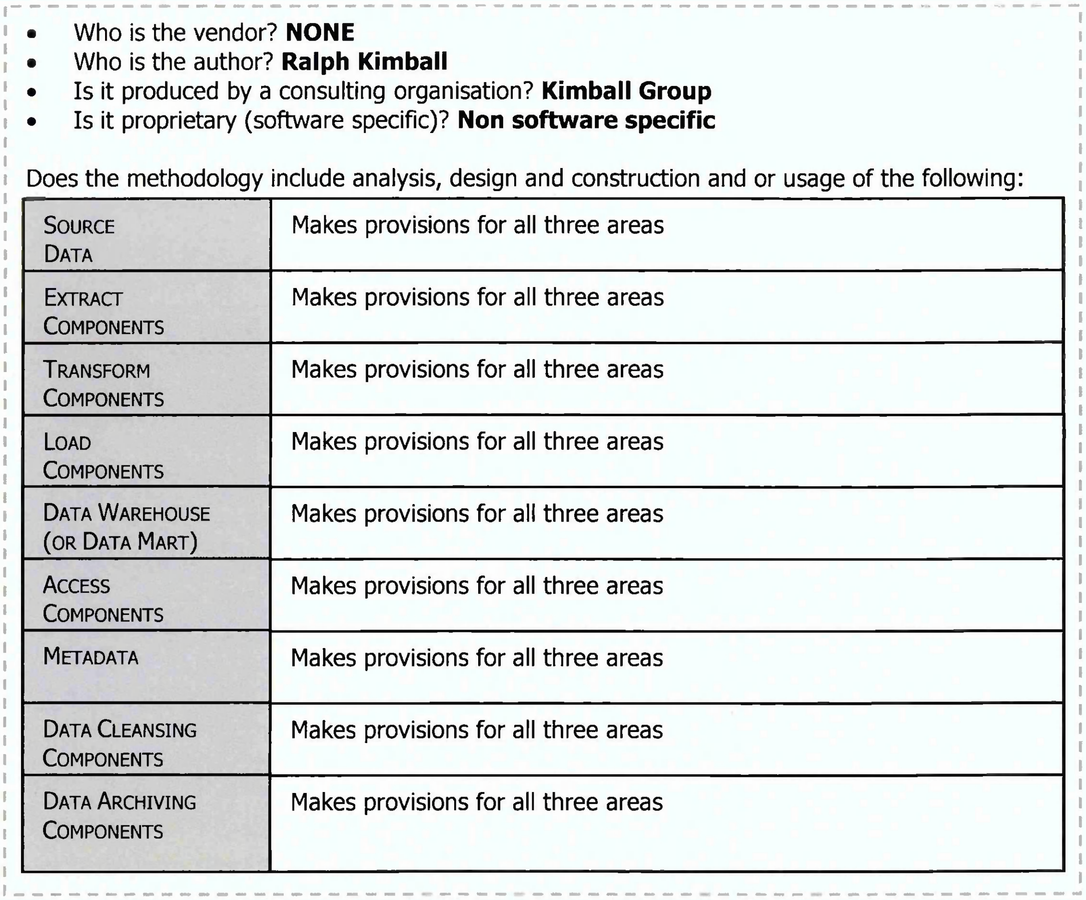

Technical Architecture Design
Architecture
Product Selection & Installation
Data Modelling
Planning & Growth
Business Requirements Definition
Dimensional Modelling
Physical Design
Data Staging Design & Development
Deployment
Analytic Application Specification
Analytic Application
Analytic Application Development
Maintenance & Growth
Project Management

Figure 4.5 – Business Dimensional Lifecycle Diagram (Kimball and Ross 2002, p332).

The Data Modelling phase has three activities that are concerned with the process of designing star schemas, indexes, aggregate tables and the data load process - staging area. The Analytic Application phase has two activities which focus on designing/constructing the application used to access the data warehouse - the OLAP (On-Line Analytical Processing) front-end. The Data Modelling phase is dependent on the completion of the Architectural phase and the Analytic Application phase is dependent on the completion of the Data Modelling phase. Once all stages have been completed the system is then Deployed. The Kimball method is an iterative process which loops back to the initial stages for Maintenance and Growth of the system (O'Donnell *et al.* 2002).

Figure 4.6, below, assesses the BDLD in regards to the criteria cited previously.



- Who is the vendor? **NONE**
- Who is the author? **Ralph Kimball**
- Is it produced by a consulting organisation? **Kimball Group**
- Is it proprietary (software specific)? **Non software specific**

Does the methodology include analysis, design and construction and or usage of the following:

SOURCE DATA	Makes provisions for all three areas
EXTRACT COMPONENTS	Makes provisions for all three areas
TRANSFORM COMPONENTS	Makes provisions for all three areas
LOAD COMPONENTS	Makes provisions for all three areas
DATA WAREHOUSE (OR DATA MART)	Makes provisions for all three areas
ACCESS COMPONENTS	Makes provisions for all three areas
METADATA	Makes provisions for all three areas
DATA CLEANSING COMPONENTS	Makes provisions for all three areas
DATA ARCHIVING COMPONENTS	Makes provisions for all three areas

Figure 4.6 – An Assessment of the Business Dimensional Lifecycle Diagram (O'Donnell et al. 2002).

4.2.3.3 SAS Rapid Data Warehousing Methodology

The SAS® Rapid Data Warehousing methodology is intended to be a top-down approach, but can also be used in a bottom-up context. Indeed, the methodology is an iterative approach that can be used to develop incremental data marts where organisations are unprepared to invest immediately in a large scale EDW. SAS® point out that such an incremental approach helps to deliver higher business value, a quick ROI and minimises the risks associated with project failure. Therefore, the methodology will be discussed as an example of a hybrid approach (SAS Institute Inc 2001).

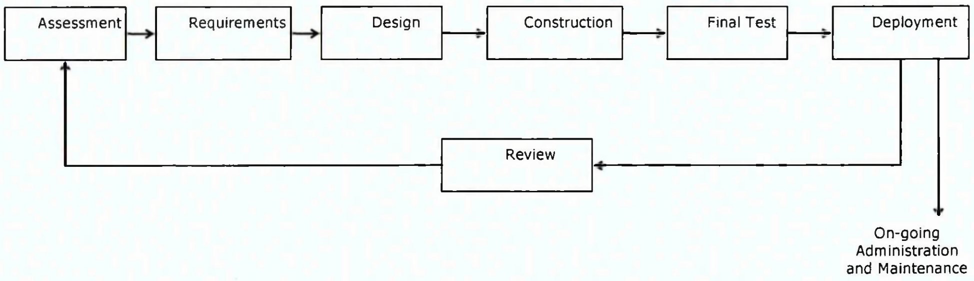


Figure 4.7 – SAS Rapid Data Warehousing Methodology (SAS Institute Inc 2001, p11).

The SAS® methodology breaks down the project into a set of builds, the build cycle consists of a number of phases called assessment, requirements, design, construction, final test and deployment. Unlike the majority of vendors SAS® provide the full end-to-end package with their Warehouse Administrator, Enterprise Guide software and their Intelligence Architecture Blueprint (SAS Institute Inc 2001). Figure 4.8, below, assesses the SAS® Rapid Data Warehousing methodology in regards to the criteria cited previously

- Who is the vendor? **SAS**
- Who is the author? **SAS**
- Is it produced by a consulting organisation? **SAS**
- Is it proprietary (software specific)? **Intended to be SAS software specific**

Does the methodology include analysis, design and construction and or usage of the following:

Source Data	Makes provisions for all three areas
Extract Components	Makes provisions for all three areas
Transform Components	Makes provisions for all three areas
Load Components	Makes provisions for all three areas
Data Warehouse (or Data Mart)	Makes provisions for all three areas
Access Components	Makes provisions for all three areas
Metadata	Makes provisions for all three areas
Data Cleansing Components	Makes provisions for all three areas
Data Archiving Components	Makes provisions for all three areas

Figure 4.8 – An Assessment of the SAS Rapid Data Warehousing Methodology (SAS Institute Inc 2001).

Having introduced the different Data Warehousing approaches, above, a decision will be made, in Chapter 5, as to which approach will be used to build the three DM marts.

4.3 DATA MINING

The common areas of DM are depicted in figure 4.9 below.

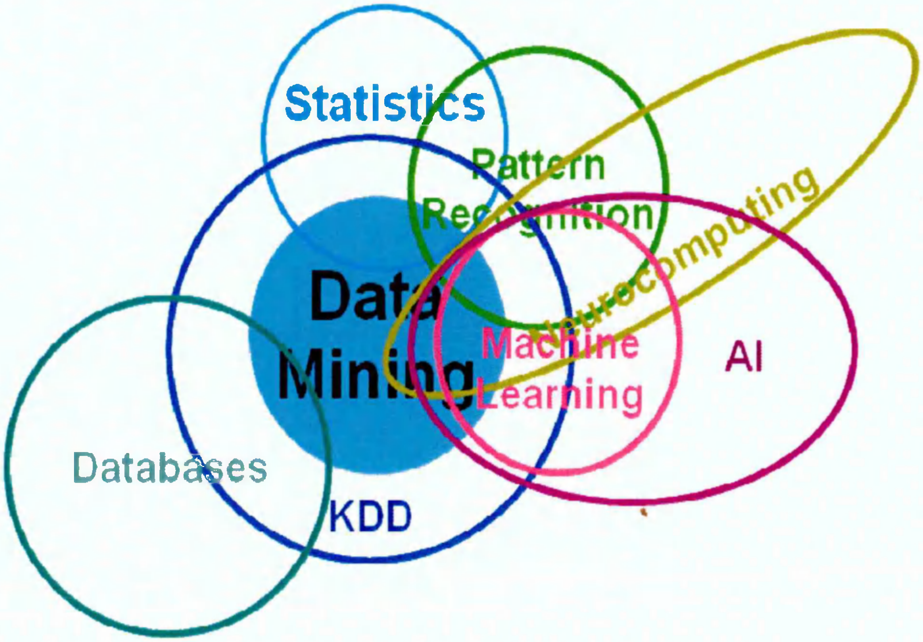


Figure 4.9 – Where Data Mining Fits In (Burley 2003)

DM is a small part of a process called Knowledge Discovery, which aids organisations in the discovery of patterns and relationships hidden within their data (Berry and Linoff 2011). Arguably it is important to understand the different DM techniques and methodologies that could be used within this research. What follows is a discussion of each of these in detail.

4.3.1 DATA MINING TECHNIQUES

A DM technique is a statistical method that becomes a DM model when coded; there can be several different models for the same technique. For example the DM technique called Rule Association (see below) has at least two DM models associated with it, Generalised Rule Induction (GRI) and Apriori (Berson *et al.* 1999). This sub-section will discuss supervised and unsupervised learning, and a number of the most prominent DM techniques, such as Clustering, Rule Induction, Decision Trees, Neural Networks, Market Basket Analysis and Genetic Algorithms.

4.3.2 Supervised V's Unsupervised Learning

Before discussing the DM techniques in detail it is imperative to define the terms supervised and unsupervised learning. Berry and Linoff (2011) define the two as follows:

> *"[...] Directed data mining [(supervised learning)] focuses on one or more variables that are targets [(output variable)], and the historical data contains examples of all target values. In other words directed data mining does not look for just patterns in the data, but for patterns that explain the target values. [...] In undirected data mining [(unsupervised learning)], there are no special roles. The goal is to find overall patterns. After patterns have been detected, it is the responsibility of a person to interpret them and decide whether they are useful."* (Berry and Linoff 2011, p81).

In supervised learning the output variable is specified before creation and in unsupervised learning the output is determined by the model. Furthermore, it is important to note that an unsupervised learning technique can be used as a precedent to a supervised learning technique when the explorer is unsure of what to look for within the data (Berson *et al* 1999).

4.3.2.1 Clustering

Clustering is the most common form of an unsupervised learning technique, it provides a high level view of what is happening within a database. Clustering works by consolidating data into high level views and grouping similar records together in a database on the basis of self-similarity. There are two main types of clustering techniques - hierarchical and non-hierarchical. There are numerous algorithms associated with clustering, one of which is Kohonen Networks, which determines clusters by using the 'Nearest Neighbour' technique. Clustering offers no explanation as to why data is grouped into a certain cluster, thus the results are generally interpreted by someone who has a knowledge and understanding of the business. Clustering is generally used as a precedent to other supervised learning techniques, such as Neural Networks (Berson *et al.* 1999).

4.3.2.2 Rule Induction

Rule Induction is an unsupervised learning technique, which identifies all possible patterns in a database. The technique gives an indication of how accurate and significant the patterns identified are and whether they are likely to occur again, it generates a number of relatively simple rules. These rules are ordered on the basis of how many times they apply and the percentage of times they are correct. One of the biggest issues with Rule Induction is that the number of rules generated can be overwhelming. Rule Induction has two DM models associated with it, GRI and Apriori, and it forms the basis of the supervised learning technique called Market Basket Analysis (Berson *et al.* 1999).

4.3.2.3 Decision Tree Analysis

Decision Tree Analysis is a supervised learning technique that can be used to either explore data or make predictions based on the data. Decisions Trees can predict both categorical and continuous variables and can be applied to both numeric and non-numeric data. They require very little data cleansing and pre-processing and their models are easily understandable, as Decision Trees consider a null value as a possible value along with their own branches and their rules can be easily translated into English and/or SQL (Structured Query Language). Decision Tree Analysis can be very complicated and the accuracy of Decision Trees can also be misleading, as they can split data in unsympathetic ways. Furthermore, the binary algorithms that they use to split data only consider one predictor variable at a time and in a specific order, which limits the number of possible splitting rules to test. This makes the relationship between the predictor variable hard to detect (Berson *et al.* 1999, Two Crows Corporation 1999).

4.3.2.4 Neural Network

A Neural Network is a supervised learning technique that can be applied to a variety of different models to identify patterns, make highly accurate predictions and learn. They can be built for classification, prediction or regression and can handle both categorical and continuous variables. A Neural Network is not 100% accurate but is more successful with large amounts of variables and data, it requires plenty of data pre-processing and time to train, as all input and output values are numeric. Therefore, the results generated have to be translated by someone who has knowledge of the business. Furthermore, Neural Networks don’t give any justification as to why the solution is valid and are limited in ease of use and deployment. It is important to note that Neural Networks require constant updating of the training set as they may age and weaken as the business environment around them changes. A common type of Neural Network is the feed-forward back propagation network (Berson *et al.* 1999, Berry and Linoff 2011).

4.3.2.5 Market Basket Analysis

Market Basket Analysis (Affinity Grouping/Rule Association) is a supervised learning technique that uses unsupervised learning techniques, such as Rule Induction, as its basis. The most common use of Market Basket Analysis is in the analysis of transactional data to help determine associations (rules). It is generally used when organisations are unsure of what to look for (for example patterns) within their data. Indeed, Market Basket Analysis is often used in retail to help retailers determine what products the customers are purchasing together (complementary products). Market Basket Analysis is one of the more popular DM techniques as association rules are generated along with the results, which show how tangible the relationship between the products and services are. It is important to be aware of the fact that the usefulness of these rules generated by the technique could be questionable. Indeed, the rules generated can be categorised as one of the following, useful, trivial and inexplicable (Berry and Linoff 2011).

4.3.2.6 Genetic Algorithms

Genetic Algorithms are used to determine a basic answer to a problem that is then continually reassessed to determine the optimal answer, in that the accuracy of the answer is increased. Genetic Algorithms or Evolutionary Algorithms are a supervised learning technique that can be applied to both classification and optimisation problems. Genetic Algorithms are one of the least used techniques as DM generally focuses on classification and prediction problems, not optimisation. A common application of Genetic Algorithms is in the training of Neural Networks (Berry and Linoff 2011).

Having introduced the different DM techniques above a decision will be made, in Chapter 7, as to which techniques will be used at different points within the research.

4.4 Data Mining Methodologies

A DM methodology is concerned with the stages of the DM process and what needs to be completed within those stages. There are numerous DM methodologies that have been developed by consultants and vendors to sell their services and products. Arguably, two of the most prominent DM methodologies are the SAS® SEMMA (Sample, Explore, Modify, Model and Assess) methodology and the CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology, this sub-section will therefore discuss these two methodologies and in an attempt to select a suitable methodology will highlight their respective strengths and limitations.

4.4.1 SEMMA

The SEMMA methodology (figure 4.10) is an award winning DM methodology that was developed by SAS®. The methodology breaks down the process of DM into five primary phases (Sample, Explore, Modify, Model and Assess) (SAS no date).

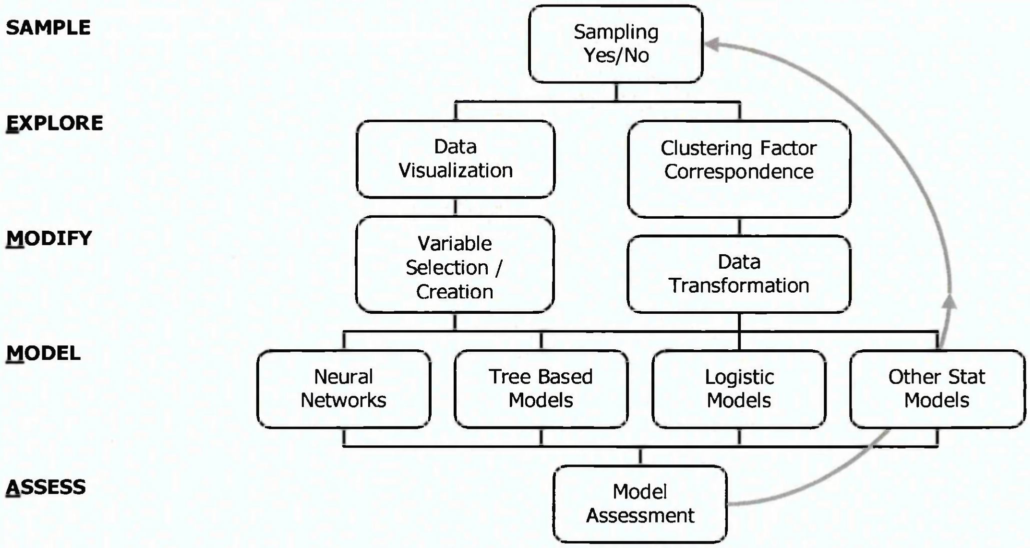


Figure 4.10 – The SAS SEMMA Methodology (SAS no date).

The Sample phase is concerned with the identification of data inputs, samples and data partitions, Explore focuses on exploring the data through graphs and statistics to determine such things as key variables. The Modify phase is where the data is transformed and prepared for analysis. Once any modifications have been carried out then a predictive Model, such a Decision Tree is fitted. The results of the modelling are then finally compared in the Assessment phase and the final model is then determined (SAS no date). The major strengths of the SEMMA methodology are its robustness and usability. Indeed, the SEMMA methodology is a tried and tested method that has won industry awards. It provides a set of logical and useful steps to aid the DM process. The main drawbacks of this method are that it mainly focuses on the analysis and interpretation of data and is SAS® specific (SAS no date).

4.4.2 CRISP-DM

CRISP-DM was developed in 1996 by DaimlerChrysler, SPSS® and NCR to create a standard process model for a growing DM market. CRISP-DM is a hierarchical process that consists of a number of tasks that are broken down into four levels of abstraction (phases, generic tasks, specialised tasks and process instances), which are then distilled into further tasks. Further to this, CRISP-DM breaks down the DM process into six phases (Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment), figure 4.11, with each phase focusing on different parts of the DM process (Chapman *et al.* 2000).

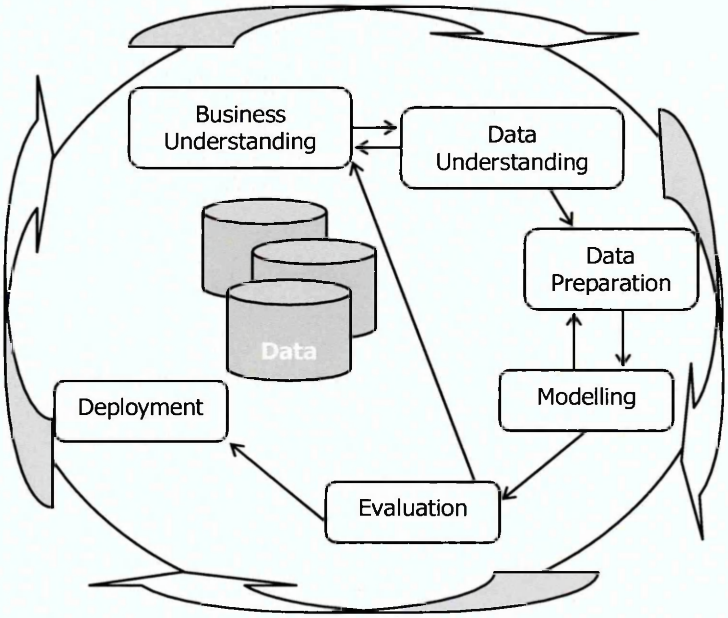


Figure 4.11 – CRISP-DM (Chapman et al. 2000).

The fundamental strengths of the CRISP-DM methodology are that it is non-platform specific as it was developed in conjunction with a number of organisations. In addition to this, it provides a structure for a developer to follow.

The method is limited by the fact that the process focuses heavily on the carrying out of tasks that have to be completed before any DM begins. The DM process is prolonged due to the amount of bureaucratic red tape (Chapman *et al.* 2000)

Having outlined the different DM methodologies above a decision will be made, in Chapter 5, as to which methodology will be used to guide the DM process.

4.5 What is New in this Research

Chapter 3 highlighted that the problem of student progression within HE is well documented and that the EDM research, in this area, has tended to focus upon assessing the applicability of DM and determining the best DM techniques. Therefore, this section will provide a brief overview as to how this research will use DW and DM to build a number of DM marts and create, using directed DM, a number of student profiles. This will conclude with a brief discussion around the appropriateness of applying quantitative methods to the problem of student progression in HE.

4.5.1 Building Data Mining Marts

McGivney (2003) suggest that there is a lack of central direction on the collection and recording of institutional information and that the accuracy of existing data is questionable, which in turn implies that the application of DW in HE is sparse. However, this research will use the DW process (outlined in section 4.2) to build a number of DM marts. It is anticipated that the data contained within the Student Information (SI) database is dirty and will require reducing and cleaning before it is analysed. This process will aid the DM stage as it will provide a more in-depth knowledge of the institution and its students. The DM marts developed here will be fundamental to the student profiling stage (see section 4.5.2 below) of the research, as they will provide a single source of transactional data for each of the models.

4.5.2 Student Profiling

One of the major differences between this research and Burley (2007) is the creation of a number of undergraduate student profiles. It is envisaged that three profiles will be created to predict:

- Award Classification
- Progression onto Postgraduate Studies at SHU
- Employment Type.

The award classification profile will generate a number of rules around students attaining a certain grade. The second profile, progression onto postgraduate studies at SHU, will create rules for undergraduate students who go onto take postgraduate at SHU. The third profile will generate rules around what type of job (such as graduate, non-graduate or unemployed) the student will obtain when they have completed their undergraduate course. It is perhaps important to point out that the predictive power, of the profiles developed, will reduce over time. Therefore, the models developed as part of this research will need to be revalidated from time to time. This is common practice in the finance and insurance sector where such models are used to predict customer behaviour.

This aspect of the research is markedly different to that of Burley's. In that it is introducing BI, through DW and DM, to the problem of behaviour (progression), which will allow for a number of predictions to be made.

4.5.3 Using Quantitative Methods to Predict Student Behaviour

A number of subject matter experts in this area disagree with using quantitative methods to predict student behaviour. Indeed, Yorke and Longden (2008) argue that quantitative results do not go far enough to understanding the problems and that a better understanding can be achieved through considering both quantitative data in conjunction with qualitative results – a mixed methods approach. This view is also implied by Moxley *et al.* (2001) who carry out qualitative research to help understand the problems associated with student progression. Furthermore, this is also echoed by McGiveny (2003), she argues that it is difficult to replicate results due to differences in how institutions measure progression and collect data, which is fundamental to the positivist stance associated with such quantitative methods. She goes on to state that results from such studies are only meaningful within the context of each individual institution or subject area. However, there is a place for quantitative methods and work carried out by Burley (2007), Luan (2006) and the University of Alabama (University Business 2004) have all demonstrated the value of such techniques within the student progression and HE domain.

4.6 Summary

This chapter sets out to provide an overview of the key concepts associated with DW and DM. It discusses two of the most prominent DW approaches, top-down and bottom-up, before going on to introduce a number of methodologies for carrying out such projects. It then goes on to discuss DM techniques and methodologies that will be useful when mining the student data in Chapter 7. It is important to note that a decision will be made in subsequent chapters (5 through to 7) as to which approach, methodology, and techniques will be applied during the DW and DM phases of this research. In this chapter, a number of DQ issues are also introduced which will be useful in Chapter 6. Further to this, the chapter outlines the differences between this and other research in this area, which centres upon the use of BI tools and the quantitative nature of the research.

5 Research Approach

The aim of this chapter is to discuss the research approach that is to be adopted during the project.

Research is:

> *"the process of arriving at dependable solutions to problems through the planned and systematic collection, analysis, and interpretation of data. It is a most important tool for advancing knowledge, for promoting progress, and for enabling man [sic] to relate more effectively to his environment, to accomplish his purpose, and to resolve his conflicts" (Cohen et al. 2000, p45).*

This definition highlights that research is an integral part of developing solutions to a given problem. According to Bryman (2012) such research should be considered from three perspectives, figure 5.1.

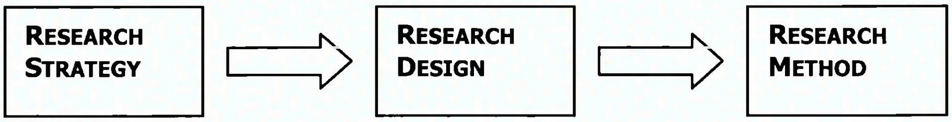


Figure 5.1 – The Fundamentals of Social Research (Bryman 2012).

The objective of the research is to *investigate large academic data sets to build predictive models of student behaviour*. The three perspectives of research, identified in figure 5.1, will be discussed in regards to achieving this objective.

5.1 Research Strategy

Cohen *et al.* (2011) attest that research is not simply a technical exercise, it is concerned with understanding the external social world. Central to this definition are the opinions of Hitchcock and Hughes (1995) who identify four fundamental concepts of research, these are summarised in figure 5.2.

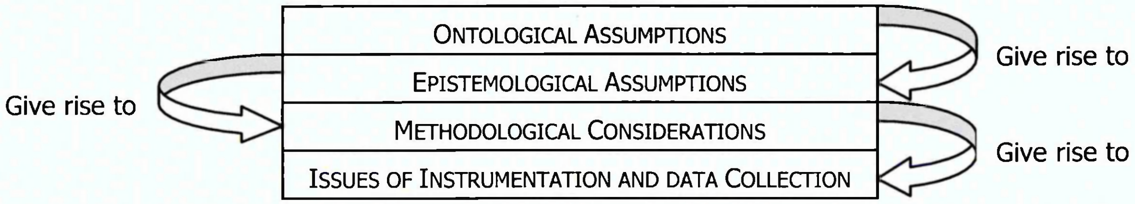

ONTOLOGICAL ASSUMPTIONS
Give rise to
EPISTEMOLOGICAL ASSUMPTIONS
Give rise to
METHODOLOGICAL CONSIDERATIONS
Give rise to
ISSUES OF INSTRUMENTATION AND DATA COLLECTION

Figure 5.2 – The Notion of Hitchcock and Hughes in Cohen et al. (2011, p03).

This notion is simplified by Bryman (2012) who highlights that there are a variety of considerations that need to be taken into account. These considerations are in relation to:

a. the type of research (quantitative or qualitative);
b. the relationship between theory and research (deductive or inductive);
c. epistemology (positivism or interpretivism); and
d. ontology (objectivism or constructionism).

There are two types of research, quantitative research and qualitative research and there is no hard and fast distinction between the two. However, it is possible to make a general distinction between them (Cohen *et al.* 2011). Indeed, quantitative research is:

> *"Where one subscribes to the view which treats the social world like the natural world - as if it were a hard, external and objective reality - then scientific investigation will be directed at analysing the relationships and regularities between selected factors in that world." (Cohen et al. 2011, p06).*

This definition suggests that there is a direct parallel between natural science and quantitative research, as quantitative research is predominately concerned with the measuring of facts. Furthermore, quantitative research is more commonly associated with the macro environment as the research population tends to be quite large and varied. Therefore, the emphasis in quantitative research tends to focus upon quantification in data collection and analysis (Byman 2012).

Conversely, in qualitative research:

> *"[...] one favours the alternative view of social reality which stresses the importance of the subjective experience of individuals in the creation of the social world [...] The principle concern is with an understanding of the way in which individuals create, modify and interpret the world in which they find themselves." (Cohen et al. 2011, p06).*

This would imply that social research is not an exact science and that human interactions and the external environment are fundamental. Qualitative research is usually focused on the micro environment as the research environment tends to be quite specific and usually places an emphasis on words (Byman 2012).

The relationship between theory and research can be considered as either deductive or inductive. In general a deductive approach is where hypotheses are deduced from theories and then used to drive the data collection process. The type of research conducted in the deductive approach is usually quantitative and is normally associated with the macro environment. The inductive approach is in sharp contrast to the deductive approach, as the research findings are generally fed back into the theory after the research has been conducted. Qualitative research is typically associated with an inductive research approach and is usually focused on the micro environment (Bryman 2012).

Given the aim and objectives of the research, Chapter 2 and having reviewed the literature in Chapter 3 it is envisaged that *the research will be mainly quantitative*. This is due to the fact that research focused on using statistical analysis and DM to determine patterns and trends in HE data sets, specifically SHU. It is envisaged that this will enable the identification of general patterns of student behaviour to be established, which could then be applied to future students. These patterns of behaviour relate to determining final award classification, progression onto postgraduate studies at SHU and employment type post undergraduate degree completion. Moreover, this type of research is better suited to the measuring of facts in a large and diverse population (*the macro environment*), such as a sample of students taken from across all SHU faculties.

Generally, the relationship between research and theory in quantitative research tends to be deductive. However, DM can be both inductive (unsupervised) or deductive (supervised) this is usually dependent upon which DM technique one wishes to apply. Arguably, a decision regarding which DM technique to use cannot be made until the data is better understood. Therefore a decision will be made in Chapter 7 as to which DM techniques will be applied to the data. Furthermore, it is envisaged that the results determined will be considered against the literature introduced in Chapter 3 – see Chapter 9, as this will help to substantiate the findings. Therefore, it is possible that *the research will have both inductive and deductive elements*.

Epistemology is concerned with "what is (or should be) regarded as acceptable knowledge in a discipline." (Bryman 2012, p27). When undertaking research there are two main epistemological positions, positivism and interpretivism. Positivism is associated with the nineteenth-century French philosopher August Comte and is concerned with the application of a scientific model to study the social world (Cohen *et al.* 2011). The fundamental idea of positivism is:

> *"[...] that the social world exists externally, and that its properties should be measured through objective methods, rather than being inferred subjectively through sensation, reflection or intuition" (Easterby-Smith et al. 2012, p22).*

The quantitative research type discussed earlier tends to be associated with positivism as it incorporates the practices and norms of the natural scientific model. In a positivist approach the transferability of the research is fundamental, as skills and knowledge acquired from one environment may be transferred to another (Bryman 2012). Interpretivism is in contrast to positivism. The key idea of interpretivism is that people and their intuitions are significantly different to that of natural sciences and any social world study should reflect this individuality (Cohen *et al.* 2011). Interpretivism is a confluence of a number of traditions mainly phenomenology, symbolic interactionism and ethnomethodology - these traditions are defined in the glossary (Cohen *et al.* 2011). Qualitative research tends to be associated with interpretivism as it allows the social world to be emphasised and interpreted. In an interpretivist approach the research is specific to a single environment and the skills and knowledge acquired are likely to be non-transferable to other environments (Bryman 2012).

The comments of Gill and Johnson (2010) were taken into consideration in determining the epistemological position of the research. They state that the most effective approach for resolving "[...] a given research question depends on a large number of variables, not least the problem itself." (Gill and Johnson 2010, p06). The epistemological position of *positivism* was thought applicable as the research is trying to establish general patterns of student behaviour and build models that can be transferred to other academic years and external organisations (other institutions).

Ontological assumptions are fundamental to the way in which research is carried out and the way research questions are formulated. Research ontology is concerned with investigating the nature or essence of social phenomena, there are two main positions, objectivism and constructionism (Bryman 2012).

Objectivism is the influencing of people through social phenomena that is beyond their reach or control; it is simply those forces that act on and inhibit people. The organisation is a good illustration of this as it has a number of predefined procedures, rules and regulations, which are stringently learned, applied and adhered to by its people. Objective research emphasises the formal properties of an organisation or the values and beliefs of members of a culture (Bryman 2012). Constructionism challenges objectivism by arguing that order in organisations is worked at and is not a pre-existing characteristic. In the case of the organisation procedures, rules and regulations are a lot more fluid as they are influenced, adapted and created by people to suit their needs as well as the organisations. In constructionist research emphasis is placed upon the involvement of people in the construction of reality (Bryman 2012).

Therefore, on the basis of the above discussion regarding ontology and the review of the literature in Chapter 3, the ontological position of *objectivism* was deemed to be more appropriate, as the review of the literature in section 3.2.3.1 highlighted that the institution has an influence on its students in terms of its policies and procedures. Such policies and procedures are usually defined by the institution and/or external HE organisations (such as HEFCE). These policies and procedures are usually less fluid and are developed to be adhered to by university staff and students. Indeed, from the student's perspective, such policies and procedures are usually set out and agreed at enrolment. Furthermore, this research places a large emphasis on generalising the behaviour of members of the student culture.

Finally, a research strategy can be roughly defined by the type of research, quantitative or qualitative. Therefore, on the basis of the discussions presented above, regarding the four considerations, a *quantitative research strategy* was thought to be the best approach in achieving the research objectives – a more detailed description of the research sequence can be found in section 5.4.

5.2 Research Design

> *"Research designs are plans and procedures for research that span the decisions from broad assumptions to detailed methods of data collection and analysis. [...] The selection of the research design is [...] based on the nature of the research problem or issue being addressed, the researchers' personal experiences, and the audiences of the study." (Creswell 2009, p03).*

Bryman (2012) suggests that there are three essential design criteria to selecting a suitable research design - these are reliability, replicability and validity. In an attempt to satisfy these criteria a *Cross-Sectional design was identified as being the most appropriate research design*. Indeed a Cross-Sectional design is concerned with the collection of data on more than one case at a single point in time (Cohen *et al.* 2011). It is envisaged that numerous exploratory variables will have to be examined for patterns and trends prior to the construction of the predictive models.

In keeping with the research design and to reduce the potential problems associated with reliability, it is envisaged that a large sample of data will be obtained from a cohort from across all SHU faculties at a single point in time. There are no reasons to suppose that this cohort would be different from any other similar time period. Furthermore, the process followed in obtaining data and building the DM mart and models from a snapshot of SHU data could easily be replicated for other years and HE institutions. It is thought that the validity of the research will be affirmed by comparing the results to previous findings introduced in Chapter 3.

5.3 Research Method

According to Bryman (2012)

> *"A research method is simply a technique for collecting data. It can involve a specific instrument such as a self-completion questionnaire or a structured interview schedule, or participation observation whereby the research listens to and watches others." (Bryman 2012, p46).*

This definition implies that the research method is more qualitative in nature. Therefore, in terms of method associated with this research, this section will provide a descriptive framework as to how the research will be organised before going on to outline the chosen DW and DM approaches and methodologies that will be followed in Chapters 6 and 7 respectively.

5.3.1 Organising the Research

Whilst a descriptive framework is more commonly associated with a Case Study research design, it is useful in this context as it will help to determine the flow of the research, figure 5.3.

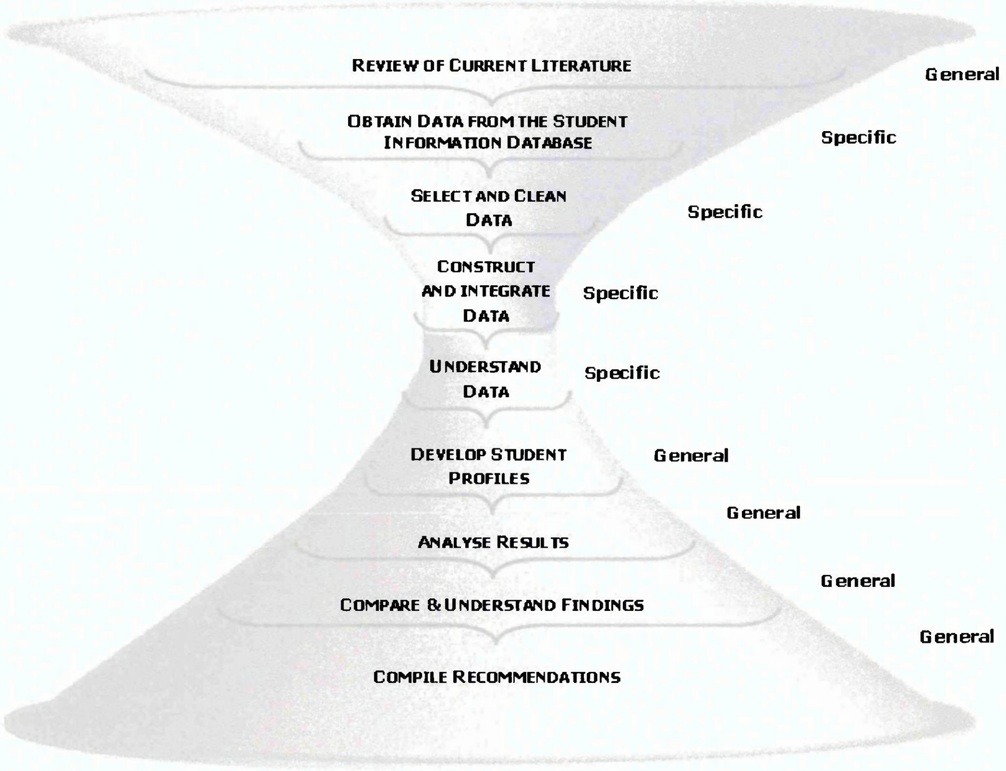


Figure 5.3 – Descriptive Framework for Organising the Research (developed by the author)

The research initiates with an extensive review of the literature, this will be carried out for two reasons. First to gain an understanding of the problems within the domain and secondly to understand the previous research that has already been conducted.

As this research is concerned with predicting student behaviour to build three student profiles (award classification, progression onto postgraduate studies at SHU and employment type), a number of meetings will be setup (see Appendix III) so that data can be obtained from the SI database.

Prior to the construction and integration of the SI data into the three DM marts the data will be selected and cleaned using the ETL (Extract, Transform and Load) process associated with DW, outlined in section 4.2.

The data within each individual DM mart will then be explored for patterns and trends in relation to each target variable. It is hoped that this understanding will help to decide upon how best to modify the data before it is modelled.

Through following an iterative process suitable DM techniques will be applied and at the modelling stage, which will then later be assessed. This will then generate a number of rules that can be used to predict student award classification, progression onto postgraduate studies at SHU and employment type.

Having analysed the results and compiled the findings a number of recommendations will be made in Chapter 9.

5.3.2 Building the Data Mining Marts

Given the definition of a DM mart in Chapter 4 and the fact that the research is concerned with building three DM marts, not an entire data warehouse, *a bottom-up approach is believed to be more suitable*.

In determining the most appropriate DW methodology (see section 4.2.3) to elect, the choice is somewhat narrowed down by selecting a bottom-up approach above. Indeed, this results in a choice between the BDLD and the SAS® Rapid Data Warehousing methodology. Ultimately, *the Business Dimensional Lifecycle Diagram was favoured* over the SAS® Rapid Data Warehousing methodology as the SAS® approach is intended to be a top-down approach that can be used in a bottom-up context. Finally, it is important to remember that the selection of a data warehousing methodology will not ensure the success of the project, as "methodologies [...] aren't magic there more like recipes [and] just because you have the cookbook doesn't mean you'll be a great chef (Watternson 1998, p63).

5.3.3 Undertaking the Data Mining

A number of DM methodologies, including their advantages and limitations, have already been outlined in section 4.4. Ultimately, the choice of methodology is determined by the software used to carry out the DM. Therefore, as the research intends to use the SAS® Enterprise Miner the SAS® SEMMA methodology will be followed during the DM phase of the research.

5.4 Research Sequence

The research execution will be broken down into seven steps as suggested by Gill and Johnson (2010), figure 5.4, theses seven steps will therefore be discussed in relation to the research.

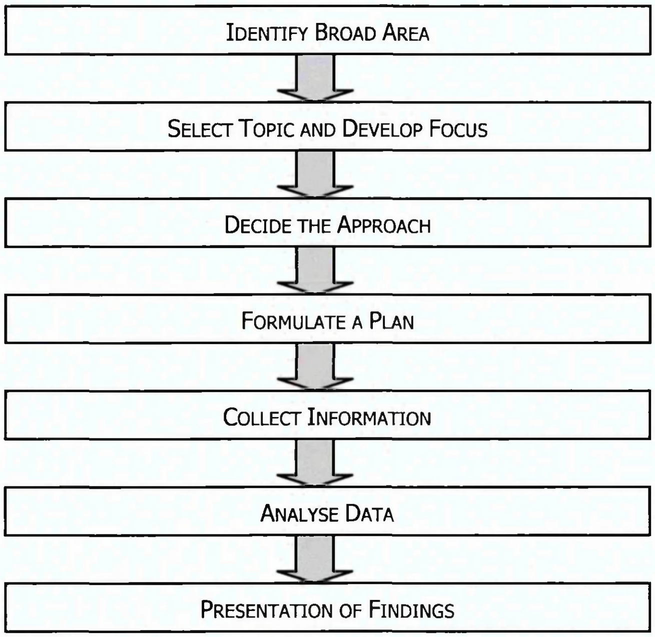


Figure 5.4 – The Research Sequence (Gill and Johnson 2010, p09).

The broad area of BI was selected as the area of investigation, as the author was keen to broaden his skills and knowledge within a new area.

In selecting the topic for the research a number of areas were considered. However, the problem of student progression was identified as an area where BI tools and techniques hadn't been utilised to any great extent (Burley 2006).

It was decided that for the manageability of the project that the research would be limited to SHU.

A research plan was formulated through developing a framework for organising the research, see figure 5.3.

Data for quantitative analysis will be gathered through the application of DW and DM techniques on the SI database. Further, information regarding the collection of data can be found in the research design and method sections.

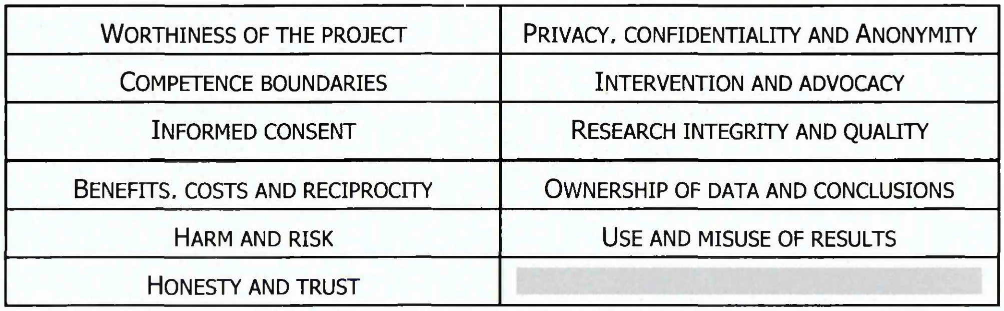
The results of the research will then be used to construct a number of user profiles to predict student behaviour and to compile a number of recommendations regarding student progression.

5.5 Research Ethics

Cohen *et al.* (2011) attest that:

> *"Ethical issues may stem from the kinds of problems investigated by social scientists and the methods they use to obtain valid and reliable data. This means that each stage in the research sequence raises ethical issues"* (Cohen *et al.* 2011, p76).

Indeed, a major ethical concern of this research is student and institutional confidentially. As a result any demographic data, which could help to identify actual students will be removed. Additionally, the ethical issues identified by Miles and Huberman (1994) will also be taken into consideration and are shown in figure 5.5.



Worthiness of the project	Privacy, confidentiality and Anonymity
Competence boundaries	Intervention and advocacy
Informed consent	Research integrity and quality
Benefits, costs and reciprocity	Ownership of data and conclusions
Harm and risk	Use and misuse of results
Honesty and trust	

Figure 5.5 – Ethical Issues in Research (Miles and Huberman 1994, p290-295).

Each of the issues identified in figure 5.5 will be considered in relation to the research.

Worthiness of the Project - the primary objective of the research is to add to the knowledge in the domain of student progression in HE, through using BI tools and techniques. Student progression has continued to improve at SHU, student progression increased by 1.1% from 91.2 in 2001-02 to 92.3% in 2004-05 (National Audit Office 2007). Therefore, it is believed that this research will add value to the institution as it will help them to understand and continue to improve their student progression and marketing. This will allow the institution to develop more flexible approaches to improving student progression and reducing the financial pressures associated with students failing to progress. Academically, it is expected that the research will contribute to the current literature and allow for further research to be conducted using BI tools and techniques. Indeed, this has been the primary motivation for doing this research.

Competence Boundaries – The author has experience of carrying out both qualitative and quantitative research. This was gained whilst studying for a Degree and a Master's Degree and through working in the NHS (National Health Service) as a statistician and whilst working as a researcher at SHU.

Informed Consent – All parties involved will be made aware of what the research requires from them, any concerns raised over confidentiality will be addressed.

Benefits, Costs and Reciprocity – It is envisaged that the research could be beneficial to a number of stakeholders. SHU has the potential to gain from the results of this research. Indeed, the minimum the institution will gain is an insight into student progression. The students could also benefit from the research as the institution may implement some of the recommendations made (see chapter 9), which could potentially improve their progression rates. In addition to this, the research will also benefit new researchers as it will demonstrate how BI can provide a better understanding of student progression. Amongst receiving a Doctorate, the author will also benefit through expanding his knowledge into new areas. There is a considerable amount of investment being made by the author in terms of time and cost. In order to better manage the cost aspect funding for the research was received from SAS®. However, the direction of the research was not influenced by SAS®.

Harm and Risk – There is a risk to the institution involved as the findings could reflect badly upon them, as they will eventually be published in journals.

Honesty and Trust – It is important that the trust placed in the author, by the institutions is maintained.

Privacy, Confidentiality and Anonymity – It is possible that the students could be identified from their data records. Therefore, it is important that all data records are made anonymous and that only the author can identify the individuals, if any further follow up work is required. Again consideration needs to be made about the institutions and how the findings will be reflected.

Intervention and Advocacy – Any confidential information gained will be treated appropriately, such as the removal of student names *et cetera*.

Research Integrity and Quality – A conscientious attempt will be made to ensure that the research is conducted thoughtfully, correctly and carefully. The integrity and quality of the research will be reinforced through a comparison to the literature on student progression and EDM.

Ownership of the Data and Conclusions – The data belongs to SHU and the project and findings belong to the author. However, as sponsorship was achieved then SAS® will also have a vested interest in the publication of journals from the research.

Use and Misuse of Results – It is important that the effect upon the institution, students and SAS® are considered before publishing the research. This may result certain aspects of the research been made anonymous.

5.6 SUMMARY

This chapter sets out to discuss the research approach that will be followed during the research phase of the project. The research approach is considered from three perspectives (research strategy, research design and research method), which resulted in a quantitative approach and a cross-sectional research design being adopted. In keeping with the research strategy and design, the research methods were outlined in relation to the DW and DM processes. These methods included a bottom-up approach, the use of the BDLD for organising the DW and the SAS® SEMMA methodology for carrying out the DW. The DM techniques that will be applied during the later stages of the research were not determined in this chapter, as these will be selected whilst carrying out the modelling of the data in Chapter 7. The chapter concludes with a discussion on research ethics, which focuses mainly upon issues of student and institutional confidentiality.

6 Building and Understanding the Data Set

This chapter will discuss the building of three DM marts that will be used at the modelling stage to predict student award classification, progression onto postgraduate studies at SHU and employment type. The process followed, Kimball *et al.* (1998) BDLD (introduced in section 4.2.3.2), in building the DM marts will be discussed. In preparation for the DM stage (Chapter 7) the chapter will also present an understanding of the data retained within each DM mart. The chapter will conclude with a discussion around the number of observations and the event rates.

6.1 Building the Data Mining Marts

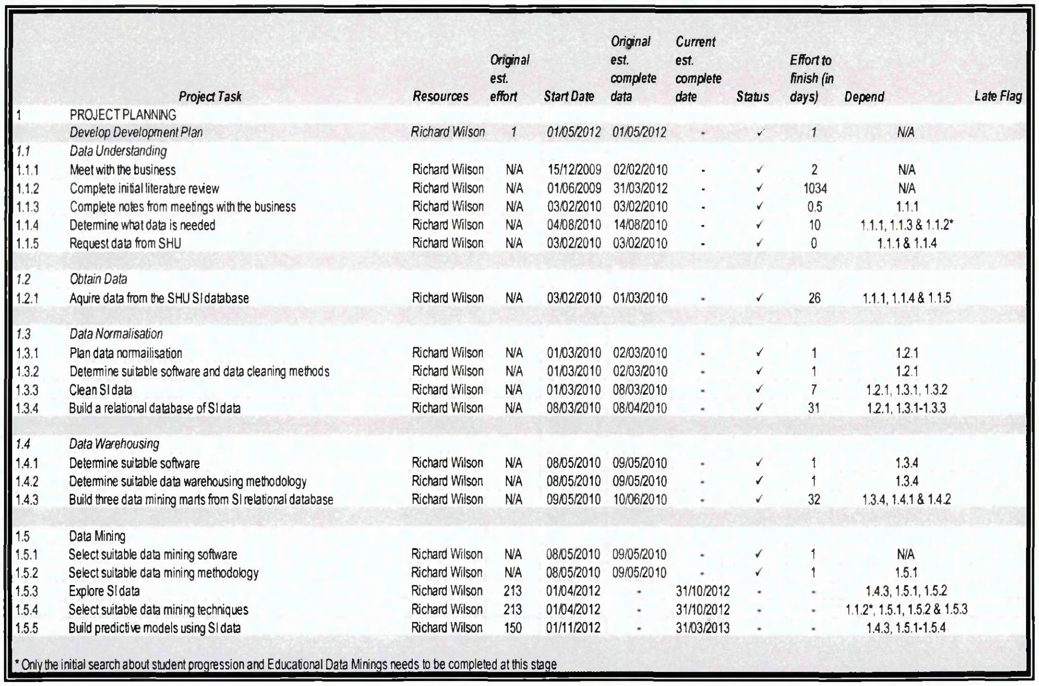
This section will discuss the following areas, in the BDLD, that are pertinent to the development of the three DM marts:

a) Project Planning and Management;
b) Business Requirements Definition; and
c) Data Modelling.

Other areas will not be discussed as these are associated with the development and deployment of a data warehouse.

6.1.1 Project Planning and Management

Kimball *et al.* (1998) point out that the most important part of this phase is to have a completed project plan, which plans the project from conception to birth. Arguably, the planning for this project will be small in comparison to the development of an entire data warehouse, as this project is only concerned with the development of three DM marts, which are reduced (only retaining the required data) and clean copies of the transactional data. Figure 6.1 gives details of the project plan for the development of the three DM marts, this has been adapted from Kimball *et al.* (1998, p75).



	Project Task	*Resources*	*Original est. effort*	*Start Date*	*Original est. complete data*	*Current est. complete date*	*Status*	*Effort to finish (in days)*	*Depend*	*Late Flag*
1	PROJECT PLANNING									
	Develop Development Plan	*Richard Wilson*	*1*	*01/05/2012*	*01/05/2012*	-	✓	*1*	*N/A*	
1.1	*Data Understanding*									
1.1.1	Meet with the business	Richard Wilson	N/A	15/12/2009	02/02/2010	-	✓	2	N/A	
1.1.2	Complete initial literature review	Richard Wilson	N/A	01/06/2009	31/03/2012	-	✓	1034	N/A	
1.1.3	Complete notes from meetings with the business	Richard Wilson	N/A	03/02/2010	03/02/2010	-	✓	0.5	1.1.1	
1.1.4	Determine what data is needed	Richard Wilson	N/A	04/08/2010	14/08/2010	-	✓	10	1.1.1, 1.1.3 & 1.1.2*	
1.1.5	Request data from SHU	Richard Wilson	N/A	03/02/2010	03/02/2010	-	✓	0	1.1.1 & 1.1.4	
1.2	*Obtain Data*									
1.2.1	Aquire data from the SHU SI database	Richard Wilson	N/A	03/02/2010	01/03/2010	-	✓	26	1.1.1, 1.1.4 & 1.1.5	
1.3	*Data Normalisation*									
1.3.1	Plan data normailisation	Richard Wilson	N/A	01/03/2010	02/03/2010	-	✓	1	1.2.1	
1.3.2	Determine suitable software and data cleaning methods	Richard Wilson	N/A	01/03/2010	02/03/2010	-	✓	1	1.2.1	
1.3.3	Clean SI data	Richard Wilson	N/A	01/03/2010	08/03/2010	-	✓	7	1.2.1, 1.3.1, 1.3.2	
1.3.4	Build a relational database of SI data	Richard Wilson	N/A	08/03/2010	08/04/2010	-	✓	31	1.2.1, 1.3.1-1.3.3	
1.4	*Data Warehousing*									
1.4.1	Determine suitable software	Richard Wilson	N/A	08/05/2010	09/05/2010	-	✓	1	1.3.4	
1.4.2	Determine suitable data warehousing methodology	Richard Wilson	N/A	08/05/2010	09/05/2010	-	✓	1	1.3.4	
1.4.3	Build three data mining marts from SI relational database	Richard Wilson	N/A	09/05/2010	10/06/2010	-	✓	32	1.3.4, 1.4.1 & 1.4.2	
1.5	Data Mining									
1.5.1	Select suitable data mining software	Richard Wilson	N/A	08/05/2010	09/05/2010	-	✓	1	N/A	
1.5.2	Select suitable data mining methodology	Richard Wilson	N/A	08/05/2010	09/05/2010	-	✓	1	1.5.1	
1.5.3	Explore SI data	Richard Wilson	213	01/04/2012	-	31/10/2012	-	-	1.4.3, 1.5.1, 1.5.2	
1.5.4	Select suitable data mining techniques	Richard Wilson	213	01/04/2012	-	31/10/2012	-	-	1.1.2*, 1.5.1, 1.5.2 & 1.5.3	
1.5.5	Build predictive models using SI data	Richard Wilson	150	01/11/2012	-	31/03/2013	-	-	1.4.3, 1.5.1-1.5.4	

* Only the initial search about student progression and Educational Data Minings needs to be completed at this stage

Figure 6.1 – Project Planning and Management.

6.1.2 Business Requirements Definition

The primary business of SHU is the education of its students. This research is concerned with understanding the educational process in terms of improving student progression. A large amount of work has already been completed into understanding the business and the issues around the progression of students at the university. This work includes the literature review and meetings with university staff members (Appendix III). In addition to this, the author also has a good understanding of the organisation and the HE environment from his experiences both as a student and as a member of staff.

6.1.3 Data Modelling

This section discusses the Data Modelling phase of the BDLD with regards to the development of the three DM marts. It is important to note that this section will only discuss the areas that are pertinent to building DM marts. Therefore, the Dimensional Modelling and Physical Design will be omitted from this section, as this is concerned with designing specific data marts for DW in relation to developing dimensions and fact tables to answer specific business questions. This section will outline the process followed in obtaining the data from the Student Information Services (SIS) department at SHU, discuss the transformation of the student data including the creation of the DM marts and it will conclude with an initial assessment of the data.

6.1.3.1 Background of the Data

The section outlines the process followed in requesting the data from the SIS department at SHU, The data is recorded by the university at student enrolment and after the student completes their course, which is input into the SI Database. The SI database holds all the student enrolment, course and employment data.

After consultation with the Information Analysts at SHU a request was made to the SIS department to obtain a large sample of data. Three datasets were requested for full-time undergraduate students, with a level 6 SCE (Student Course Enrolment) records, in 2006-07 and one data set relating to SHU postgraduate students for the academic years 2007/08, 2008/09 and 2009/10. The undergraduate datasets reflected the student's demographics, academic assessments and entry qualifications. Based on the meeting with the information department at SHU the following data items were requested:

Student ID number	Degree classification (class / rank)
Students faculty	Course name
SCE enrolment status	SCE result code
Entry qualification type	SCE end date
UCAS / tariff point **total**	Ethnic group
Age on entry to SHU	Gender
Home postcode	Fee status
Disability status (all categories)	Last institution / school attended
Reason for withdrawal (RFT on SCE)	Local authority
Learning contract (Yes/No)	Full final award title (SHU qualification obtained and award title)
Nationality	Final award date
Sponsor name (for funding)	Final mark *
Socio-economic group (post '02)	Level 6 average mark^{+}
SOC code	
** Final mark = mark which has determined students degree classification (SAWS screen).* $^{+}$*Level 6 average mark = mean mark of all modules undertaken at Level 6*	

Figure 6.2 – Dataset 1 (all registered full-time undergraduate level 6 students 06/07).

student id number	students course name
module name	students programme name
faculty which owns module	programme area which owns module
module credit rating	students overall module mark
module assessment pattern *	students overall module result $^{+}$
** module assessment pattern = i.e. is the module following Model A (requiring only an overall pass at 40%), or Model B (requiring 40% passes in each component)?* $^{+}$*needs to indicate whether the module is pass, refer, defer, compensated pass etc.*	

Figure 6.3 – Dataset 2 (all level 6 module taking records for all registered full-time undergraduate level 6 students, 2006/07).

student id number	qualification result (e.g. “Pass”)
entry qualification category	qualification tariff points awarded
qualification name (e.g. “A level”)	date qualification awarded (“Examination date”)
qualification title (e.g. “Maths”)	school / institution last attended
qualification grade (e.g. “B”)	students UCAS / tariff point total

Figure 6.4 – Dataset 3 (all entry qualifications for all registered full-time undergraduate level 6 students, 2006/07).

The following data items were then requested for the postgraduate data sets.

student id number	students faculty
academic year (e.g. 2007/08) *	students SCE enrolment status (for each year registered) *
students course name	
∗ *Some students may appear in more than one row, if they are registered in more than one academic year.*	

Figure 6.5 – Dataset 4 (all registered postgraduate students, 2007/08, 2008/09, 2009/10).

Through the meetings with SHU information department it transpired that the university was keen to include Undergraduate Destination of Leavers (DOL) as part of the research. After some careful consideration it was agreed to include DOL as part of the beyond undergraduate studies element of the project. This helped to provide an additional dimension to the research whilst also adding further value to the project for SHU.

The following data items, figure 6.6, were provided by SIS for the DOL table for undergraduate students that graduated between 2006 and 2008.

STU_CODE	QualObt	EMPNAM	B14_CAREER12	C21_PROFSOCT
COYEAR	QualObt_bin	B08_NHSORG	B14_CAREER6	C22_INSTPROV
CRS_NAME	FacultyAll	MAKEDO	B14_CAREER8	C23_SECINT1
Gender	Origin	SIC_rc	B14_CAREER9	C23_SECINT2
Age	Department	LOCEMP	B14_CAREER10	C23_SECINT3
Age_bin	A01_EMPCIR	B10_LOCEMP_r	JOBRES	C23_SECINT4
Age_coll	A02_MODSTUDY	B11_EMPSIZE	B16_PREVEMP	C23_SECINT5
Disability_full	DESTINATION	B12_QUALREQ	B16c_PREVEMP	C23_SECINT6
Disability_coll	A_HESACAT	B13_EMPIMP	B17_PREVCAT1	C23_SECINT7
Disability_bin	JOBTITLE	B14_CAREER7	B17_PREVCAT2	C23_SECINT9
Ethnicity_full	DUTIES	B14_CAREER1	B17_PREVCAT3	C24_FUNDSTDY
Ethnicity_coll	B04_GradJob	B14_CAREER3	B17_PREVCAT4	
Ethnicity_bin	B04_SocHE	B14_CAREER4	B17_PREVCAT5	
SI_Level	B04_Occupation	B14_CAREER5	B17_PREVCAT6	
Mode	B05_Contract	B14_CAREER11	C18_NATSTUDY	
ModeC	SALARY	B14_CAREER2	C19_TYPEQUAL	

Figure 6.6 – Dataset 5 (Destination of Leavers 2006 – 2008).

The data was extracted by the SI department, using a number of SQL queries, into a number of CSV files. A copy of these files was made and placed into a folder called raw data. This was done so that the data could be manipulated without contaminating the original source data. Should any problems occur in the cleansing process then it would be possible to start again with the original source data. Three areas were then setup in Microsoft Access called raw, staging and final. The raw data was where a copy of the SI data was stored, in Access, the staging area was where the data was processed before it was loaded into the DM tables, which was located in the final area.

6.1.3.2 Data Staging and Development

This sub-section outlines the process followed in developing the three DM marts, from the raw data descriptions introduced above, and provides some initial analysis into that data. This sub-section breaks down the data staging and development into five areas: selecting and cleaning the data; constructing and integrating the data; formatting the data; understanding the data; and data items and group.

6.1.3.2.1 SELECT AND CLEAN DATA

This section outlines the process followed in exporting, transforming and cleaning the data. A description of the types of data quality errors, identified during this process, can be found in section 4.2. The request for data returned five comma separated files that needed to be transformed into a suitable format for DM. The first part of the process involved removing repeated values in the data through normalisation. A diagram of the database was created to plan the development, figure 6.7.

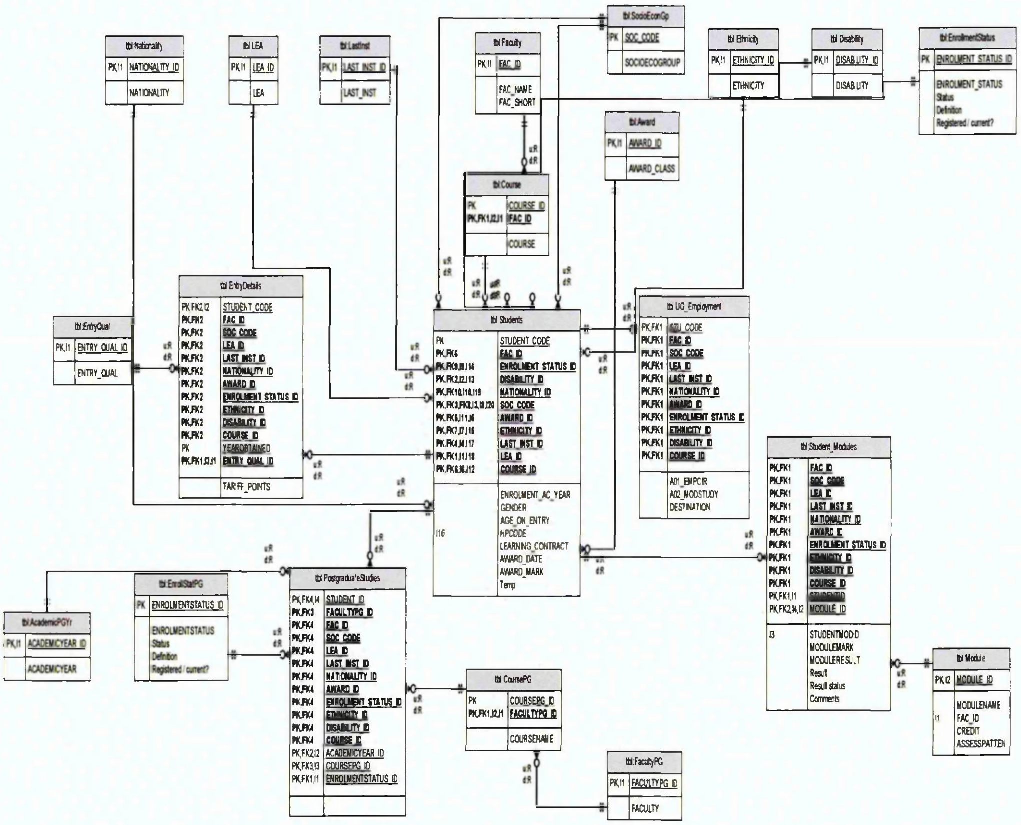
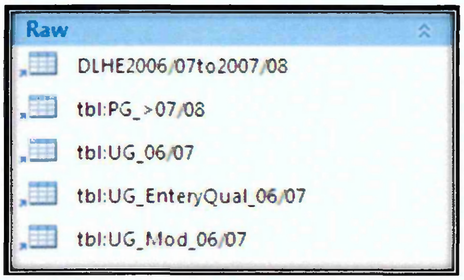


Figure 6.7 – Database Tables.

Data items were selected based upon the research conducted so far and the requirement to predict award classification, progression onto postgraduate studies at SHU and employment type. Having planned the development of the database, the data was then imported into Microsoft Access. A number of areas (raw, staging, final) were then setup within the database and the imported data was assigned to the raw area, figure 6.8.



Raw

DLHE2006/07to2007/08

tbl:PG_>07/08

tbl:UG_06/07

tbl:UG_EnteryQual_06/07

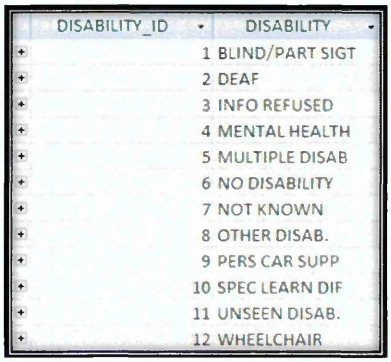
tbl:UG_Mod_06/07

Figure 6.8 – Raw Data Tables.

The development process focused upon the transformation, including the cleaning, of the data. This initially centred upon the raw Undergraduate 2006/07 data and involved the creation of eleven tables, these were:

- Disability;
- Award;
- Entry_Qualification;
- Ethnicity;
- Last_Institution;
- Local_Education_Authority;
- Nationality;
- Socioeconomic_Group;
- Enrolment_Status;
- Students; and
- Entry_Details.

The Disability table was created by updating any blank data values, in the Disability field, to 'not known'. A table was then created by grouping the data values in the disability column and adding a Disability_ID field. The identifier (ID) field was created using the auto number function in Microsoft Access, which later became the primary ID for the table. Figure 6.9 below shows the disability table that was created through this process.

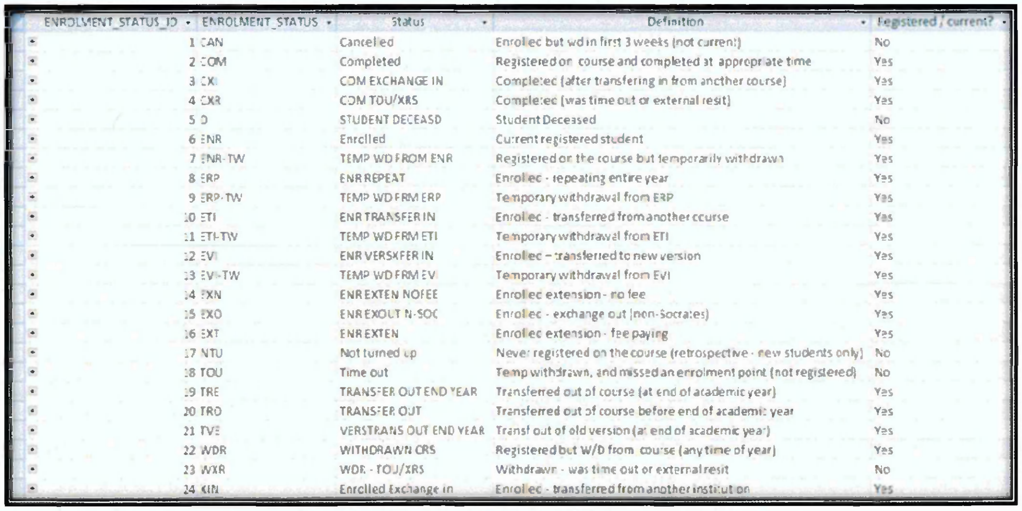


DISABILITY_ID	DISABILITY
1	BLIND/PART SIGT
2	DEAF
3	INFO REFUSED
4	MENTAL HEALTH
5	MULTIPLE DISAB
6	NO DISABILITY
7	NOT KNOWN
8	OTHER DISAB.
9	PERS CAR SUPP
10	SPEC LEARN DIF
11	UNSEEN DISAB.
12	WHEELCHAIR

Figure 6.9 – Disability Table.

The Award table was created using the Award_Class field and replacing any blank values in this field with 'unknown'. Again the data items were grouped and an auto number was created as a unique ID for the table. The same process was then followed in the creation of the Entry_Qualifications, Ethnicity, Last_Institution, Local_Education_Authority, Nationality and Socio_Economic_Group tables.

The Enrolment_Status table was formed by checking for data quality errors, in the Enrolment_Status field, and then grouping the values. This resulted in one Enrolment_Status value per row. An ID field was then added along with additional information. This additional information, figure 6.10, was obtained from SIS and it helped to understand the student's current status.



ENROLMENT_STATUS_ID	ENROLMENT_STATUS	Status	Definition	Registered / current?
1	CAN	Cancelled	Enrolled but wd in first 3 weeks (not current)	No
2	COM	Completed	Registered on course and completed at appropriate time	Yes
3	CXI	COM EXCHANGE IN	Completed (after transfering in from another course)	Yes
4	CXR	COM TOU/XRS	Completed (was time out or external resit)	Yes
5	D	STUDENT DECEASD	Student Deceased	No
6	ENR	Enrolled	Current registered student	Yes
7	ENR-TW	TEMP WD FROM ENR	Registered on the course but temporarily withdrawn	Yes
8	ERP	ENR REPEAT	Enrolled - repeating entire year	Yes
9	ERP-TW	TEMP WD FRM ERP	Temporary withdrawal from ERP	Yes
10	ETI	ENR TRANSFER IN	Enrolled - transferred from another course	Yes
11	ETI-TW	TEMP WD FRM ETI	Temporary withdrawal from ETI	Yes
12	EVI	ENR VERSXFER IN	Enrolled – transferred to new version	Yes
13	EVI-TW	TEMP WD FRM EVI	Temporary withdrawal from EVI	Yes
14	EXN	ENR EXTEN NOFEE	Enrolled extension - no fee	Yes
15	EXO	ENR EXOUT N-SOC	Enrolled - exchange out (non-Socrates)	Yes
16	EXT	ENR EXTEN	Enrolled extension - fee paying	Yes
17	NTU	Not turned up	Never registered on the course (retrospective - new students only)	No
18	TOU	Time out	Temp withdrawn, and missed an enrolment point (not registered)	No
19	TRE	TRANSFER OUT END YEAR	Transferred out of course (at end of academic year)	Yes
20	TRO	TRANSFER OUT	Transferred out of course before end of academic year	Yes
21	TVE	VERSTRANS OUT END YEAR	Transf out of old version (at end of academic year)	Yes
22	WDR	WITHDRAWN CRS	Registered but W/D from course (any time of year)	Yes
23	WXR	WDR - TOU/XRS	Withdrawn - was time out or external resit	No
24	XIN	Enrolled Exchange in	Enrolled - transferred from another institution	Yes

Figure 6.10 – Enrolment Status Table.

The student table was created using a number of staging tables, these tables were used to separate the student data into clean and dirty data. This allowed the dirty student data to be cleaned by fixing any data quality errors and removing any duplicate records. This process involved grouping the values in the Student_Code field and counting the number of times a repeated Student_Code appeared. A table of ID's with their respective counts was then created. Any repeated student codes were moved into a Dirty_Student_Records table and the remaining values were retained in a Clean_Student_Records table. The dirty data items were then resolved by assessing each row manually. Whilst this process was time consuming, it was speeded up by understanding the enrolment status codes of each individual student (figure 6.10), which enabled the last record for each student to be identified. The two clean tables were then merged to create a Student_Staging table. The unique IDs from the other tables were then added to a Students table. This allowed the table to be linked to the other tables that were created previously, figure 6.11.

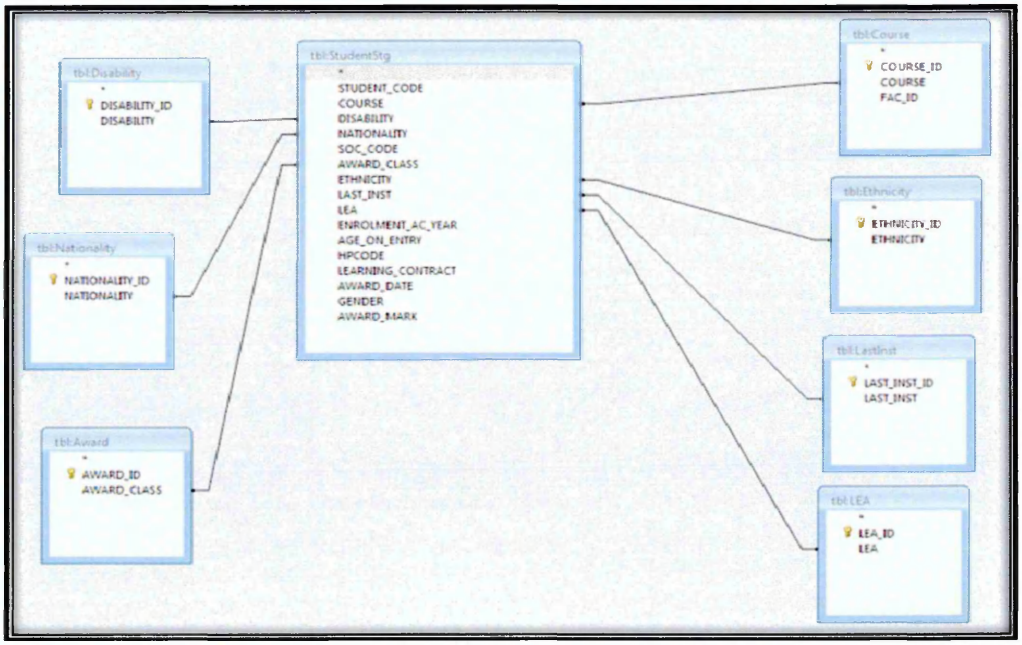
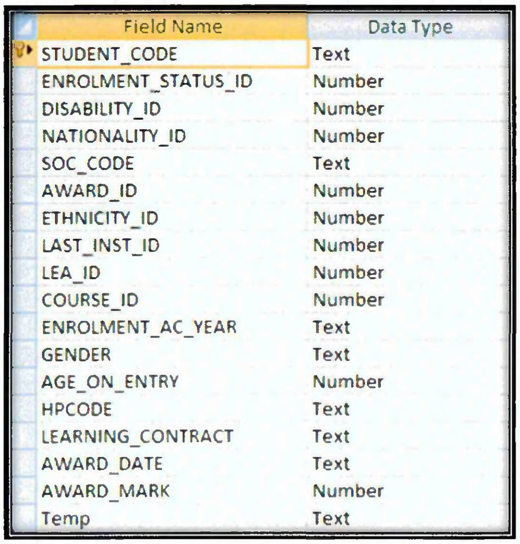


Figure 6.11 – Creating the Student Table.

A number of extra fields were then added to the table to create the final Students table, figure 6.12.

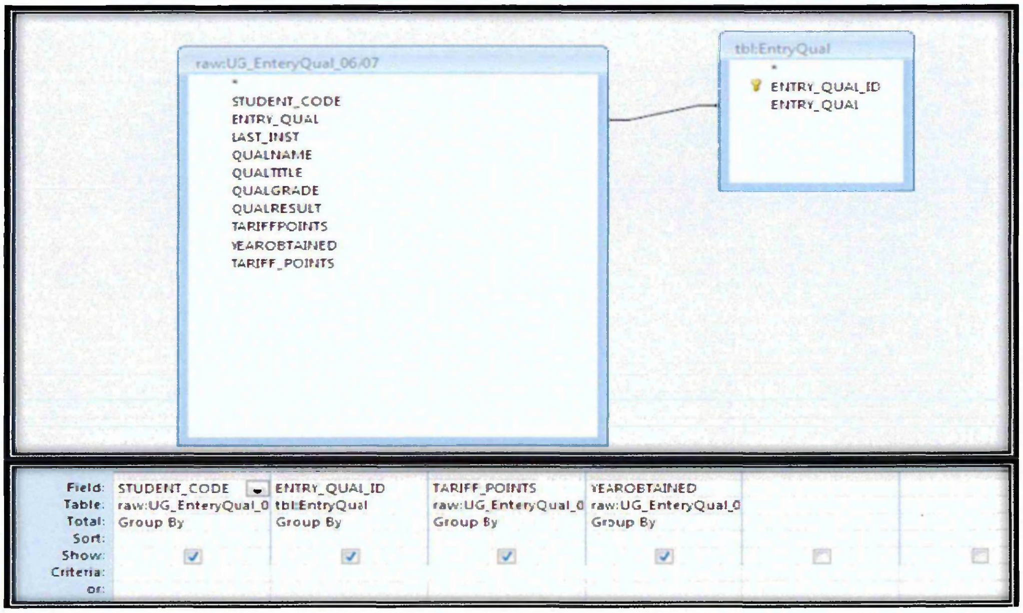


Field Name	Data Type
STUDENT_CODE	Text
ENROLMENT_STATUS_ID	Number
DISABILITY_ID	Number
NATIONALITY_ID	Number
SOC_CODE	Text
AWARD_ID	Number
ETHNICITY_ID	Number
LAST_INST_ID	Number
LEA_ID	Number
COURSE_ID	Number
ENROLMENT_AC_YEAR	Text
GENDER	Text
AGE_ON_ENTRY	Number
HPCODE	Text
LEARNING_CONTRACT	Text
AWARD_DATE	Text
AWARD_MARK	Number
Temp	Text

Figure 6.12 – Final Students Table.

Having previously resolved the data quality issues the Student_Code was then selected as the primary key.

The next stage involved creating an Entry_Details table using the raw Undergraduate Entry Qualifications 2006/07 data. The Student_Code, Entry_Qualification_ID, Tariff_Points and Year_Obtained fields were then grouped to produce a single row of data per entry. The Entry_Qualification_ID was added to the table from the Entry_Qualifications table to enable the two tables to be linked, figure 6.13.



raw:UG_EnteryQual_06/07

*
STUDENT_CODE
ENTRY_QUAL
LAST_INST
QUALNAME
QUALTITLE
QUALGRADE
QUALRESULT
TARIFFPOINTS
YEAROBTAINED
TARIFF_POINTS

tbl:EntryQual

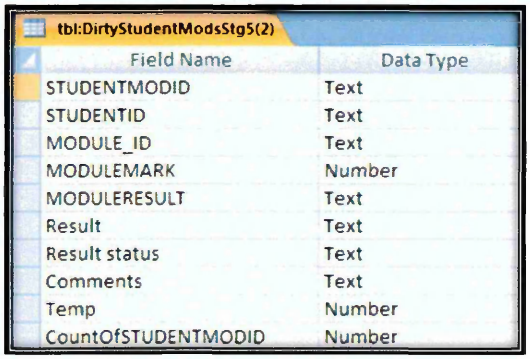
*
ENTRY_QUAL_ID
ENTRY_QUAL

Field:	STUDENT_CODE	ENTRY_QUAL_ID	TARIFF_POINTS	YEAROBTAINED		
Table:	raw:UG_EnteryQual_0	tbl:EntryQual	raw:UG_EnteryQual_0	raw:UG_EnteryQual_0		
Total:	Group By	Group By	Group By	Group By		
Sort:						
Show:	☑	☑	☑	☑	□	□
Criteria:						
or:						

Figure 6.13 – Creating the Enrolment Status Table.

A composite key was then created on the Student_Code and Year_Obtained, as it was acceptable to expect repeat student codes within the table for different years.

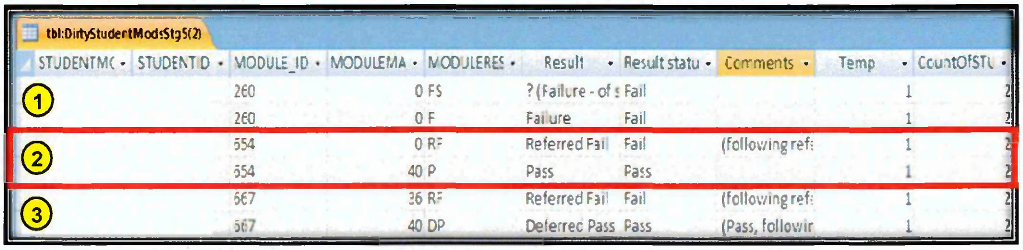
The raw Undergraduate Modules 2006/07 data was then used to create tables for Course, Faculty, Module and Student_Modules. The Faculty table was built by grouping the data in the Fac_Name. Again an ID (Fac_ID) was added to the table along with an abbreviated faculty name. A Course table was then created in the same way and Fac_Name was also added to the table. This allowed the Fac_ID field to be added to the Course table, which generated a relationship of many courses to one faculty. A Module table was then created following a similar method. A Student_Module_Staging table was then created by using the Module table to incorporate the Moudle_ID field into the Student_Module table. A unique data item was then created by concatenating the Student_Code and Module_ID fields together. This field was then sorted in ascending order and a count of the concatenated value was made. Rows were removed, into a Dirty_Student_Modules table (figure 6.14), where the count of the value was greater than or equal to two. The remaining data values were exported into a Clean_Student Modules table.


tbl:DirtyStudentModsStg5(2)

Field Name	Data Type
STUDENTMODID	Text
STUDENTID	Text
MODULE_ID	Text
MODULEMARK	Number
MODULERESULT	Text
Result	Text
Result status	Text
Comments	Text
Temp	Number
CountOfSTUDENTMODID	Number

Figure 6.14 – Dirty Student Modules.

The Dirty_Student_Modules were then manually updated so that there was one Student_Code to one Module_ID. The majority of repeats in the table were due to students failing their course and subsequently passing it when they repeated the exam or coursework. Figure 6.15, below, is an example of the dirty data found within the raw Undergraduate Modules 2006/07 data.


tbl:DirtyStudentModsStg5(2)

STUDENTMC	STUDENTID	MODULE_ID	MODULEMA	MODULERES	Result	Result statu	Comments	Temp	CountOfSTU
		260	0	FS	? (Failure - of s	Fail		1	2
		260	0	F	Failure	Fail		1	2
		654	0	RF	Referred Fail	Fail	(following ref	1	2
		654	40	P	Pass	Pass		1	2
		667	36	RF	Referred Fail	Fail	(following ref	1	2
		667	40	DP	Deferred Pass	Pass	(Pass, followir	1	2

Figure 6.15 – Example Dirty Data.

The example above shows that student two failed module 654 and subsequently passed with an overall mark of 40.

The Student_Modules table was formed by merging the cleaned Dirty_Student_Modules with the Clean_Student_Modules data. A composite key was then created using the Student_Code and Module_ID fields. Links were then established between the Students and Course tables, and between the Course and Faculty tables. The Student_Code in the Students table was then used to generate a link to the Student_Modules, which was later linked to the Modules table.

The raw Postgraduate 2007 to 2010 data was then used to create five Postgraduate tables. These tables were Postgraduate_Academic_Year, Postgraduate_Course, Postgraduate_Enrolment_Status and Postgraduate_Studies. All tables were created in the same way as previous tables but using the postgraduate data. The Postgraduate_Studies table was built using the Student_Code and the ID fields from each of the Postgraduate tables. A composite key was then created using Student_ID, Academic_Year_ID, Postgraduate_Course_ID and Enrolment_Status_ID. The data, in the Postgraduate_Studies table, was then reduced by identifying which undergraduate students had gone on to study at Postgraduate level, figure 6.16.

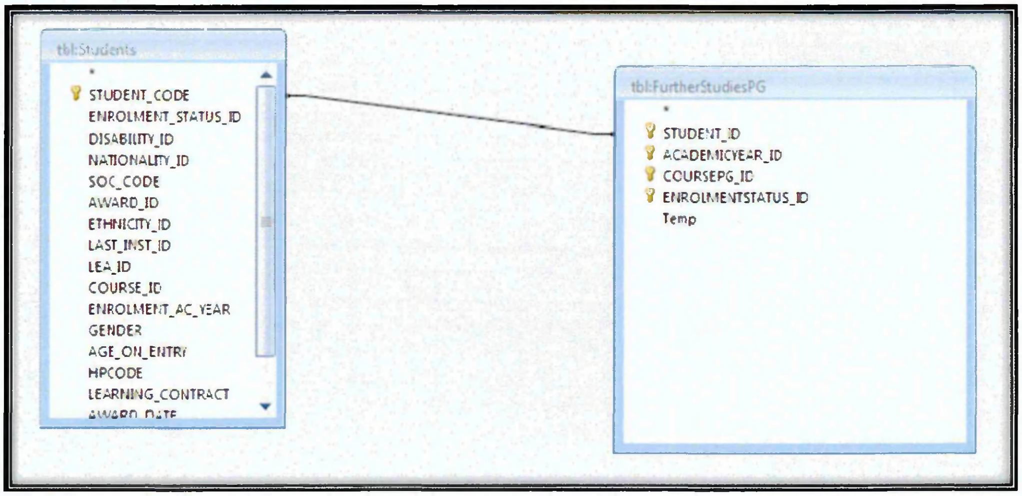
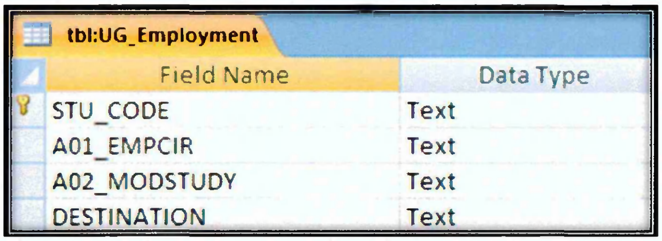


Figure 6.16 – Reducing the Postgraduate Data Set.

Figure 6.16, above, shows the Postgraduate_Studies table linked to the Students table using Student_Code. The remaining data items, within the Postgraduate_Studies table, were then linked to the other Postgraduate tables created previously.

The final table, Undergraduate_Employment, was built from the DOL 2006 to 2008 data. Based on the meetings with the SIS department, the following fields were selected, figure 6.17, to create the table



tbl:UG_Employment

Field Name	Data Type
STU_CODE	Text
A01_EMPCIR	Text
A02_MODSTUDY	Text
DESTINATION	Text

Figure 6.17 – Data Items for Undergraduate Employment Table.

Repeated records were removed from the data set by grouping on all data fields within the table. This resulted in one record per student in the table and reduced the chances of repeated values skewing the final DM models. The students in the Undergraduate_Employment table were further reduced by identifying the students that appeared in both the Undergraduate_Employment and Students table.

The final data tables were then copied into another database as this kept the transformation process separate from the process of building the data marts. Figure 6.18 shows the final normalised database structure that was created following the processes detailed above.

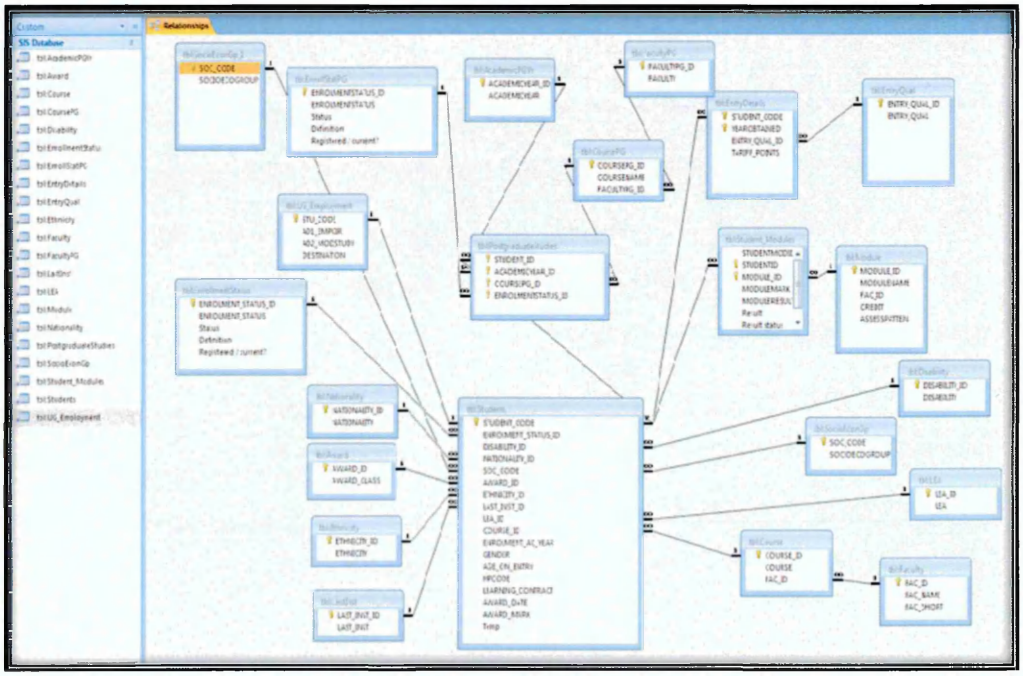


Figure 6.18 – Normalised Database Structure.

The normalised database was then exported into another Microsoft Access Database where queries were created to build three different DM marts.

6.1.3.2.2 CONSTRUCT AND INTEGRATE DATA

This outlines the processes followed in constructing and integrating the data into a single DM mart, which will be used to predict student award classification, progression onto postgraduate studies at SHU and employment type. The DM mart was created using the select and clean data process as outlined above.

1. Undergraduate Award Classificaiton

The data required to predict award classification was added by selecting fields from the student table that were thought to be useful in predicting whether the student would receive an honours degree. All students were included in the data set regardless of whether they received an honours classification. Figure 6.19 shows the tables used to create the award classification DM mart.

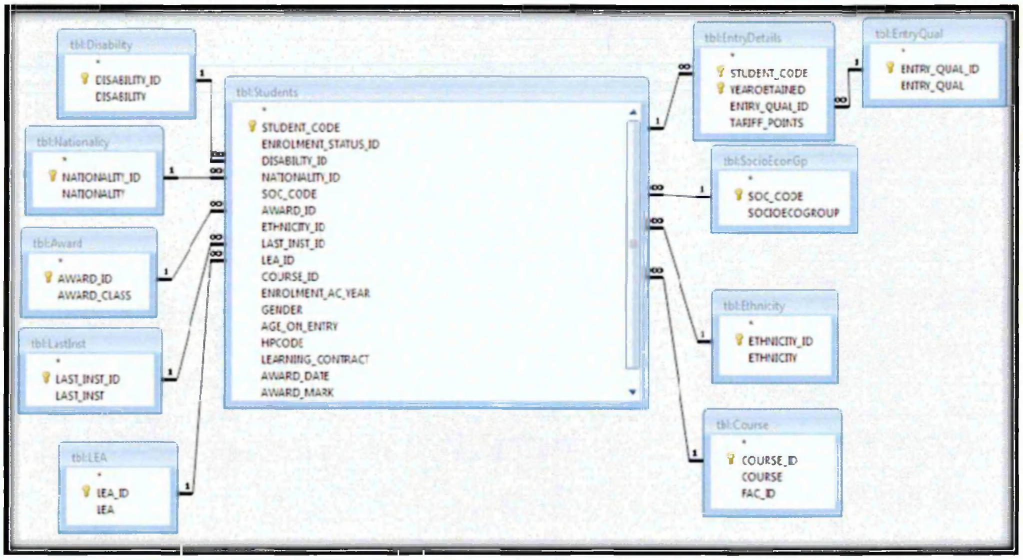


Figure 6.19 –Undergraduate Mart Tables.

The following sixteen fields, figure 6.20, were selected from these tables and aliases were created.

Field No.	Field Name	Alias	Criteria
1	STUDENT_CODE	StudentNumber	
2	ENROLMENT_AC_YEAR	EnrolmentYr	
3	ENTRY_QUAL	EntryQualification	
4	DISABILITY	Disability	
5	NATIONALITY	Nationality	
6	ETHNICITY	Ethnicity	
7	SOCIOECOGROUP	SocioEconomicGP	
8	GENDER	Gender	
9	HPCODE	HomePostcode	
10	LEA	LEA	
11	AGE_ON_ENTRY	EntryAge	
12	TARIFF_POINTS	EntryPoints	
13	COURSE	Course	
14	AWARD_DATE	AwardDate	
15	AWARD_CLASS	AwardClassification	
16	AWARD_MARK	AwardMark	

Figure 6.20 – Award Classification Mart Fields.

This generated four thousand and twenty three records for the award classification DM mart.

2. Postgraduate Studies

The postgraduate studies data was selected by adding the postgraduate studies table and the postgraduate academic year table to the undergraduate DM mart tables in figure 6.19. Figure 6.21 shows the postgraduate DM mart tables, here the undergraduate tables are linked to the postgraduate tables where the same student number appears in both tables.

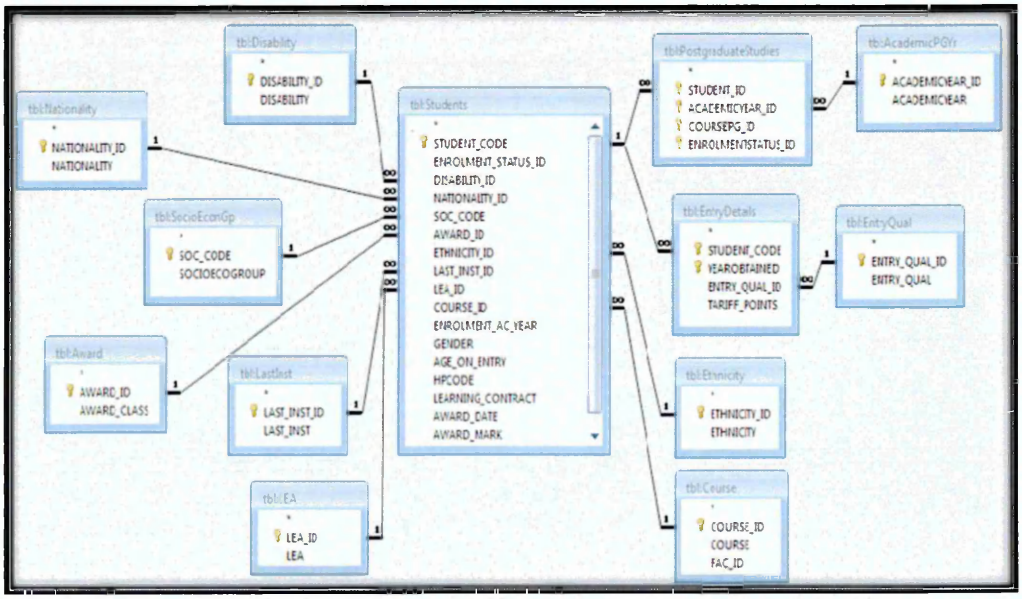


Figure 6.21 –Postgraduate Mart Tables.

The following fourteen fields, figure 6.22, were selected from these tables and aliases were created. Given that the research is interested in determining whether a student had progressed onto postgraduate studies at SHU. The criteria for postgraduate academic year was set to identify the last year that the student was enrolled on a postgraduate course. The reason for this was that some students were enrolled over two academic years and this created a repeat record.

Field No.	Field Name	Alias	Criteria
1	STUDENT_ID	StudentNumber	
2	ACADEMICYEAR	PGEnrolmentYr	Max
3	AWARD_DATE	UGAwardDate	
4	AWARD_CLASS	UGClassification	
5	ENTRY_QUAL	UGEntryQualification	
6	TARIFF_POINTS	UGEntryPoints	
7	COURSE	UGCourse	
8	DISABILITY	Disability	
9	NATIONALITY	Nationality	
10	ETHNICITY	Ethnicity	
11	SOCIOECOGROUP	SocioEconomicGP	
12	GENDER	Gender	
13	HPCODE	HomePostcode	
14	LEA	LEA	

Figure 6.22 – Postgraduate Studies Mart Fields.

This generated three hundred and twenty two rows of data for the progression onto postgraduate studies at SHU mart.

3. Employment Type

The final DM mart created through this process was the employment mart, this used a combination of the undergraduate mart tables, in figure 6.19, with the addition of the undergraduate employment table. In figure 6.23, below, the students and employment tables are linked using the student number. The tables were joined so that all undergraduate students would be included in the employment mart. If there was no employment record available for a student the destination field was updated to not recorded.

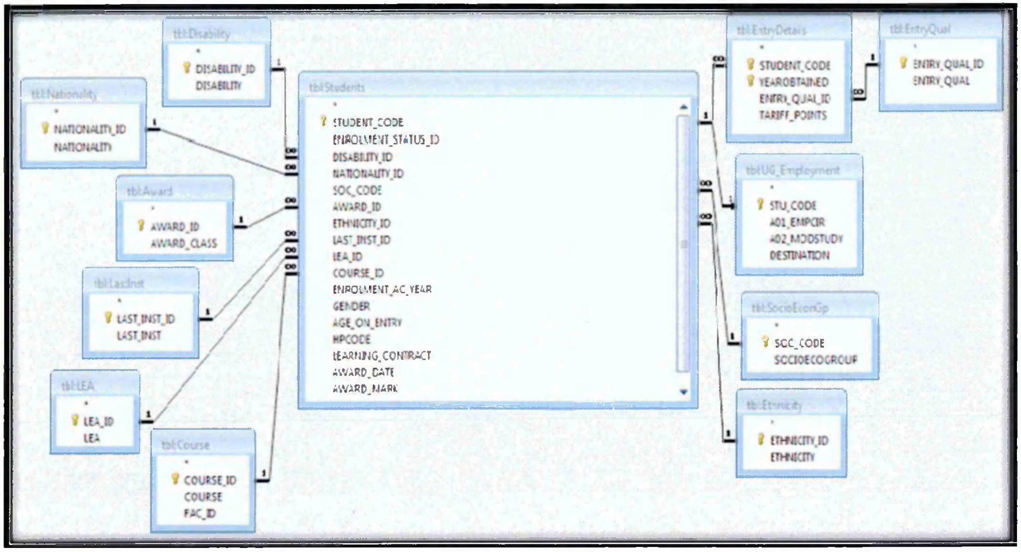


Figure 6.23 –Employment Mart Tables.

Seventeen fields, figure 6.24, were included in the employment mart and aliases for the field names were created.

Field No.	Field Name	Alias	Criteria
1	STUDENT_CODE	StudentNumber	
2	ENROLMENT_AC_YEAR	EnrolmentYr	
3	ENTRY_QUAL	EntryQualification	
4	DISABILITY	Disability	
5	NATIONALITY	Nationality	
6	ETHNICITY	Ethnicity	
7	SOCIOECOGROUP	SocioEconomicGP	
8	GENDER	Gender	
9	HPCODE	HomePostcode	
10	LEA	LEA	
11	AGE_ON_ENTRY	EntryAge	
12	TARIFF_POINTS	EntryPoints	
13	COURSE	Course	
14	AWARD_DATE	AwardDate	
15	AWARD_CLASS	AwardClassification	
16	AWARD_MARK	AwardMark	
17	DESTINATION	PostUGDestination	

Figure 6.24 –Employment Mart Field.

A total of four thousand and twenty three records were generated.

Two additional fields were later added based on the HEFCE work around POLAR2 (Participation of Local Areas), which looks at the participation of young people in HE based on their location (postcode) (QYPR) and the number of adults with HE qualifications who live in a certain area (postcode) (QAHE).

These measures rate participation and adult qualifications on a scale of 1 to 5, where 1 is lowest participation/number of qualifications and 5 relates to high participation/number of qualifications (HEFCE 2012).

Through carrying out the initial understanding of the data, in the data understanding section (below), and given the student demographics data was the same, in each data mart. It was decided that all three DM marts could be merged into a single table with the correct fields added for award classification, progression onto postgraduate studies at SHU and employment type. Furthermore, additional variables were also added as potential replacements for the Course variable. Therefore, JACS Subject and Faculty (see Appendix I for a list of faculty departments) were added to the single table view for further analysis.

6.1.3.2.3 Format Data

Having built the required DM mart the final part of the data preparation was to convert the data into a suitable format that could be accepted by SAS® Enterprise Miner. The acceptable format of SAS® is a single CSV file with a tab, comma or space delimiter. Comma delimiters were chosen as it was the default for Microsoft Access. Furthermore, the DM mart was built specifically to be used in SAS® as the original source of the data was a relational database, which had to be changed to a single flat for DM, hence the creation of the DM mart.

6.1.3.2.4 Understand Data

This section was added by the author as it is important to understand the data before the DM models are built. It is important to note that certain variables, such as date and student number, were omitted from this section as these will be rejected at the start of the DM process.

1. Initial Data Set

The student dataset has 26 variables and 4023 observations

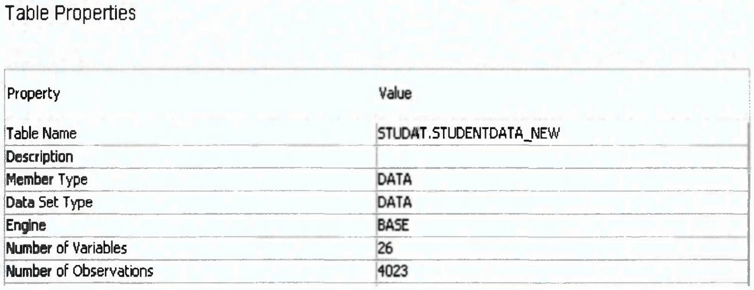
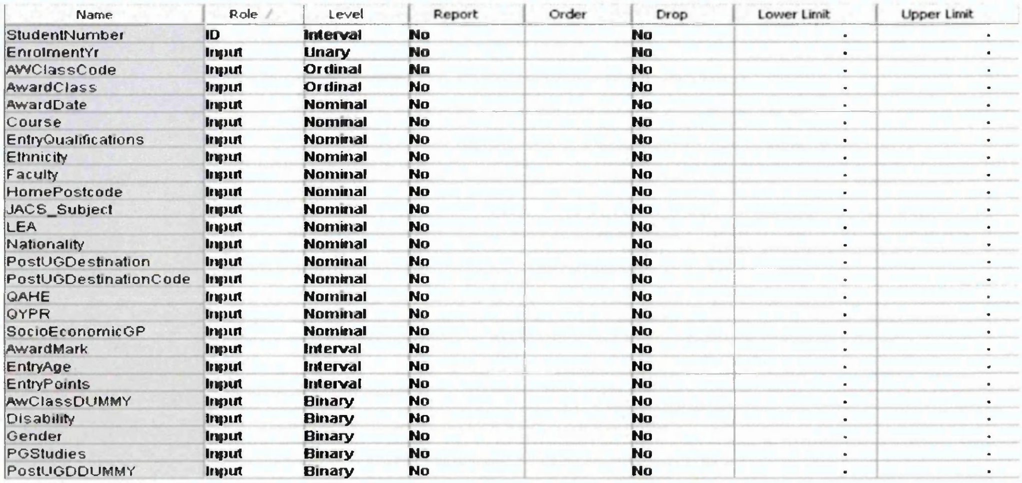


Table Properties

Property	Value
Table Name	STUDAT.STUDENTDATA_NEW
Description	
Member Type	DATA
Data Set Type	DATA
Engine	BASE
Number of Variables	26
Number of Observations	4023

Figure 6.25 - Initial Data Set Table Properties.

The data set holds all three target variables AwardClass, PGStudies and PostUGDestination.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
StudentNumber	ID	Interval	No		No	.	.
EnrolmentYr	Input	Unary	No		No	.	.
AWClassCode	Input	Ordinal	No		No	.	.
AwardClass	Input	Ordinal	No		No	.	.
AwardDate	Input	Nominal	No		No	.	.
Course	Input	Nominal	No		No	.	.
EntryQualifications	Input	Nominal	No		No	.	.
Ethnicity	Input	Nominal	No		No	.	.
Faculty	Input	Nominal	No		No	.	.
HomePostcode	Input	Nominal	No		No	.	.
JACS_Subject	Input	Nominal	No		No	.	.
LEA	Input	Nominal	No		No	.	.
Nationality	Input	Nominal	No		No	.	.
PostUGDestination	Input	Nominal	No		No	.	.
PostUGDestinationCode	Input	Nominal	No		No	.	.
QAHE	Input	Nominal	No		No	.	.
QYPR	Input	Nominal	No		No	.	.
SocioEconomicGP	Input	Nominal	No		No	.	.
AwardMark	Input	Interval	No		No	.	.
EntryAge	Input	Interval	No		No	.	.
EntryPoints	Input	Interval	No		No	.	.
AwClassDUMMY	Input	Binary	No		No	.	.
Disability	Input	Binary	No		No	.	.
Gender	Input	Binary	No		No	.	.
PGStudies	Input	Binary	No		No	.	.
PostUGDDUMMY	Input	Binary	No		No	.	.

Figure 6.26 - Initial Variables and Roles.

The StudentNumber is a unique identification number that is used to identify the students. This has been assigned the role of ID as it is not pertinent to the DM analysis. The majority of the other variables (with the role of input) will be assessed through the data understanding and SEMMA process, as to their suitability for predicting the target variable. These variables include:

- EnrolmentYr - is a unary variable and it contains the value 2006/07;
- AwardClass - is the actual award the student received;
- AwardDate - is the date that the student received the award;
- Course - is the course the student was on in the final year of their degree;
- EntryQualifications - the initial type of qualification that the student entered onto their course with;

- Ethnicity - the students common culture/ancestry and physical appearance;
- Faculty - the department which the students course belongs to;
- HomePostcode - relates to the students home postcode;
- JACS Subject - is a group of courses, this provides a much higher level of information than course but more detail than faculty;
- LEA - location based variable that relates to the students home Local Education Authority;
- Nationality - relates to those students who share a common characteristic based on certain criteria, such as language, ethnic identity and/or culture *et* cetera;
- PostUGDestination - is the target variable that indicates what type of employment the student went onto after their undergraduate studies;
- QAHE - a location variable that pertains to adults with HE qualifications by postcode;
- QYPR - a location based variable that relates to participation of young people in HE by postcode;
- SocioEconomicGP - used to determine the students background based on their parents social economic group;
- AwardMark - the final mark received by the student which is used to determine the award classification;
- EntryAge - the students age when they entered onto their course;
- EntryPoints - the UCAS points that the student entered onto their course with;
- Disability - relates to whether the student has a disability;
- Gender - used to determine whether the student is male or female;
- PGStudies - is the target variable that indicates whether the student went onto postgraduate studies;

Four extra variables have been created (to explore ways of automatically grouping values), two for AwardClass - AwardClassCode and AWClassDummy and two for PostUGDestination – PostUGDDUMMY and PostUGDestinationCode.

- AWClassCode is a code that relates to the different award classes, which could be used at the modelling stage. These codes are 1=1st, 2=2:1, 3=2:2, 4=3rd and 5=unknown.
- The AWClassDUMMY variable was created so that Interactive Binning/Grouping could be carried out, as SAS® will be used to help aid the decision process in determining the most appropriate groups for the different variables. The SAS® Interactive Binning/Grouping node requires a binary target variable. Hence a AWClassDUMMY variable was created with the following values:
    - <=2:1 (relates to those students who obtained a 1st or a 2:1);
    - >2:1 (pertains to those student who received a 2:2 or a 3rd)
    - Missing values are dealt with automatically by the node, as observations are assigned to a separate branch.
- PostUGDDUMMY variable was created so that Interactive Binning/Grouping could be carried out, as SAS® will be used to help aid the decision process in determining the most appropriate groups for the different variables. The SAS® Interactive Binning/Grouping node requires a binary target variable. Hence a PostUGDDUMMY variable was created with the following values:
    - Employed = (relates to students obtaining a Graduate or a Non-Graduate job);
    - Unemployed (pertains to Other, Study, Unemployed)
    - Missing values are dealt with automatically by the node, as observations are assigned to a separate branch (these include the unknow and not recorded values).
- PostUGDestinationCode is a code that relates to the different employment types, which could be used at the modelling stage. These codes are 1=Graduate Job, 2=Non-Graduate Job, 3=Other, 4=Study, 5=Unemployed, 6=Not Recorded.

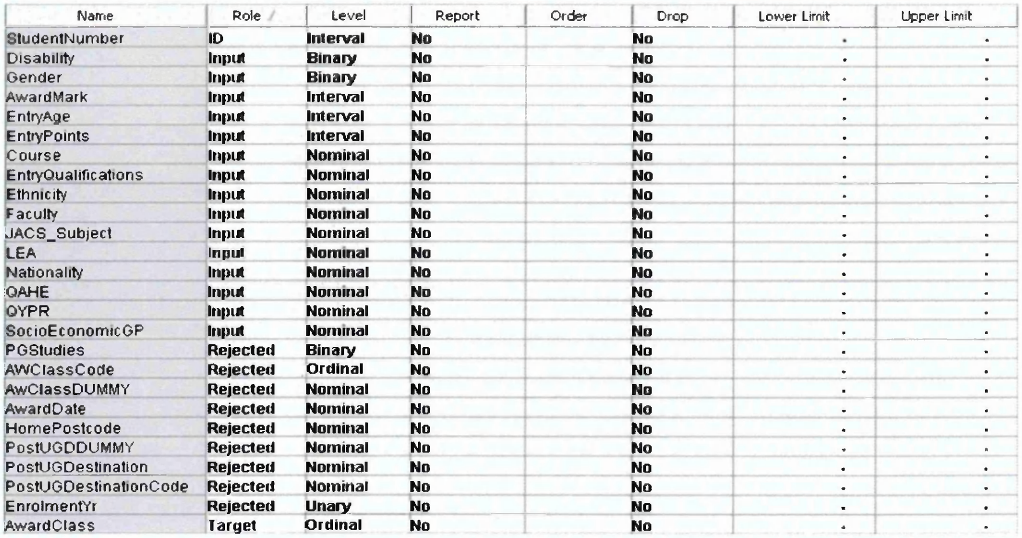
The role and levels of these variables are

Role	**Level**	**No. of variables**
ID	Interval	1
Input	Nominal	14
Input	Binary	5
Input	Interval	3
Input	Unary	1
Input	Ordinal	2
Total No. Variables		**26**

Figure 6.27 - Initial Roles, Levels and Counts.

2. Undergraduate Award Classificaiton

Using AwardClass as the target variable (rejecting the other two target variables (PGStudies and PostUGDestination) the list below shows the variables to be assessed.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
StudentNumber	ID	Interval	No		No	.	.
Disability	Input	Binary	No		No	.	.
Gender	Input	Binary	No		No	.	.
AwardMark	Input	Interval	No		No	.	.
EntryAge	Input	Interval	No		No	.	.
EntryPoints	Input	Interval	No		No	.	.
Course	Input	Nominal	No		No	.	.
EntryQualifications	Input	Nominal	No		No	.	.
Ethnicity	Input	Nominal	No		No	.	.
Faculty	Input	Nominal	No		No	.	.
JACS_Subject	Input	Nominal	No		No	.	.
LEA	Input	Nominal	No		No	.	.
Nationality	Input	Nominal	No		No	.	.
QAHE	Input	Nominal	No		No	.	.
QYPR	Input	Nominal	No		No	.	.
SocioEconomicGP	Input	Nominal	No		No	.	.
PGStudies	Rejected	Binary	No		No	.	.
AWClassCode	Rejected	Ordinal	No		No	.	.
AwClassDUMMY	Rejected	Nominal	No		No	.	.
AwardDate	Rejected	Nominal	No		No	.	.
HomePostcode	Rejected	Nominal	No		No	.	.
PostUGDDUMMY	Rejected	Nominal	No		No	.	.
PostUGDestination	Rejected	Nominal	No		No	.	.
PostUGDestinationCode	Rejected	Nominal	No		No	.	.
EnrolmentYr	Rejected	Unary	No		No	.	.
AwardClass	Target	Ordinal	No		No	.	.

Figure 6.28 - Undergraduate Award Classification Variables and Roles.

The role and levels of these variables are:

Role	**Level**	**No. of variables**
ID	Interval	1
Input	Nominal	10
Input	Binary	2
Input	Interval	3
Rejected	Binary	1
Rejected	Nominal	6
Rejected	Ordinal	1
Rejected	Unary	1
Target	Ordinal	1
Total No. Variables		**26**

Figure 6.29 - Undergraduate Award Classification Roles, Levels and Counts.

Rejected variables include:

- PGStudies - the target variable for another model;
- AWClassCode - relates to the different award classes;
- AWClassDummy - used to determine potential groups;
- AwardDate - the date that the student received the award, this has no significance to the AwardClass received by the student;
- PostUGDUMMY - used to determine potential groups for another model;
- PostUGDestination - the target variable for another model;
- PostUGDestinationCode - relates to the different postgraduate desinations for another model;
- EnrolmentYr - a unary variable;
- HomePostcode - there are other location based variables in the dataset such as QYPR, QAHE and LEA.

An initial assessment of the histograms indicates that the data is mainly nominal (see glossary page ix), below.

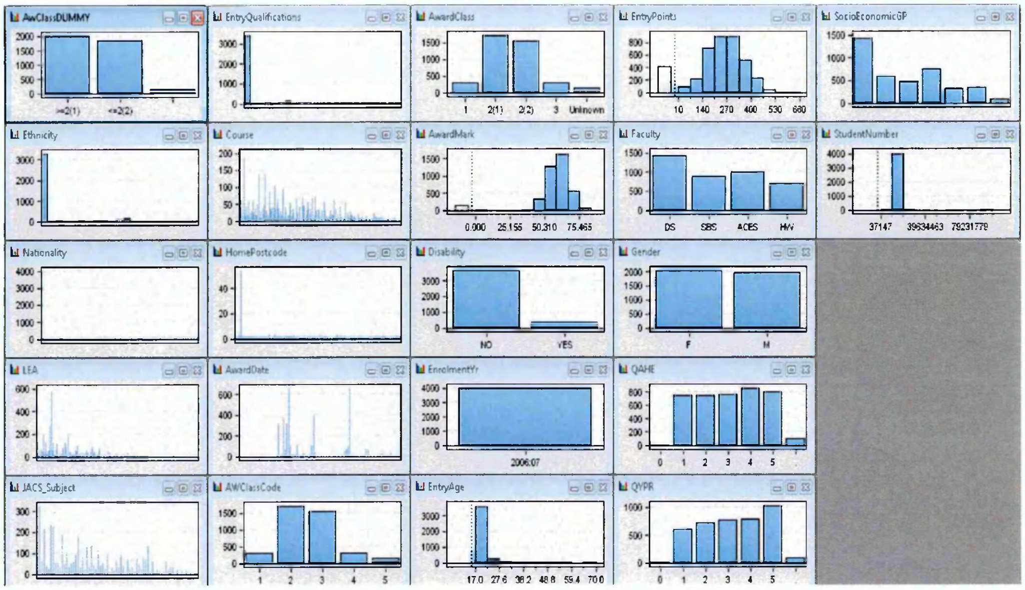


Figure 6.30 - Undergraduate Award Classification Initial Histograms.

The above graphs also highlight that:

- AwardDate, JACS Subject, EntryQualifications, LEA, Course, Ethnicity and Nationality have too many overall levels;
- EntryAge is positively skewed and there are two potential outliers (17 & 70);
- AwardMark is negatively skewed;
- SocioEconmicGP, LEA, QYPR and QAHE are all location based variables; and
- Some variables contain missing values, which will have to be resolved.

Each variable will now be examined individually.

AWARD CLASSIFICATION

The graphs below show AwardClass, AWClassCode and AWClassDUMMY against Frequency.

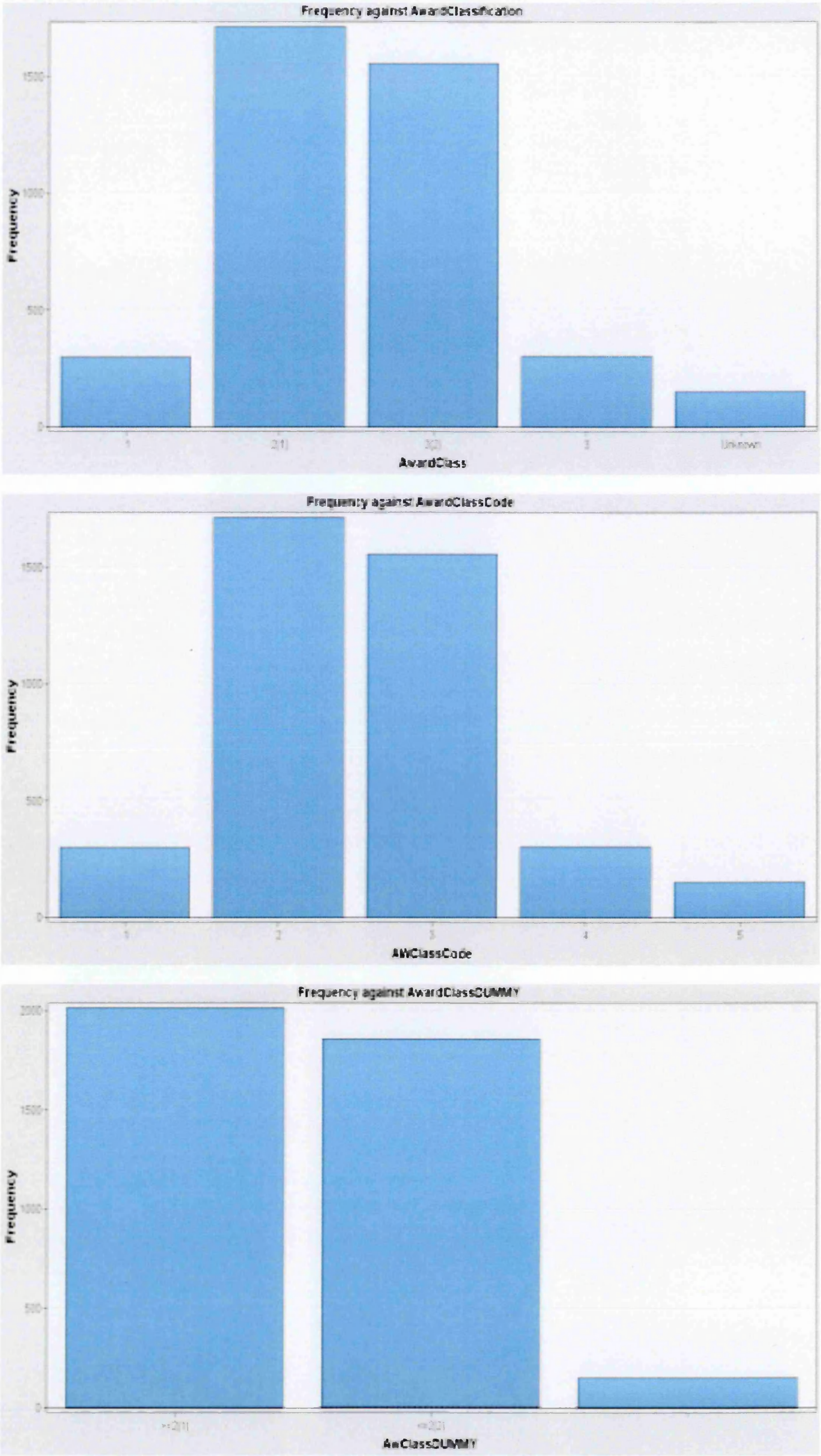


Figure 6.31 - Award Classification Histograms.

Please note that the top bar chart presents the data in reverse order, in that the award classifications decrease when moving from left (1^{st}) to right (3^{rd}). Of the 4023 students 3873 obtained an honours degree (1^{st}, 2:1, 2:2 or 3^{rd}) this suggests that 96% obtained an honours degree and 150 of the results are unknown. The AwardClass and AwardClassCode graphs show that in general more students obtain a 2:1 and that a first is less common. It is also important to note that the missing values have been coded as unknown. The AWClassDUMMY variable contains more students who obtained a grade greater than or equal to a 2:1, this is due mainly to the large number of students obtaining a 2:1.

AWARD MARK

The AwardMark is plotted against the Frequency below.

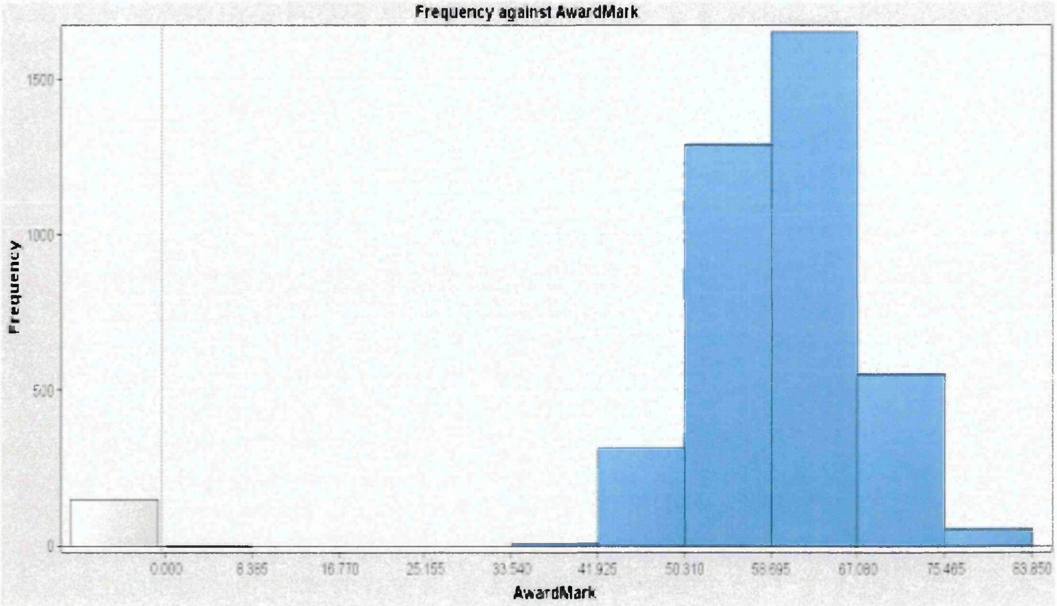


Figure 6.32 - Award Mark Histogram.

This bar chart reflects the award classification graphs above as it shows that the majority of students received an AwardMark between 58 and 67. There are also some missing values that would need to be resolved. This variable will be rejected at the modelling stage as it is used to calculate the AwardClassification target variable.

ENTRY AGE

A bar chart of EntryAge against Frequency and AwardClass against EntryAge is plotted below.

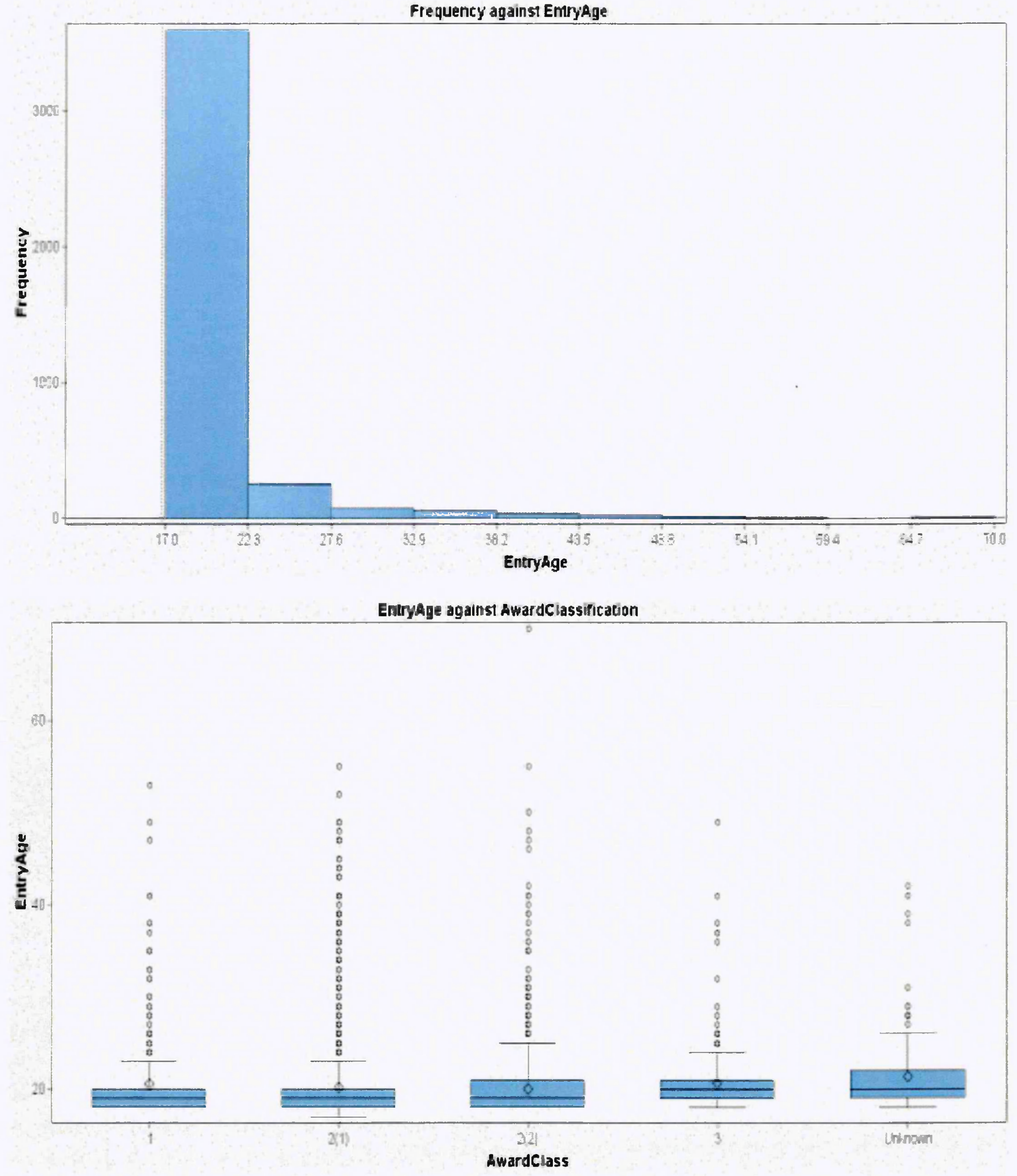


Figure 6.33 - Entry Age Histograms.

An initial assessment of EntryAge shows that the student's entry age varies from 17 to 70. The majority of students entered university between the age of 17 and 22 and very few students entered onto a degree after the age of 27. Assessing EntryAge against award classification suggests that the average entry age of students who obtain a 3rd is 20. Furthermore those students who were awarded a 1st and a 2:1 tended to be younger than those who obtain a 2:2 or a 3rd. Looking at EntryAge against EntryQualificiations, below, it can be inferred that younger students are recruited from A-Level/GNVQ3 and ONC/OND.

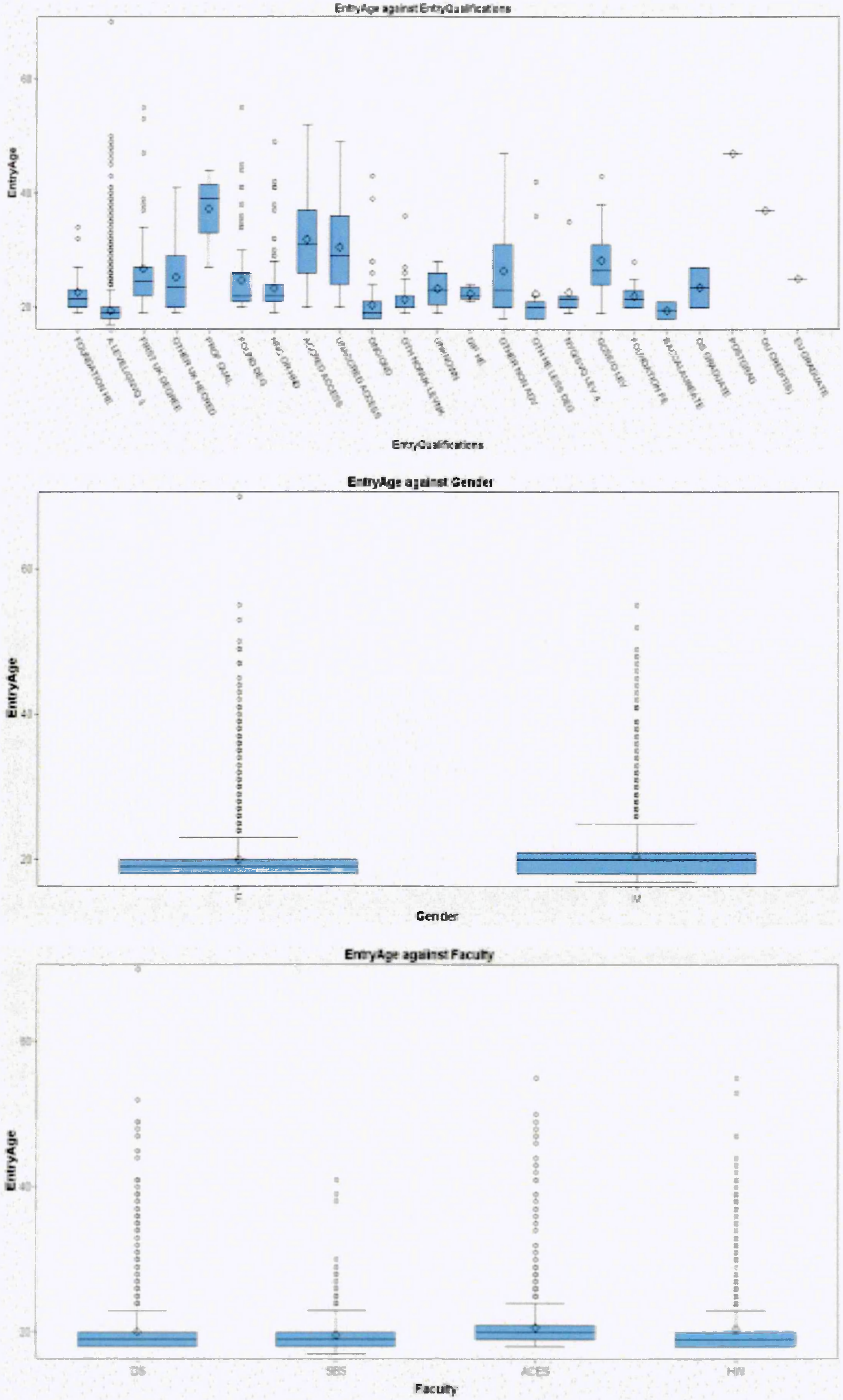


Figure 6.34 - Entry Age Box Plots.

When considering EntryAge against Gender, on average more males entered their degrees at the age of twenty whereas more females entered at an average age of less than twenty. Moreover, students who entered the Faculty of ACES, on average, tended to be older than those on Health and Wellbeing courses. Further investigation was carried out into two potential entry age outliers, 17 and 70, this determined that there were in fact two students who entered onto a full-time degree course at the ages of 17 and 70.

NATIONALITY

The Nationality of the students is plotted in both a bar chart and mosiac below.

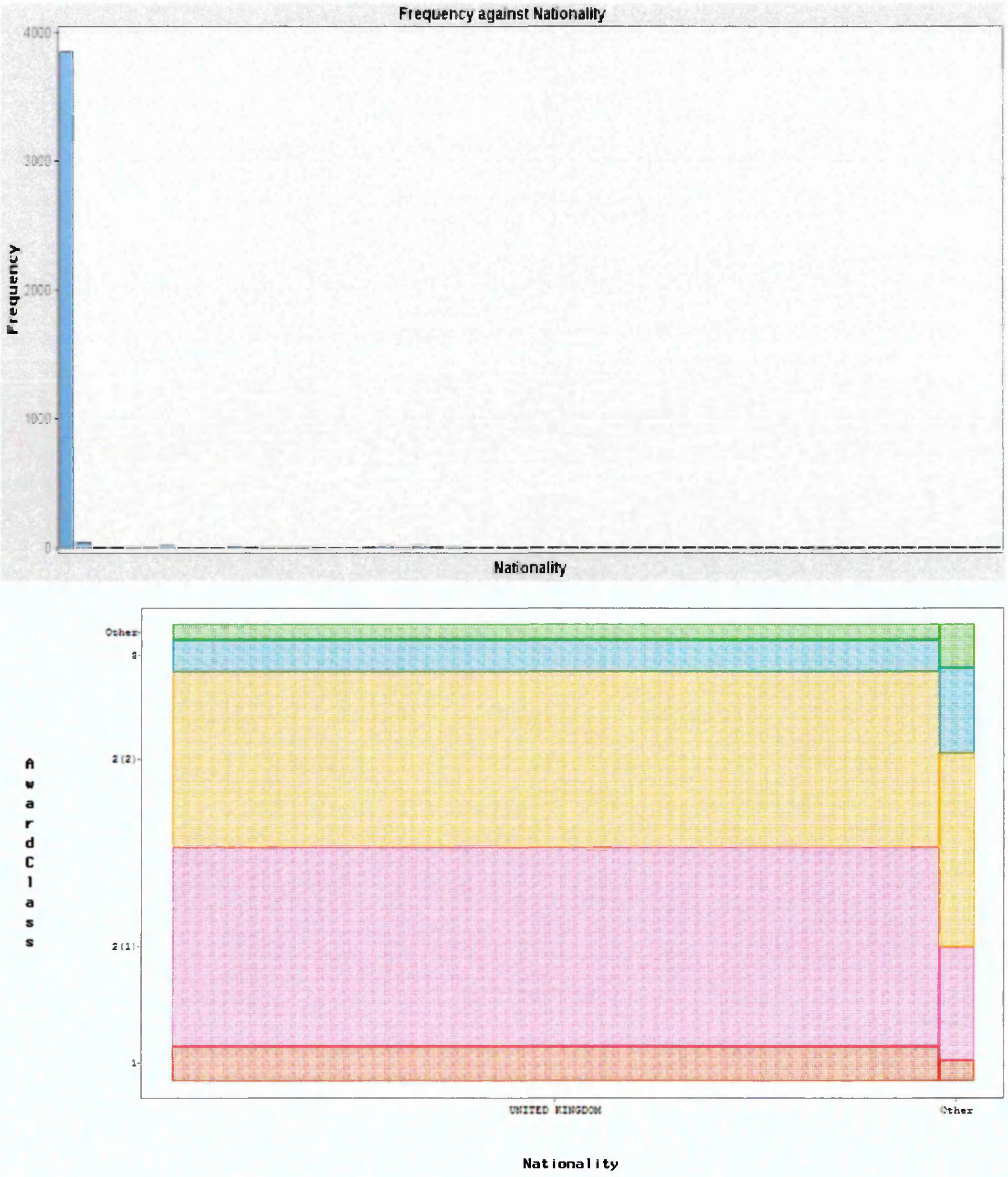


Figure 6.35 - Nationality Bar Chart and Mosiac Plot.

These suggest that there are too many categories to the Nationality variable. The majority of the students are from the UK. The mosaic plot above indicates that UK students tend to achieve better grades as the majority of UK students achieve either a 2:1 or a 2:2. Non UK students tended to obtain more 2:2 and 3rd class classifications.

ETHNICITY

The Ethnicity of the students is plotted in both a bar chart and mosiac below.

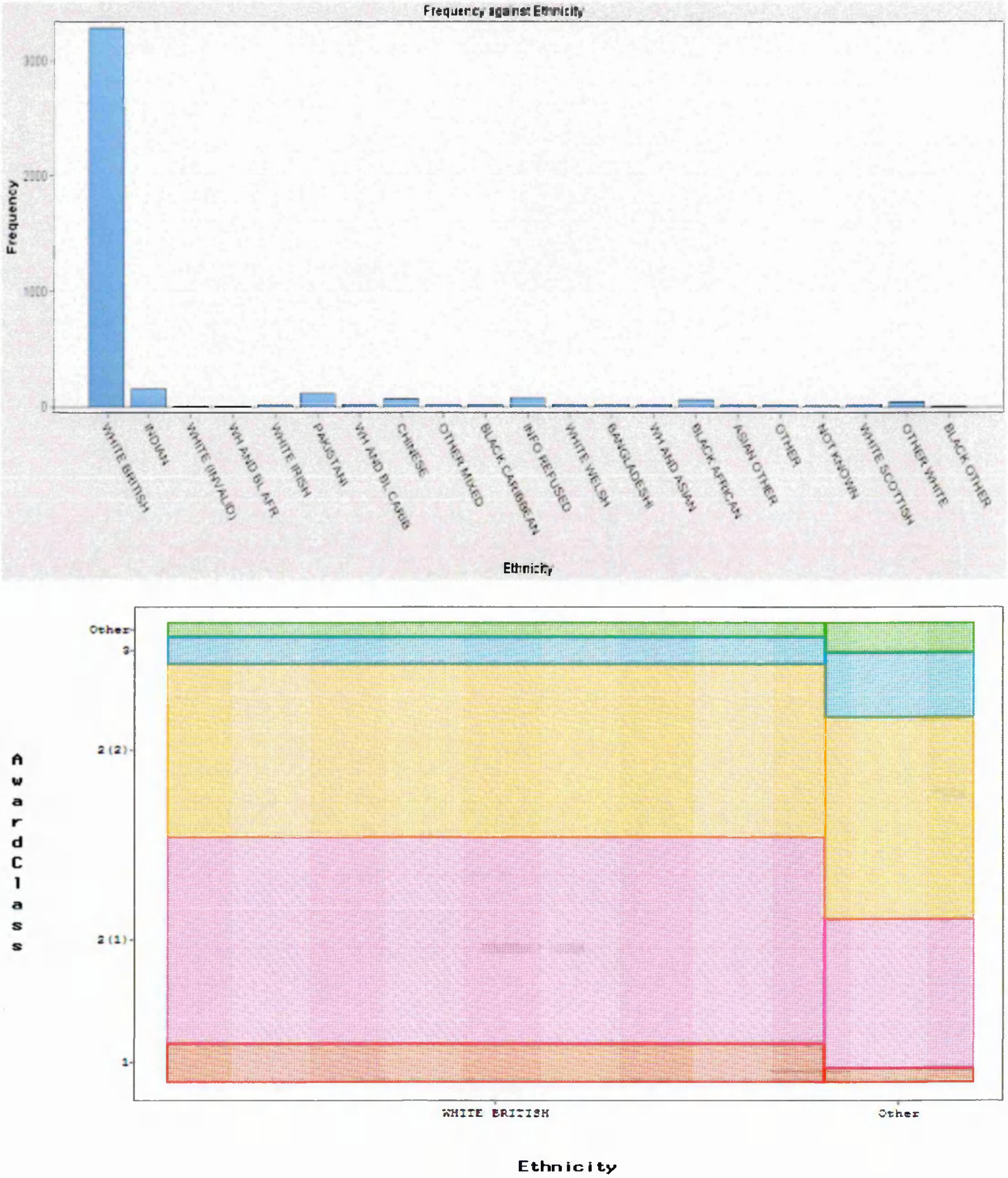


Figure 6.36 - Ethnicity Bar Chart and Mosiac Plot.

Again Ethnicity appears to have too many overall categories, due to the large number of different ethnic groups. The mosaic plot indicates that White British students tend to achieve a better classification than the others.

ENTRY QUALIFICATIONS

Students EntryQualifications are plotted against Frequency and AwardClass below.

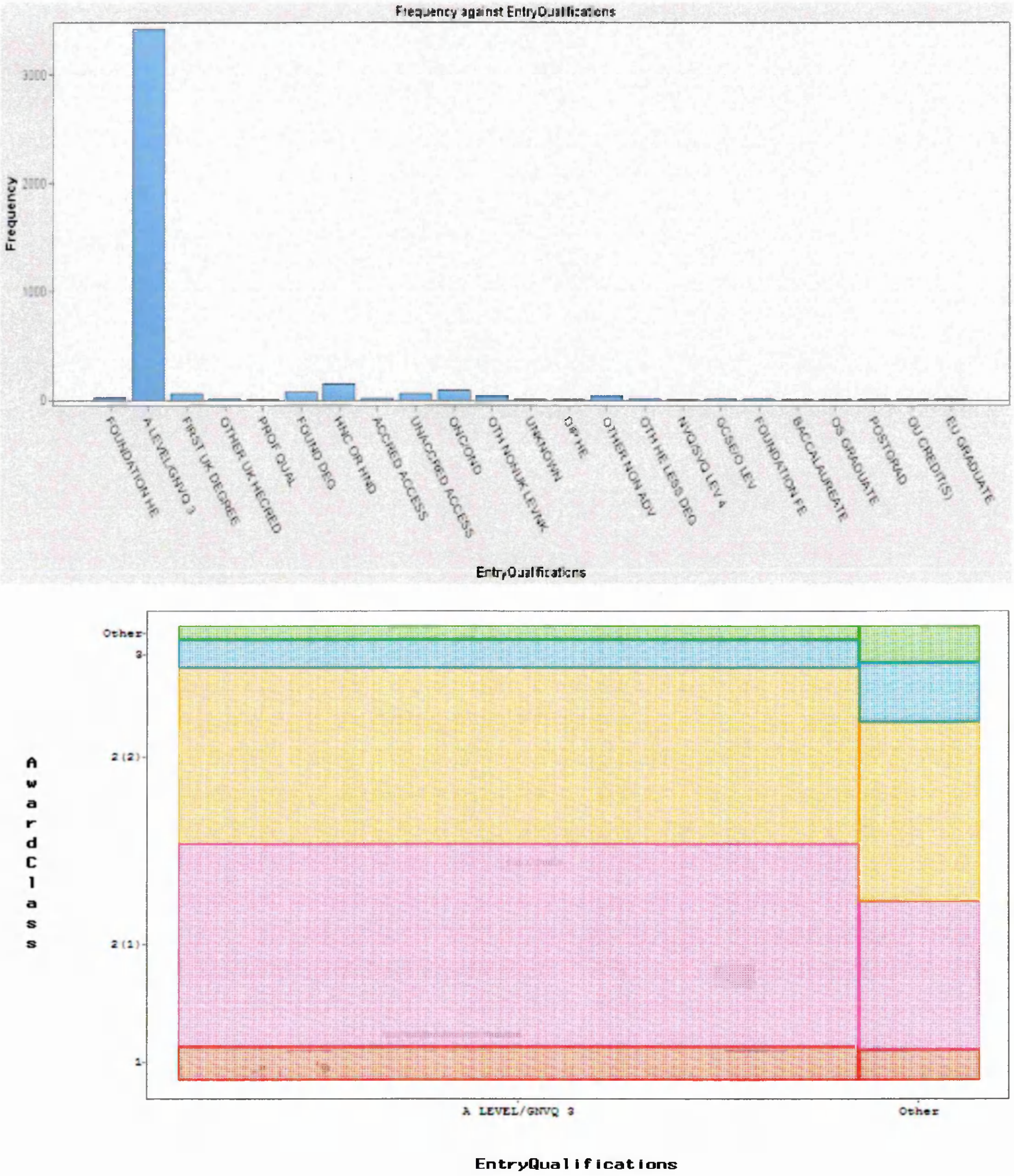


Figure 6.37 – Entry Qualifications Bar Chart and Mosiac Plot.

Arguably, there are too many different EntryQualifications and there is a large proportion of A-Level/GNVQ 3, which provides some justification for the large proportion of students entering their undergraduate degrees between the ages of 17 and 22. The mosiac plot suggests that students who enter their degrees with A-Level/GNVQ 3 qualifications tended to achieve a better award classification than those who entered with other qualifications. These students obtained more 1st and 2:1 classifications whereas those students who entered their degrees with other qualifications tended to achieved more 3rd class honours then those with A-Level/GNVQ 3 qualifications.

ENTRY POINTS

Students EntryPoints are plotted against Frequency and AwardClass below.

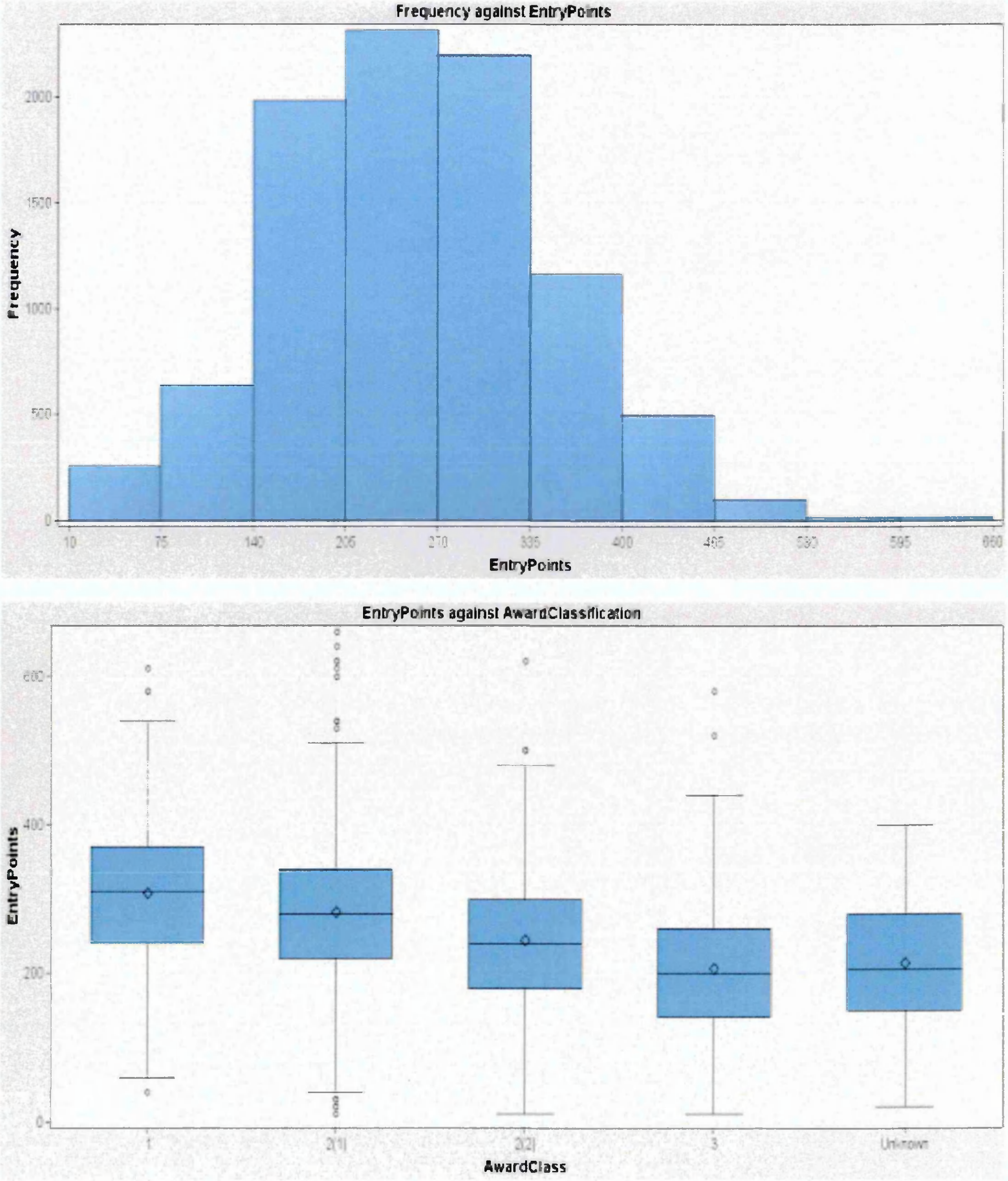


Figure 6.38 – Entry Points Bar Chart and Box Plot.

This highlights that the majority of students entered their degrees with between 205 and 270 entry points - potential outliers include 10-75 and 530-660. In addition to this, there is an association between student entry points and award classification. It is perhaps prudent to note that going forward the EntryPoints groupings will change due to the introduction of the A* A-Level and any models built may have to be changed to reflect this. As stated above, there is also a relationship between EntryPoints and EntryQualifications variable. However, the box plot above indicates that there is a relationship between the number of entry points and final award classification.

COURSE

The Course variable is plotted against Frequency below.

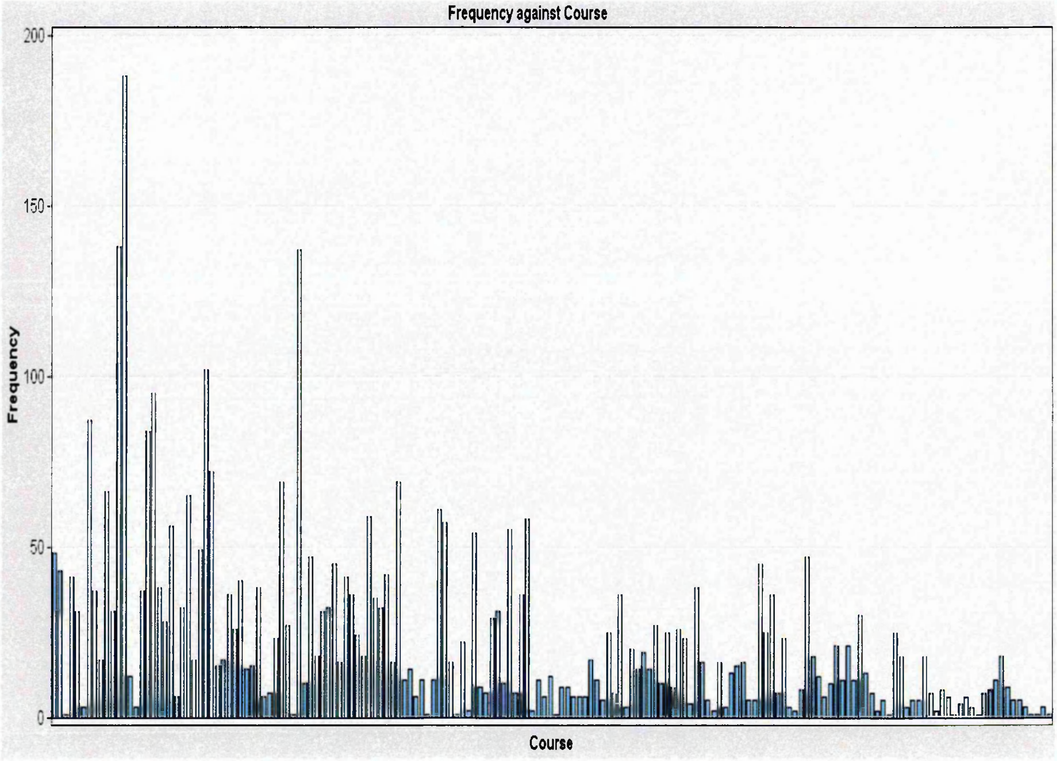


Figure 6.39 – Course Bar Chart.

There are too many categories to the Course variable and its inclusion in the modelling stage is questionable. Therefore, a method for grouping this variable or the inclusion of other variables (such as Faculty or JACS Subject) should be considered instead of the Course variable.

FACULTY

The graphs below are used to assess the Faculty variable against Frequency, EntryPoints and AwardMark.

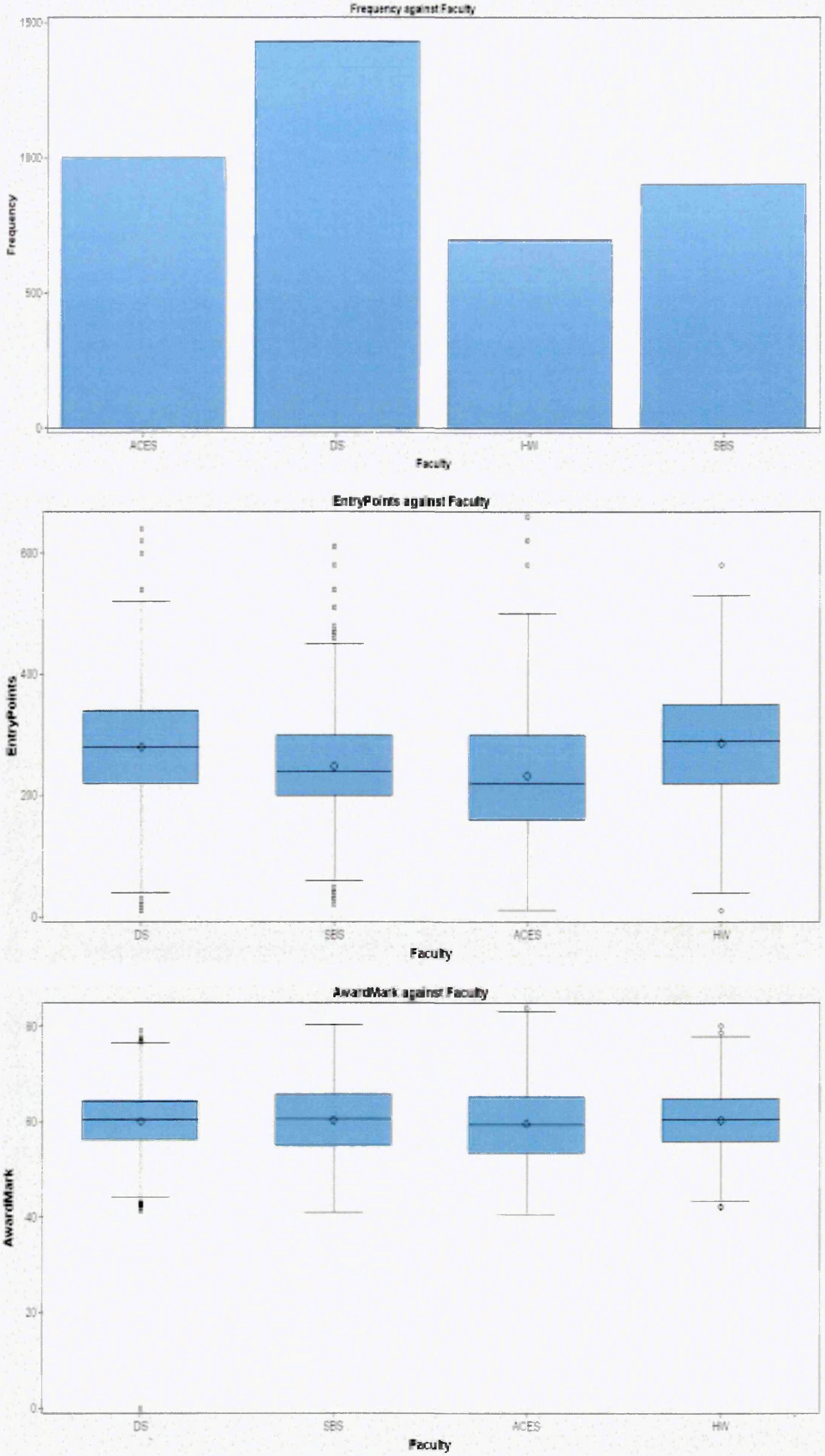


Figure 6.40 – Faculty Bar Chart and Box Plots.

The Faculty variable has four levels ACES, DS (Development and Society), HW (Health and Wellbeing) and SBS (Organisation and Management). Looking at EntryPoints and AwardMark against Faculty graphs, it is apparent that the average entry points in ACES tends to be lower which is refelected in the average AwardMark. Although, the students in the ACES faculty have also achieved higher individual award marks than students in any other faculty. The mosiac plot of Faculty against AwardClassification, below, confirms that ACES students tend to get more 1st and 3rd class classifications than any other faculty and that DS has the largest proportion of students and they tend too achieve more 2:1 and 2:2 classifications.

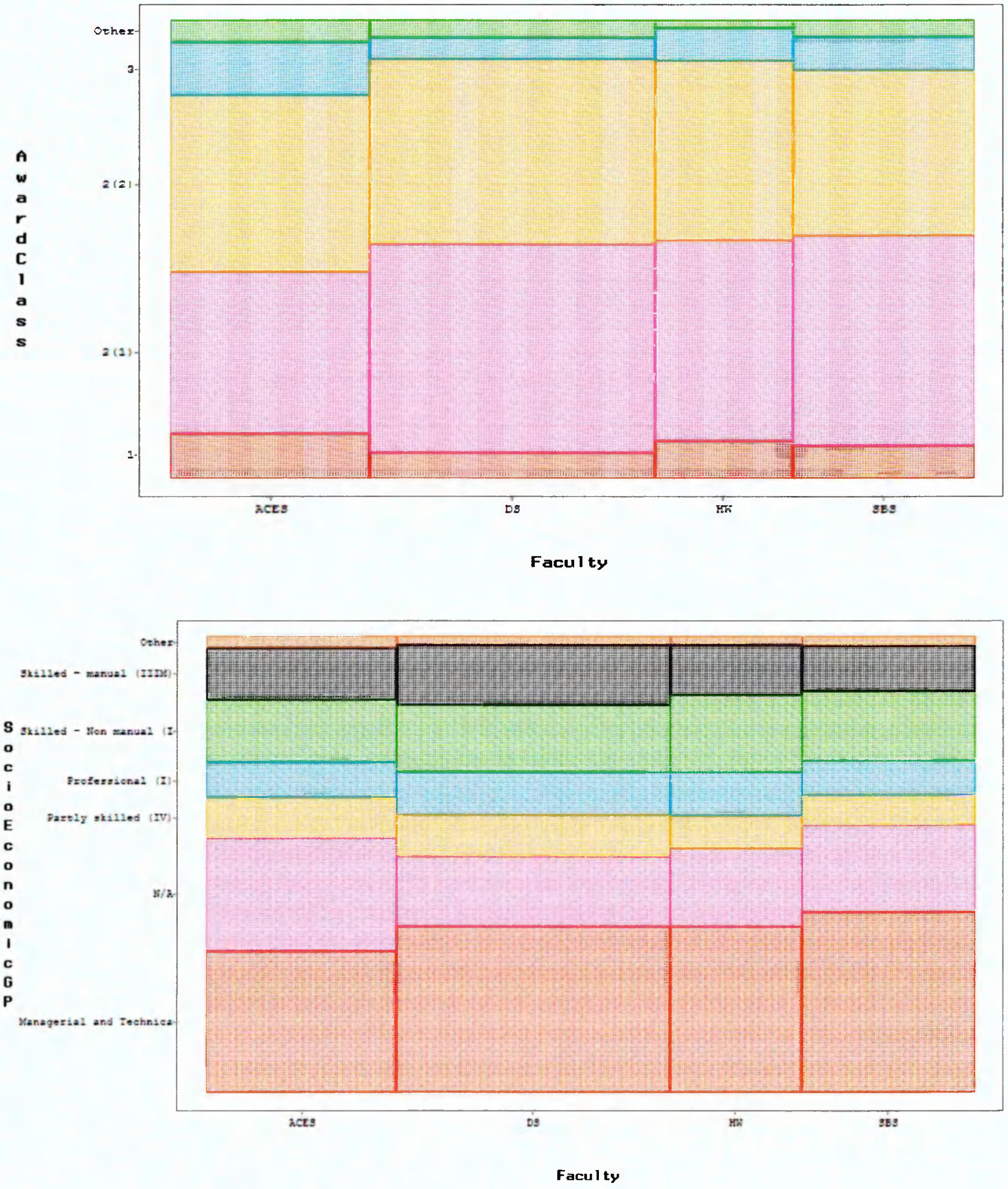
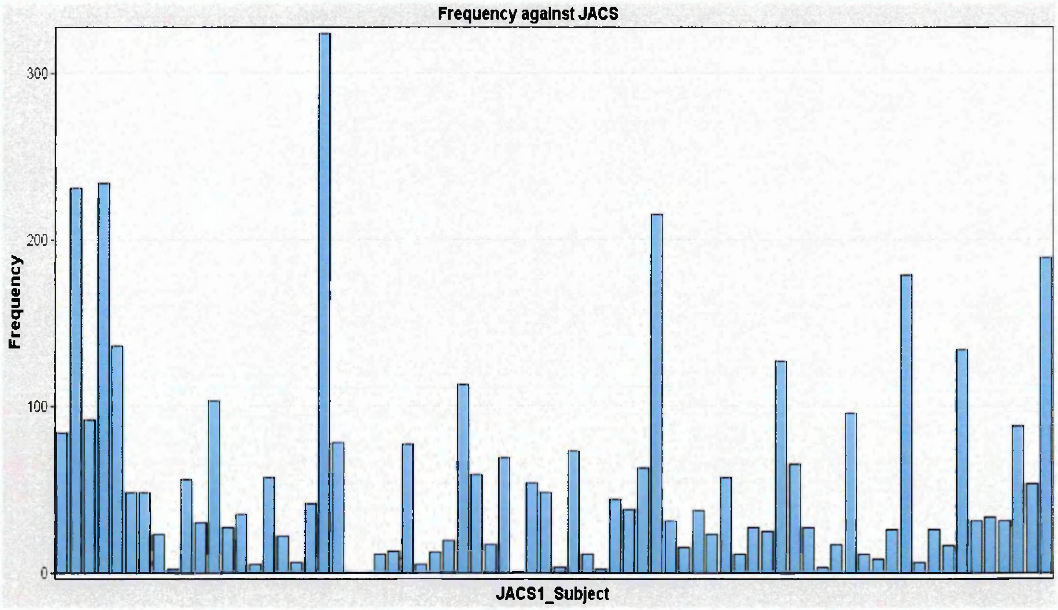


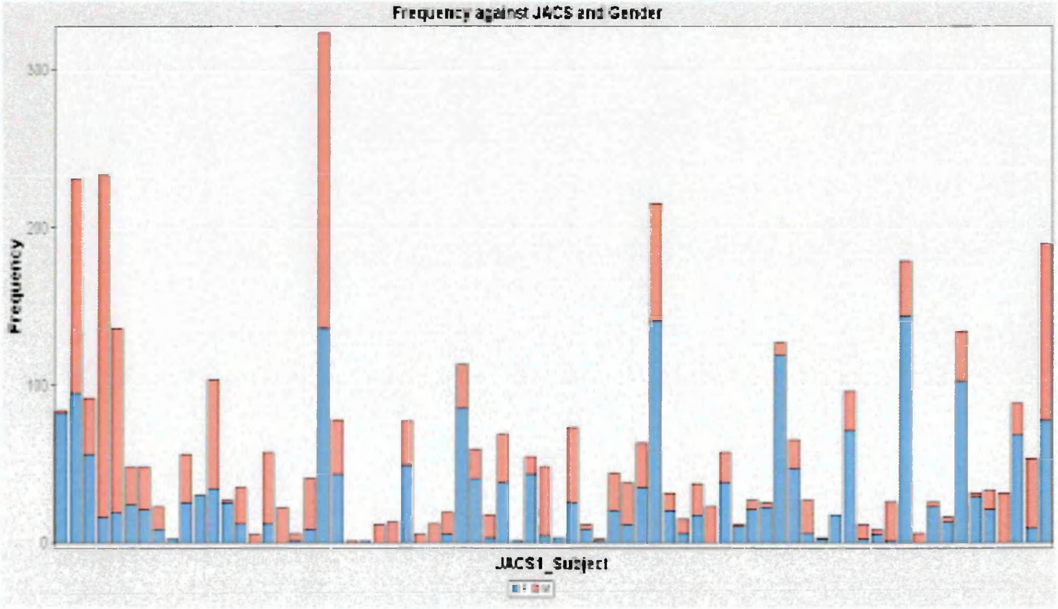
Figure 6.41 – Faculty Mosiac Plots.

In terms of SocioEconomicGP SBS tends to have more students from a technical and managerial socioeconmic group and DS and HW also have the highest proportion of students from a professional socioeconmic group. The inclusion of this variable at the modelling stage is questionable as there are probably too few levels in that any models built using this variable could be to general.

JACS – Subject

The graphs, below, show JACS Subject against Frequency, Gender and AwardClass.





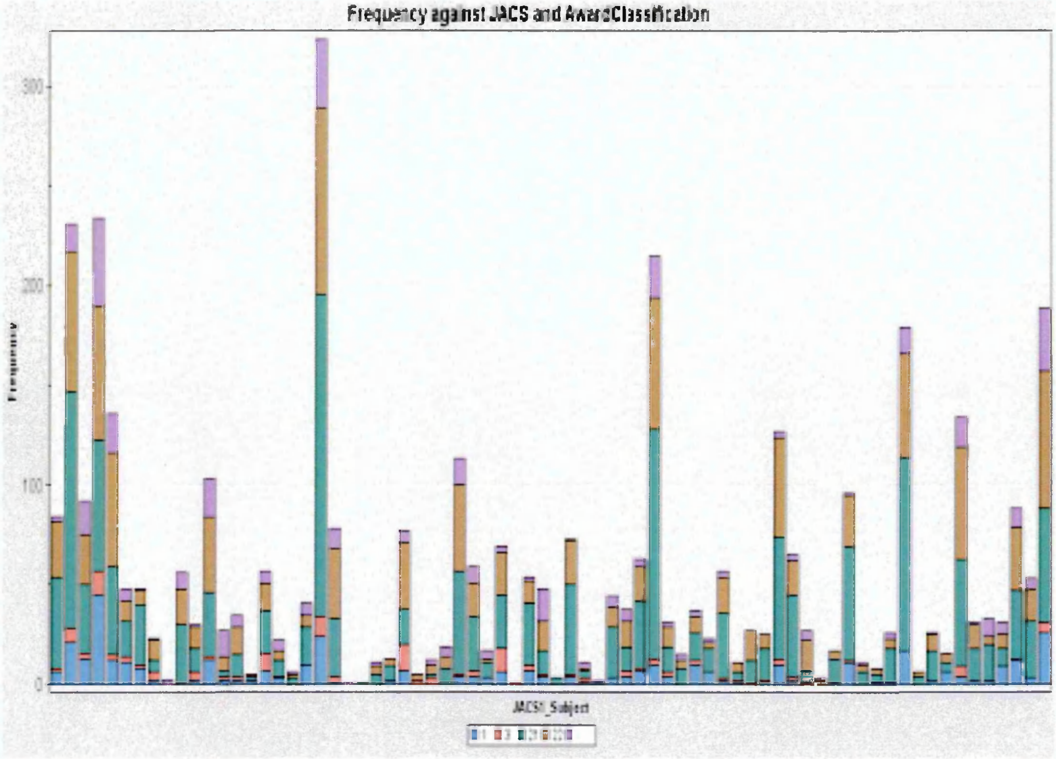


Figure 6.42 – JACS Subject Bar Charts.

The JACS Subject variable is a half-way house between Course and Faculty. However, there still appears to be a large number of overall levels. The graphs show that some of the JACS subjects are mainly taken by males and others females. In addition to this, they also highlight that students in some JACS subjects appear to achieve a better overall award classification then students in others

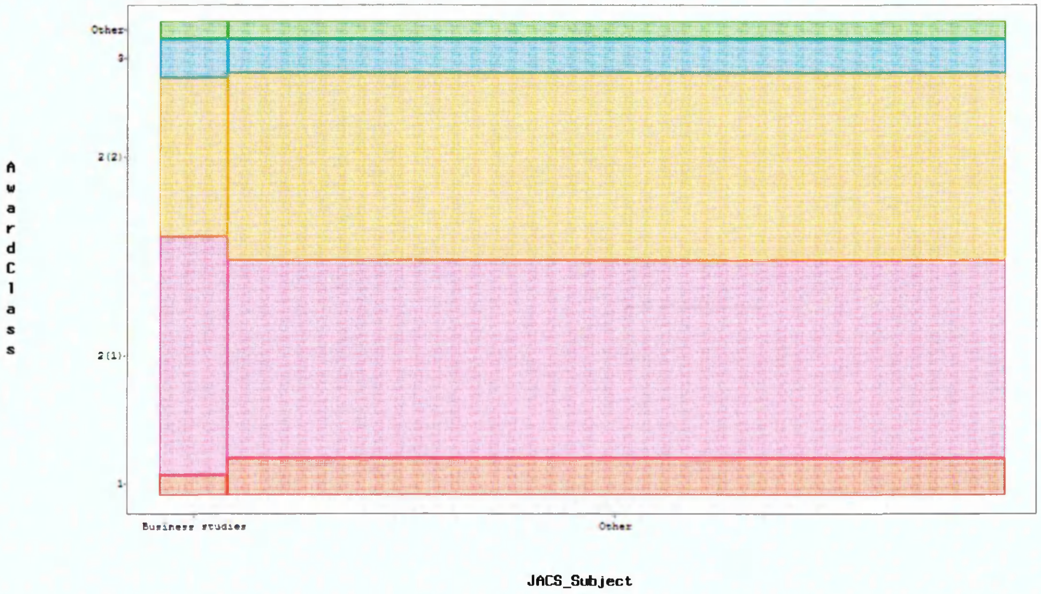
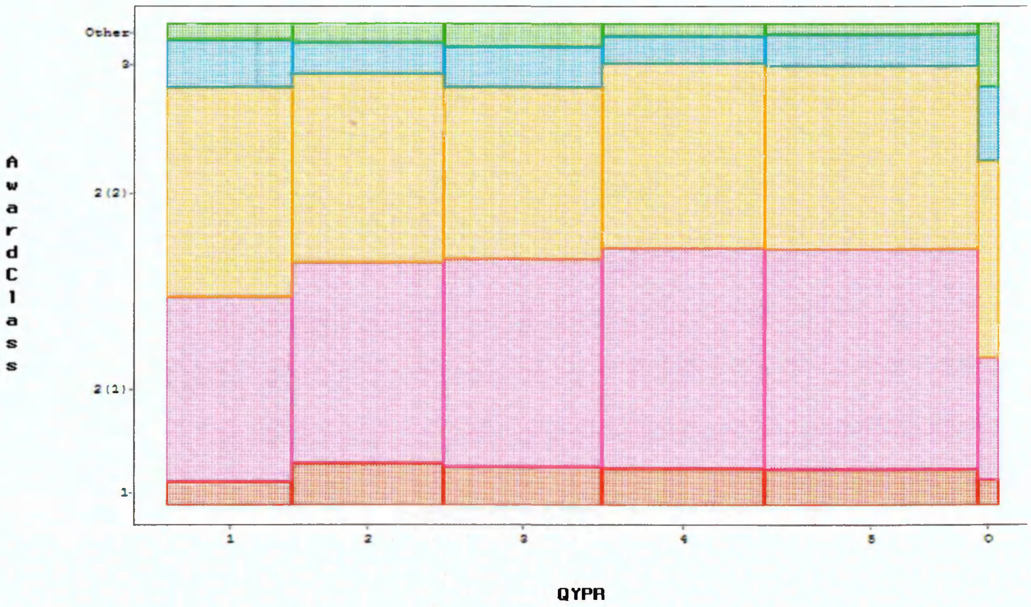


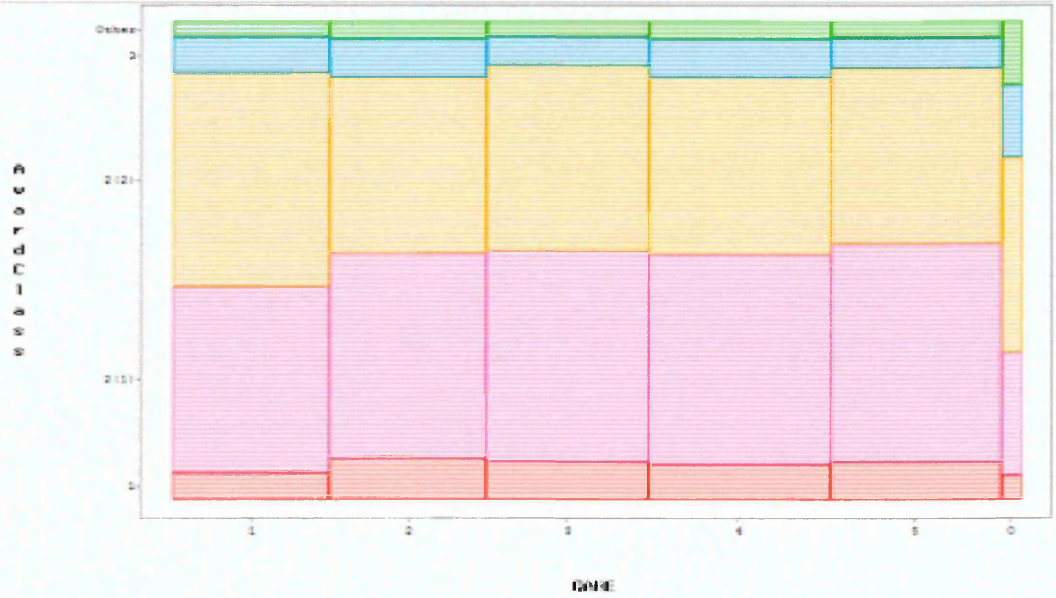
Figure 6.43 – JACS Subject Mosiac Plot.

Indeed, the mosaic plot, above, of AwardClassification against JACS Subject singles out the business studies subject which shows that students in this group are more likely to achieve a 2:1 and a 3rd than in any other subject. It also shows that, in relation to other subjects, students taking business studies are less likely to achieve a 1st classification.

Socioeconomic Groupings

AwardClass is plotted against QYPR, QAHE and SocioEconomicGP, below.





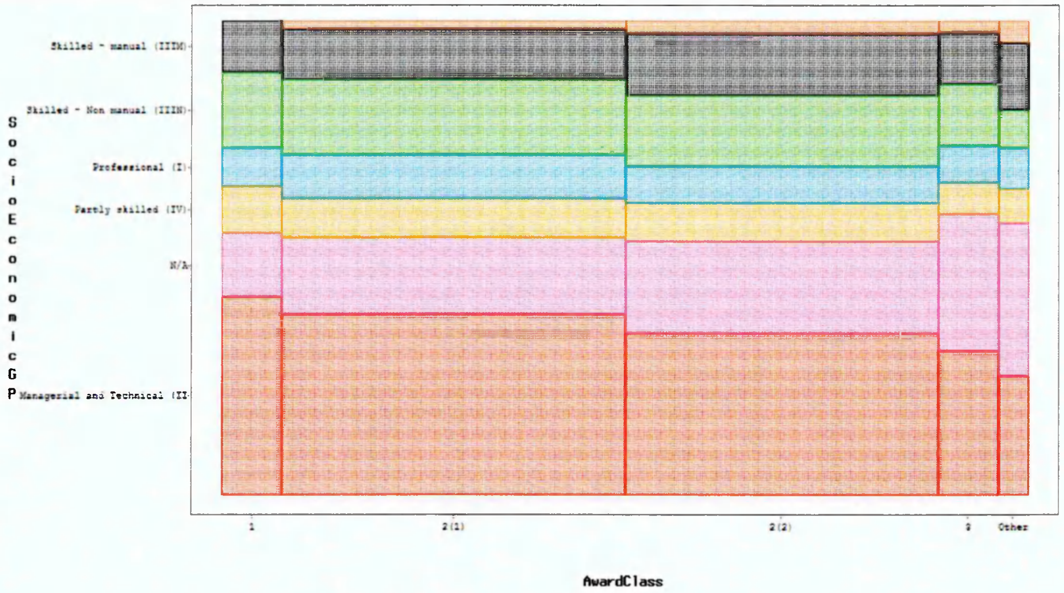


Figure 6.44 – Socioecominic Groupings Mosiac Plots.

Participation of Young People in HE by Postcode (QYPR) - the mosiac plot of AwardClassification against QYPR indicates that more students were from areas where QYPR was rated has a 5 (high), this group had the largest proportion of 2:1 classifications. Those student with a QYPR of 2 (low-medium) have more students who obtain a 1st classification, this is likely to be due to the non-traditional type of students catered for by SHU.

Adults with HE Qualifications by Postcode (QAHE) - the mosiac plot of AwardClassification against QAHE suggests that students from areas where QAHE was 1 (low) had more students obtaining a 2:2 and where QAHE was 2 (low-medium) there were more 1st and 3rd classifcations. The largest group appears to be where QAHE is 4 (medium-high).

Socioeconmic Group of parents - the mosaic plot of AwardClassification against SocioEconmicGP highlights that more students are from the managerial and technical group. Students in this group tend to achieve a classified degree, from across the entire degree classification spectrum, than any other group. Students from the skilled groups tend to obtain more 2:1 and 2:2 awards.

The inclusion of all three variables (QYPR, QAHE, SocioEconmicGP) should be assessed at the DM stage.

GENDER

Gender is plotted against AwardClasss, EntryQualifications and SocioEconomicGP, below.

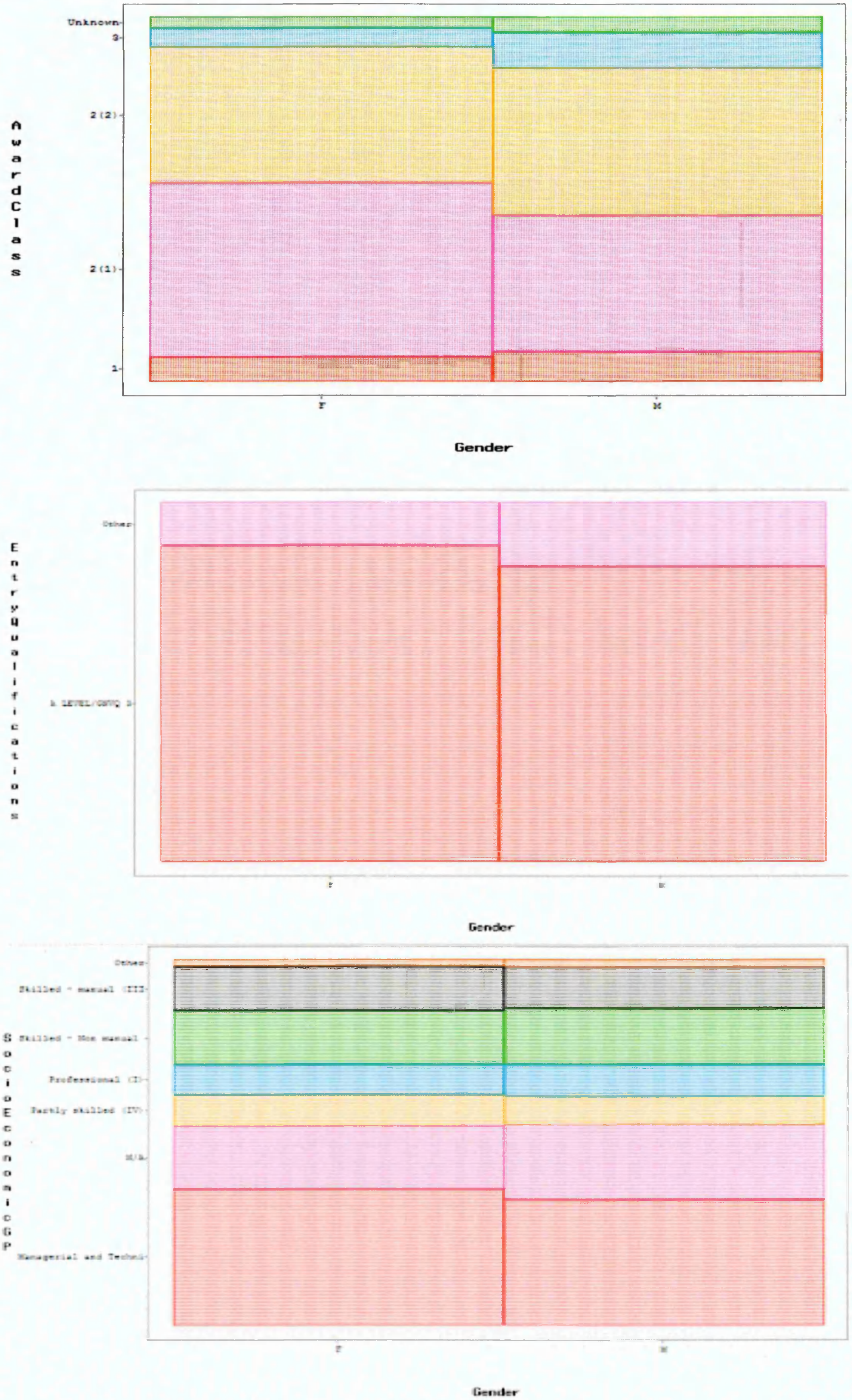


Figure 6.45 – Gender Mosiac Plots.

These graphs show that there are more females than males in the dataset. They also show that males tend to achieve more 1st, 2:2 and 3rd class classifications and females achieve more 2:1 classifications. The second mosaic plot indicates that more females enter their degrees with A-Level/GNVQ 3 qualifications then males. The final mosaic plot suggests that there are more females from the socioeconomic group of managerial and technical than males and there are more males from the socioeconmic group skilled non-manual than females.

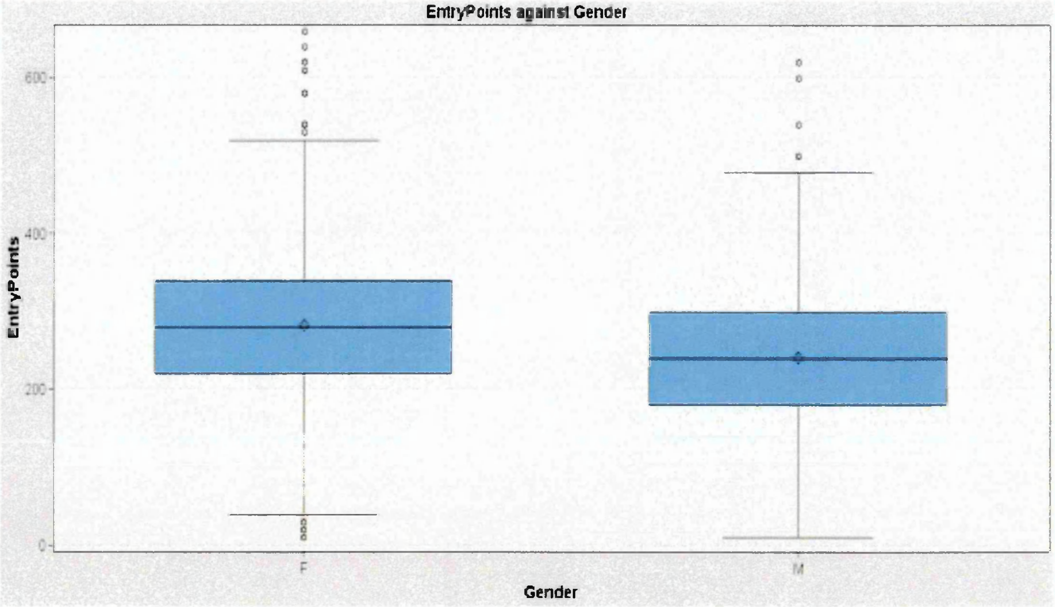


Figure 6.46 – Gender Box Plot.

The above box plot shows that the average EntryPoints for females is higher than males.

DISABILITY

Disability is plotted against AwardClass, Gender and Faculty, below.

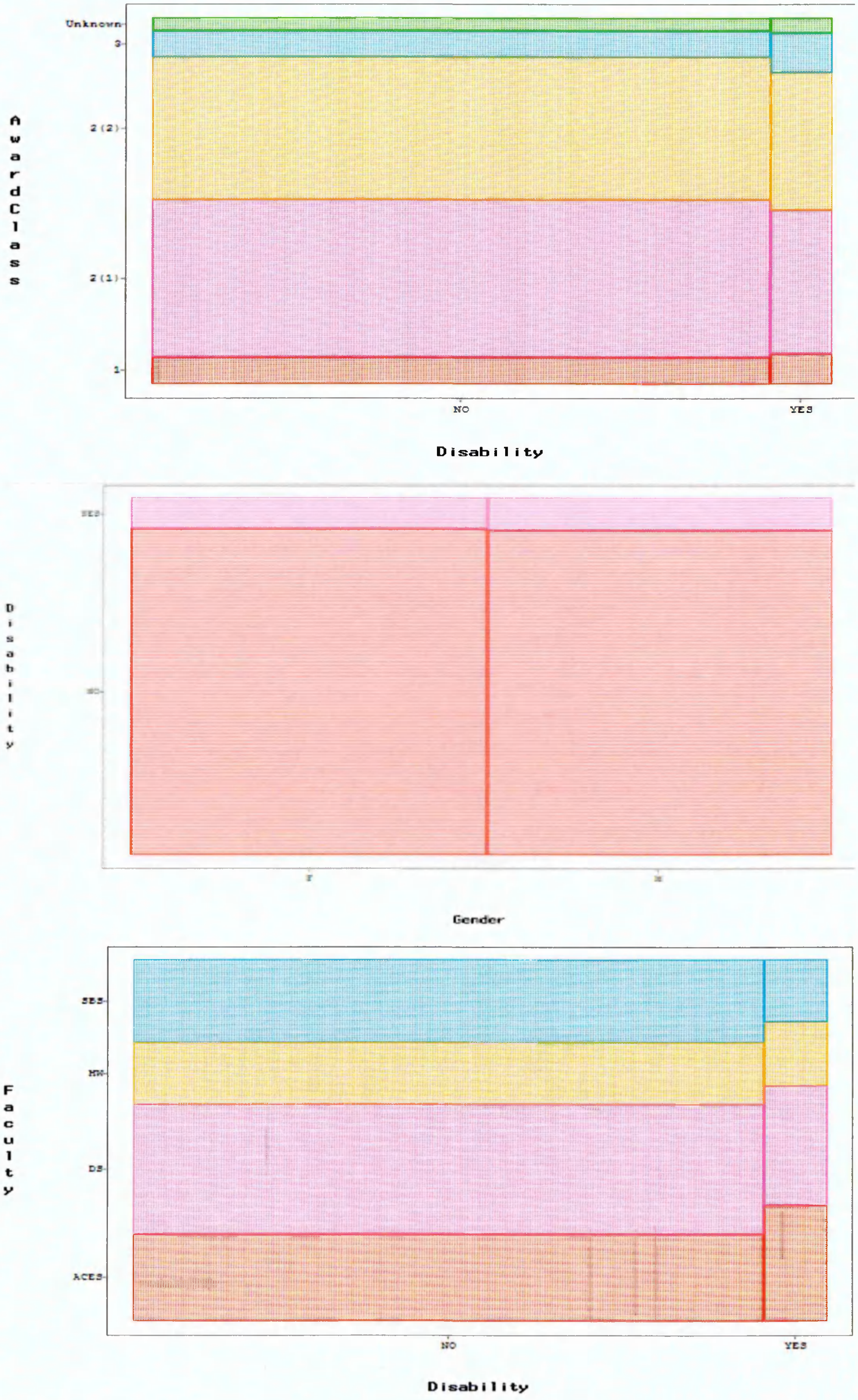


Figure 6.47 – Disability Mosiac Plots.

The disability graphs, above, show that there are significantly more non-disabled students in the dataset and that disabled students tend to obtain more 1st and 3rd class classifications. There are slightly more disabled males in the data set than disabled females and the faculty of ACES tends to have more disabled students. The inclusion of Disability in the final model is questionable, given the small amount of disabled students in the dataset.

LEA

LEA is plotted against Frequency and AwardClass, below.

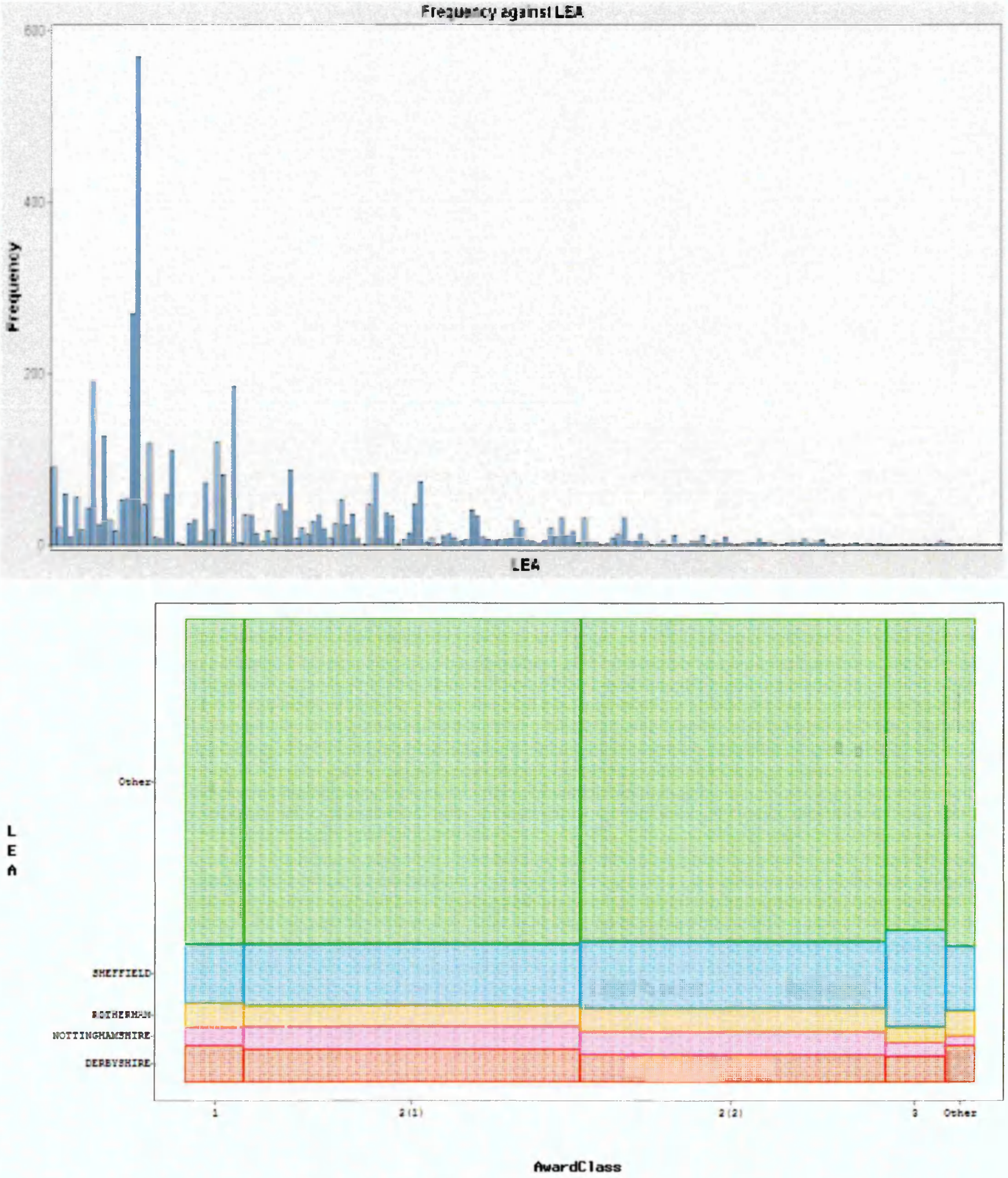
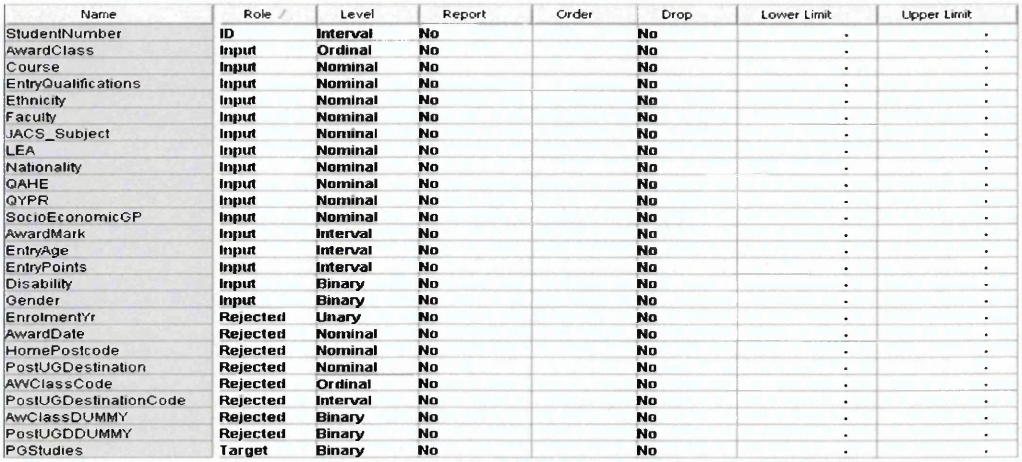


Figure 6.48 – LEA Bar Chart and Mosiac Plot.

LEA as too many categories and there are other location based variables (QYPR, QAHE and SocioEconomicGP. Therefore, this variable won't be discussed any further in the proceeding sections.

3. Postgraduate Studies

Using PGStudies as the target variable (rejecting the other two target variables (AwardClass and PostUGDestination) the list below shows the variables to be assessed.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
StudentNumber	ID	Interval	No		No	.	.
AwardClass	Input	Ordinal	No		No	.	.
Course	Input	Nominal	No		No	.	.
EntryQualifications	Input	Nominal	No		No	.	.
Ethnicity	Input	Nominal	No		No	.	.
Faculty	Input	Nominal	No		No	.	.
JACS_Subject	Input	Nominal	No		No	.	.
LEA	Input	Nominal	No		No	.	.
Nationality	Input	Nominal	No		No	.	.
QAHE	Input	Nominal	No		No	.	.
QYPR	Input	Nominal	No		No	.	.
SocioEconomicGP	Input	Nominal	No		No	.	.
AwardMark	Input	Interval	No		No	.	.
EntryAge	Input	Interval	No		No	.	.
EntryPoints	Input	Interval	No		No	.	.
Disability	Input	Binary	No		No	.	.
Gender	Input	Binary	No		No	.	.
EnrolmentYr	Rejected	Unary	No		No	.	.
AwardDate	Rejected	Nominal	No		No	.	.
HomePostcode	Rejected	Nominal	No		No	.	.
PostUGDestination	Rejected	Nominal	No		No	.	.
AWClassCode	Rejected	Ordinal	No		No	.	.
PostUGDestinationCode	Rejected	Interval	No		No	.	.
AwClassDUMMY	Rejected	Binary	No		No	.	.
PostUGDDUMMY	Rejected	Binary	No		No	.	.
PGStudies	Target	Binary	No		No	.	.

Figure 6.49 – Postgraduate Studies Variables and Roles.

The role and levels of these variables are:

Role	Level	No. of variables
ID	Interval	1
Input	Nominal	10
Input	Ordinal	1
Input	Binary	2
Input	Interval	3
Rejected	Binary	2
Rejected	Nominal	4
Rejected	Ordinal	1
Rejected	Unary	1
Target	Nominal	1
Total No. Variables		**26**

Figure 6.50 – Postgraduate Studies Roles, Levels and Counts.

Rejected variables include:

- AWClassCode - relates to the different award classes for another model;
- AWClassDummy - used to determine potential groups for another model;
- AwardDate - the date that the student received the award this has no significance to the AwardClass received by the student;
- PostUGDUMMY - used to determine potential groups for another model;
- PostUGDestination - the target variable for another model;
- PostUGDestinationCode - relates to the different postgraduate destinations for another model;
- EnrolmentYr - a unary variable;
- HomePostcode - there are other location based variables in the dataset.

The PGStudies variable is given the role of target for the data understanding and modelling stage, as this is a binary target variable there is no need to create any dummy variables.

The distribution of the data, shown below, is exactly the same as previously discussed. Therefore this section will look at the individual variables in relation to the PGStudies target variable.

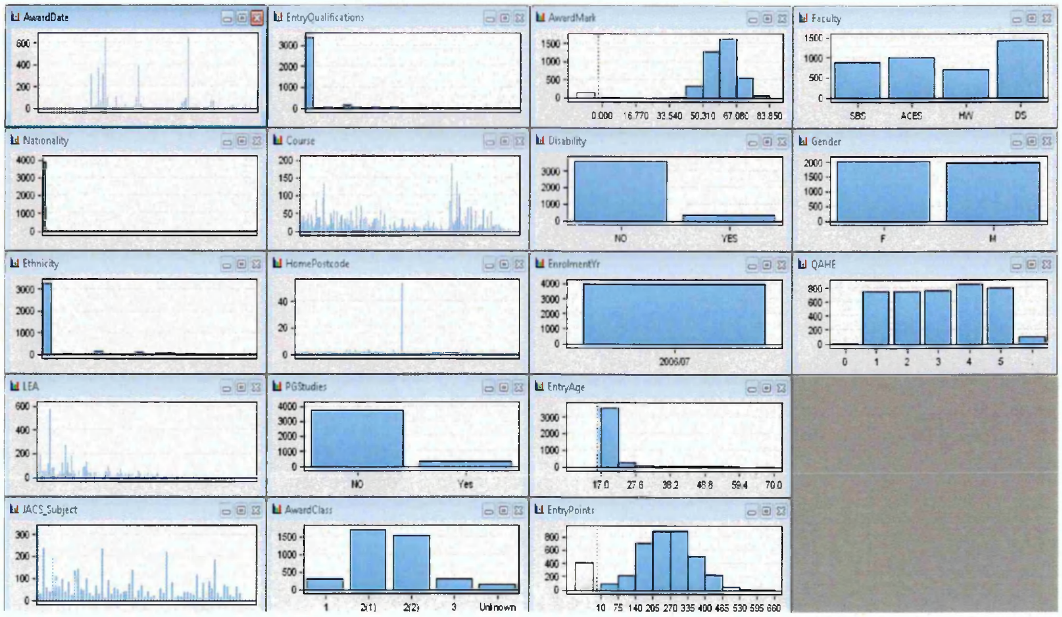


Figure 6.51 – Postgraduate Studies Initial Histograms.

As highlighted in the honours award classificaiton section above there are a number of variables that have too many overall levels. These will be addressed through assessing the variables/values using variable selection, categorical input consolidation and interactive binning/grouping in the building the models section.

Postgraduate Studies

The frequency of the PGStudies variable is plotted below.

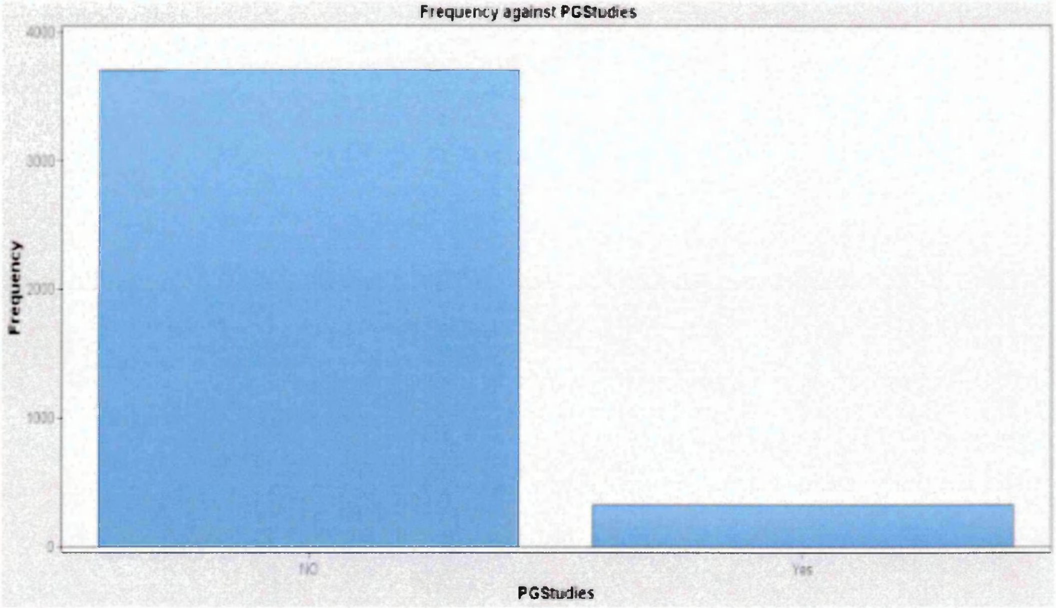
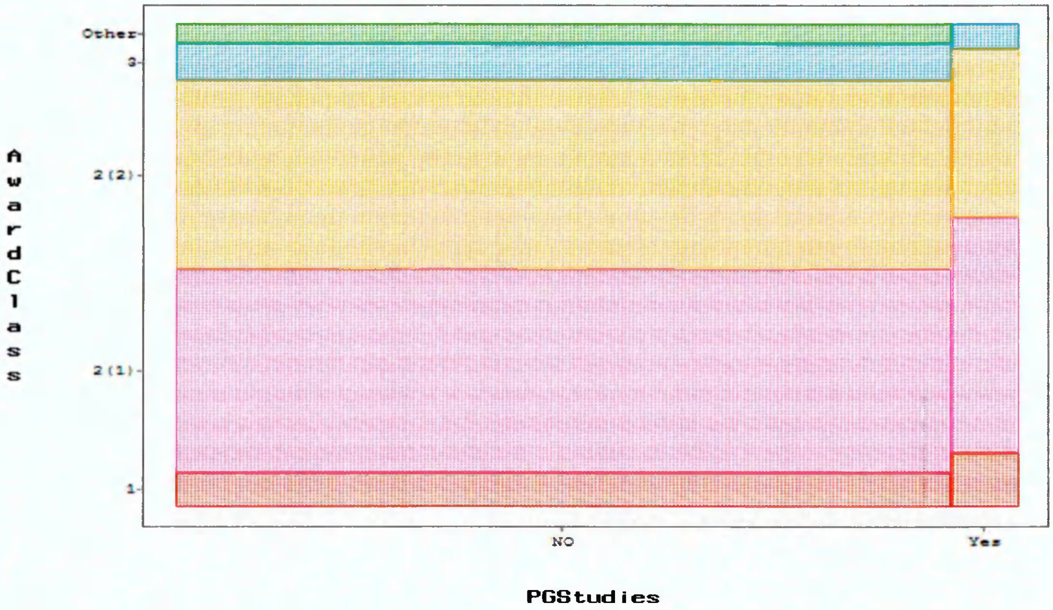


Figure 6.52 – Postgraduate Studies Bar Chart.

Of the 4023 students who completed their degrees in 2006/07 322 went onto postgraduate studies in 2007/08, 2008/09 2009/10 - at the point when the data was sampled from the SI system. Due to the large number of students who didn't go on to undertake a postgraduate degree during this time the data will have to be oversampled, so that the 'Yes' value is not overshadowed by the large amount of 'No' values - as a result prior probabilities will have to be built before the data is oversampled.

Award Classification and Award Mark

Below is a plot of the student's award classifications and award marks.



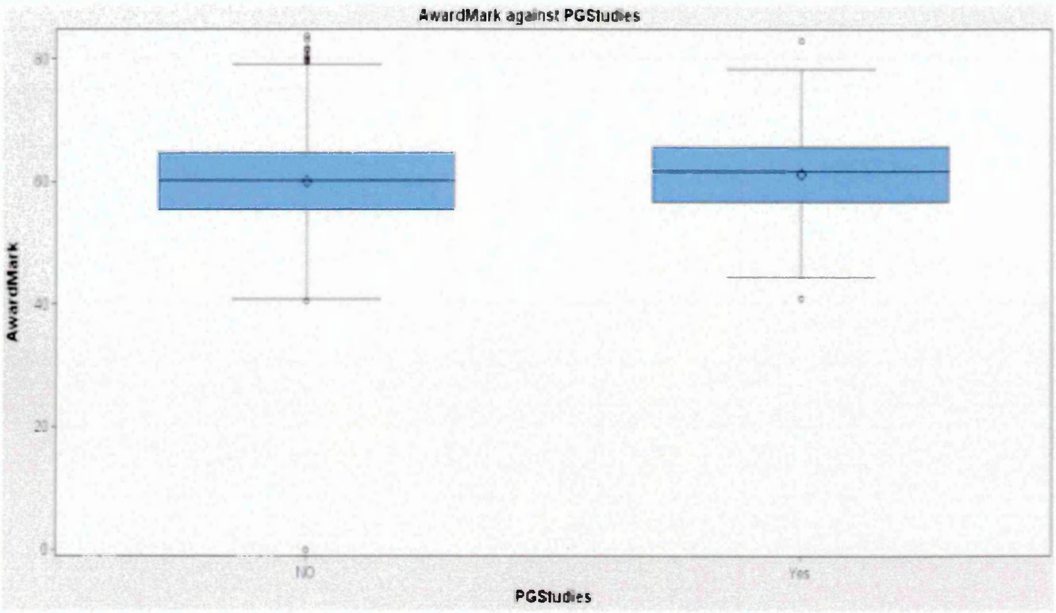


Figure 6.53 – Award Classification and Mark Mosiac Plot and Bar Chart.

Award Classification - the mosaic plot of PGStudies against AwardClassification shows that the most common AwardClassification obtained by students who go on to take a postgraduate degree is a 2:1. In addition to this, the above graph highlights that more students who obtained a 1st in there undergraduate degree are more likely to go on to undertake postgraduate studies. The inclusion of this variable in the modelling stage is questionable as AwardMark is used to calculate the AwardClassification.

Award Mark - the box plot of AwardMark against PGStudies shows that the average AwardMark of students who go on to undertake postgraduate studies is higher than those who don't. The AwardMark variable will be included at the modelling stage in favour of AwardClassification as it is used to calculate the classification.

Entry Age

The box plot, below, shows student entry age against whether or not students went on to take PGStudies at SHU.

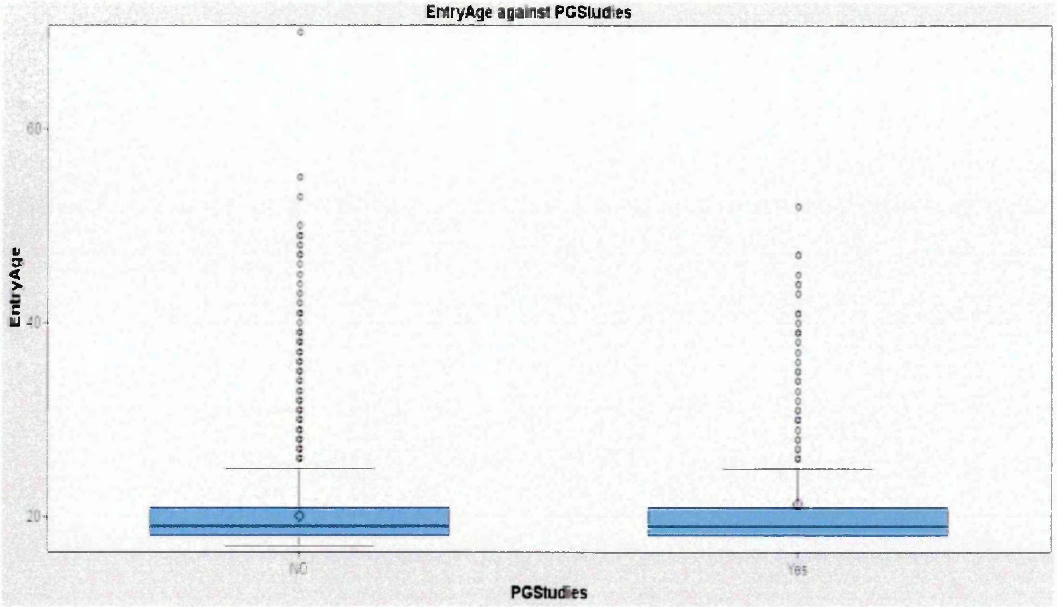
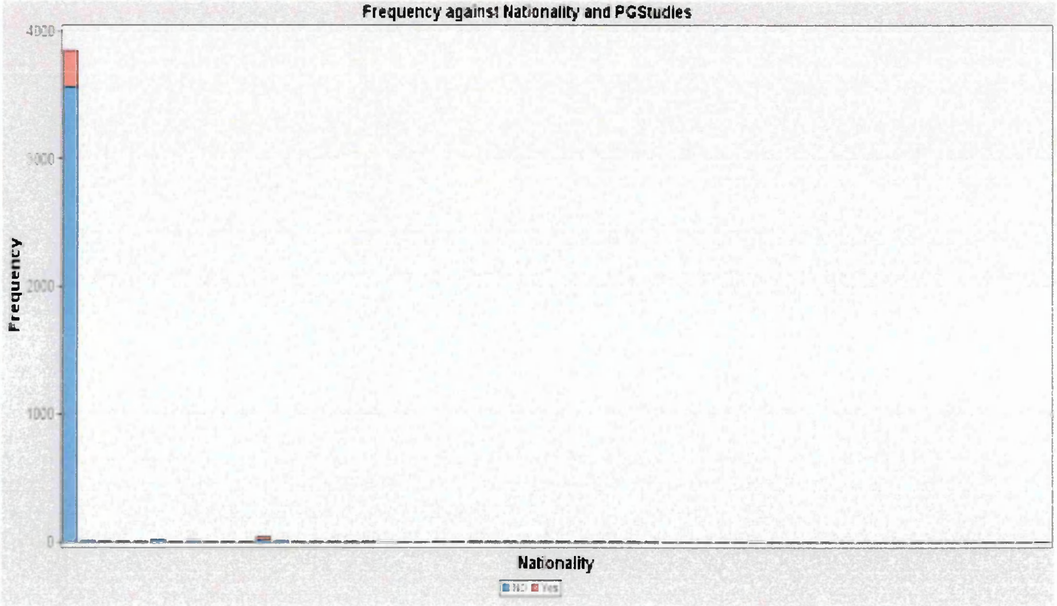


Figure 6.54 – Entry Age Box Plot.

The above box plot of EntryAge against PGStudies shows that the average age of those students who entered their original undergraduate degree and went onto postgraduate studies was less than those who didn't go on to take a postgraduate qualification.

Nationality and Ethnicity

The frequency of Nationality and Ethnicity are plotted below.



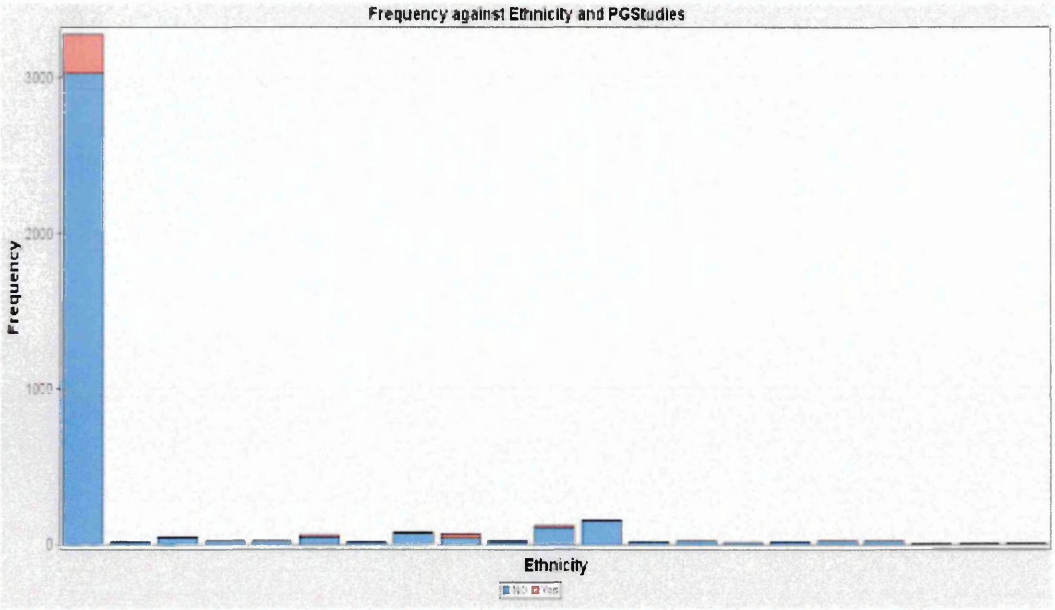


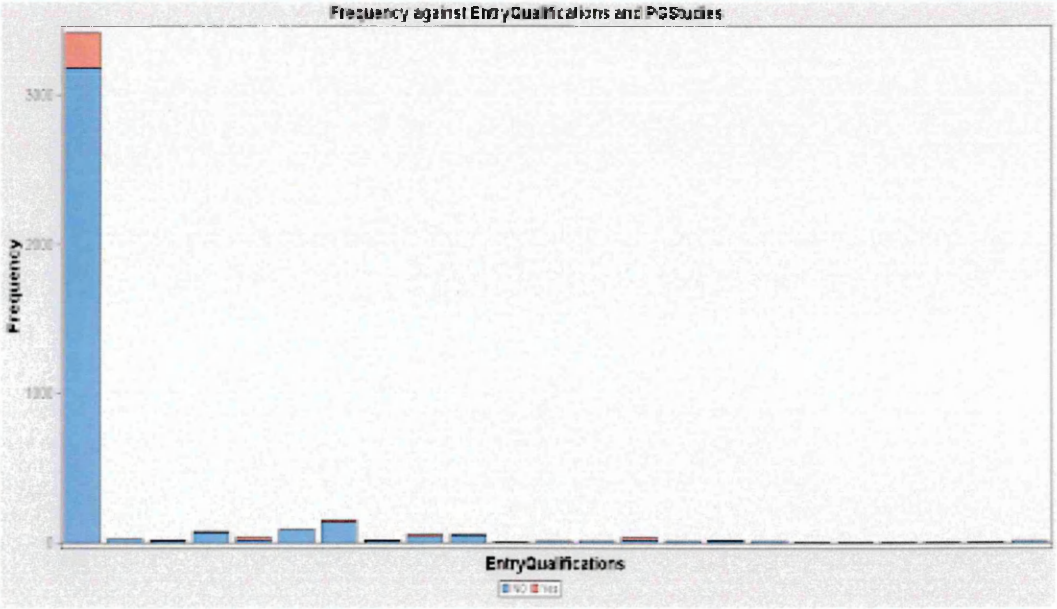
Figure 6.55 – Nationality and Ethnicity Bar Charts.

Nationality - again there are too many categories to the <u>Nationality</u> variable. The majority of the students who went on to study a postgraduate degree were from the UK.

Ethnicity - again <u>Ethnicity</u> appears to have too many categories, due to the large number of different ethnic groups. The graph indicates that more White British students tend to go on to take postgraduate studies.

Entry Qualifications and Entry Points

Progression onto <u>PGstudies</u> at SHU is plotted against the students <u>EntryQualifications</u> and <u>EntryPoints</u> below.



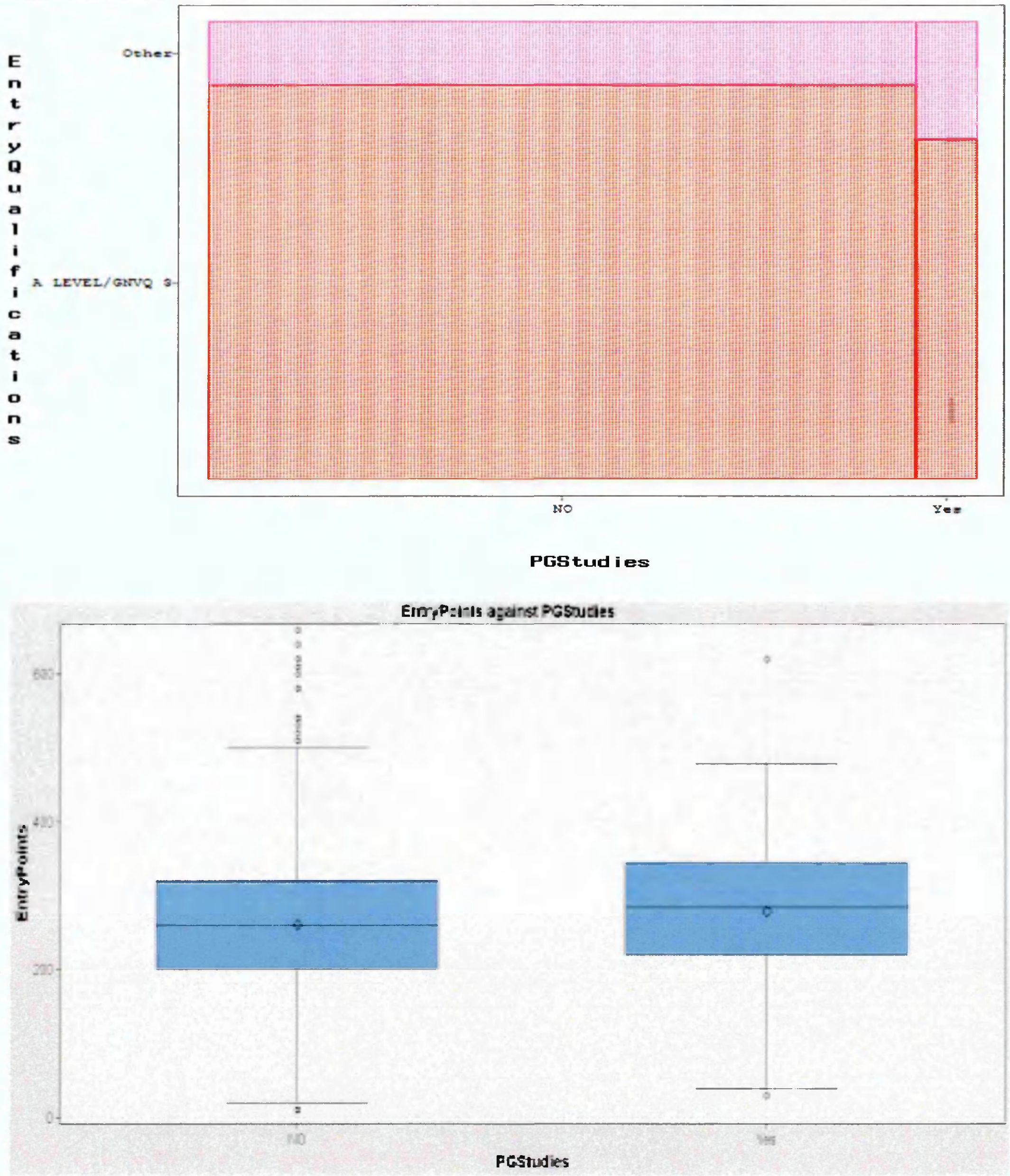


Figure 6.56 – Entry Qualifications and Points Bar Chart, Mosiac Plot and Box Plot.

Entry Qualifications - there are too many different EntryQualifications and the largest proportion of students who went on to take postgraduate studies entered university with A-Level/GNVQ 3.

Entry Points - The box plot above shows that students who go on to study a postgraduate degree tend to enter their undergraduate degrees with higher average EntryPoints than those who don't go on to take a postgraduate degree. As before it is perhaps prudent to note that going forward the EntryPoints groupings will change due to the introduction of the A* A-Level and any models built may have to be changed to reflect this. As stated above, there is also a relationship between EntryPoints and EntryQualifications variable. However, the box plot above suggests that there is a relationship between the number of entry points and those students who go on to take postgraduate studies at SHU.

COURSE, FACULTY AND JACS SUBJECT

This section considers Course, Faculty and JACS Subject in relation to the target variable of PGStudies.

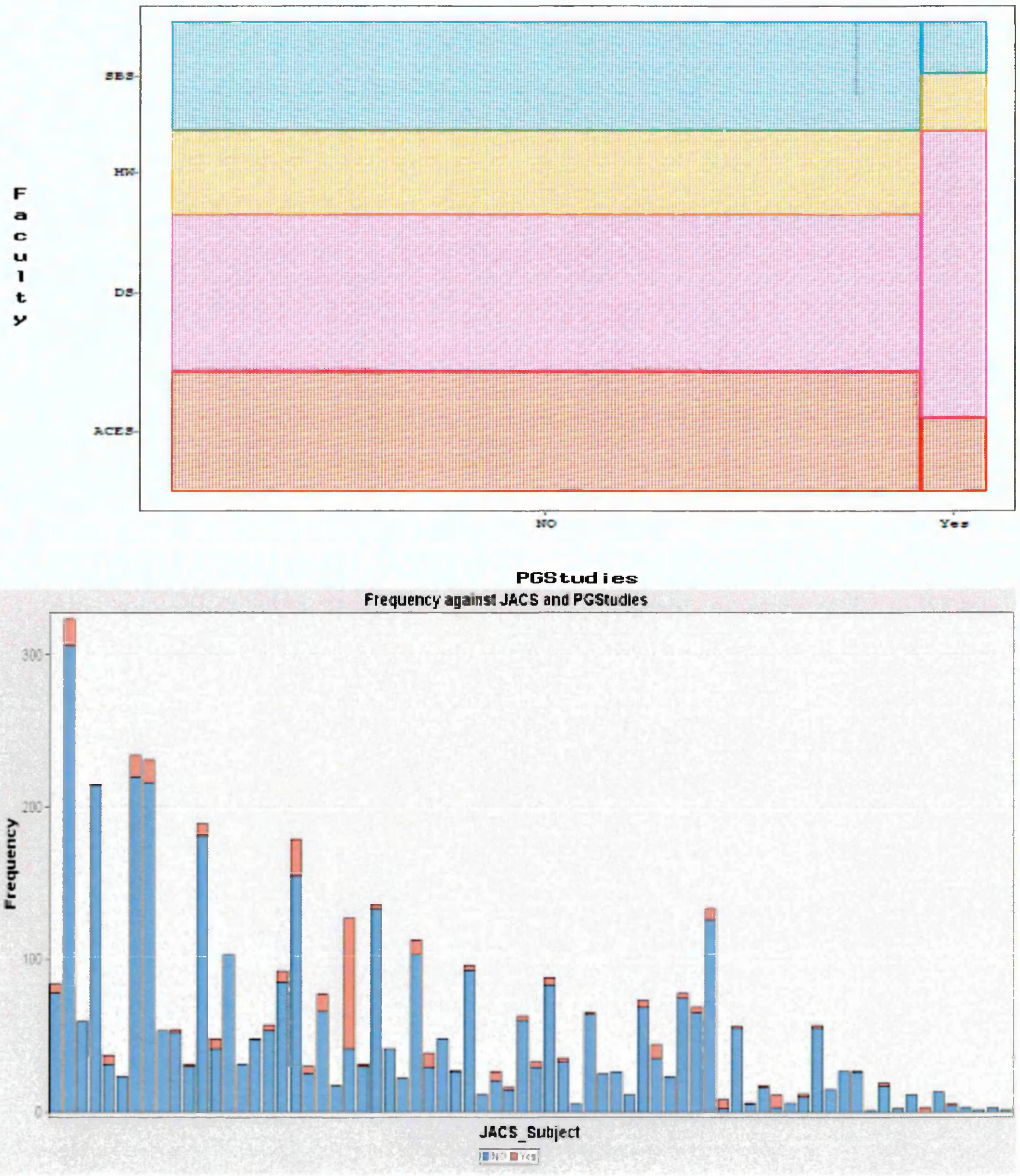


Figure 6.57 – Faculty and JACS Subject Mosiac Plot and Bar Chart.

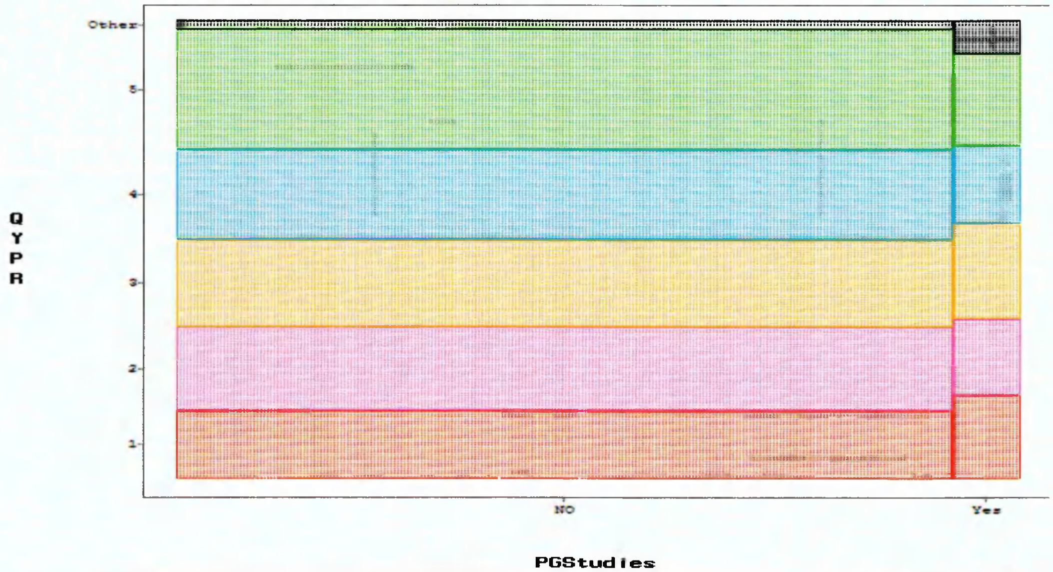
Course - Any graph of this variable would make it difficult to determine the different categories of students going onto postgraduate studies.

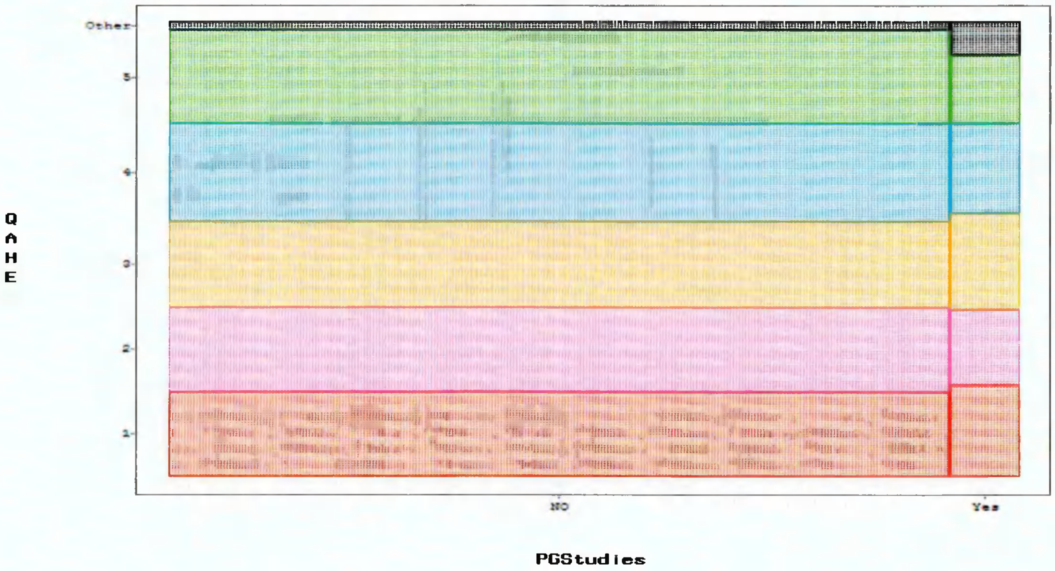
Faculty - a larger proportion of undergraduates go on to study a postgraduate degree in the faculty of Development and Society. However, inclusion of this variable at the modelling stage is questionable as there are probably too few levels for the modelling stage in that any models built using this variable would be to general.

JACS Subject - The JACS Subject group variable provides a little more detail than the Faculty variable and a little less detail than Course. Whilst it's difficult to interpret the above graph due to the number of levels. It does show that some JACS Subject groups have a large proportion of undergraduates who go on to study a postgraduate degree and that some of the JACS Subject groups have no students who go on to study at postgraduate level.

SOCIOECONOMIC GROUPINGS

QYPR, QAHE and SocioEconomicGP are plotted against PGStudies, below.





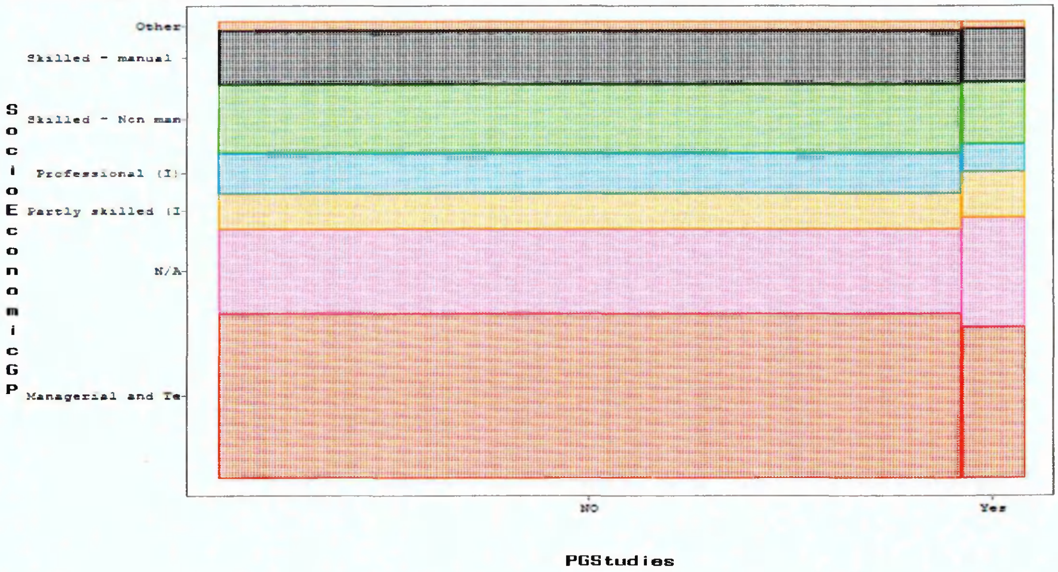


Figure 6.58 – Socioeconomic Groupings Mosiac Plots.

Participation of Young People in HE by Postcode (QYPR) - The graph above suggests that more students go on to study a postgraudate degree from areas where QYPR was rated has a 1 (low).

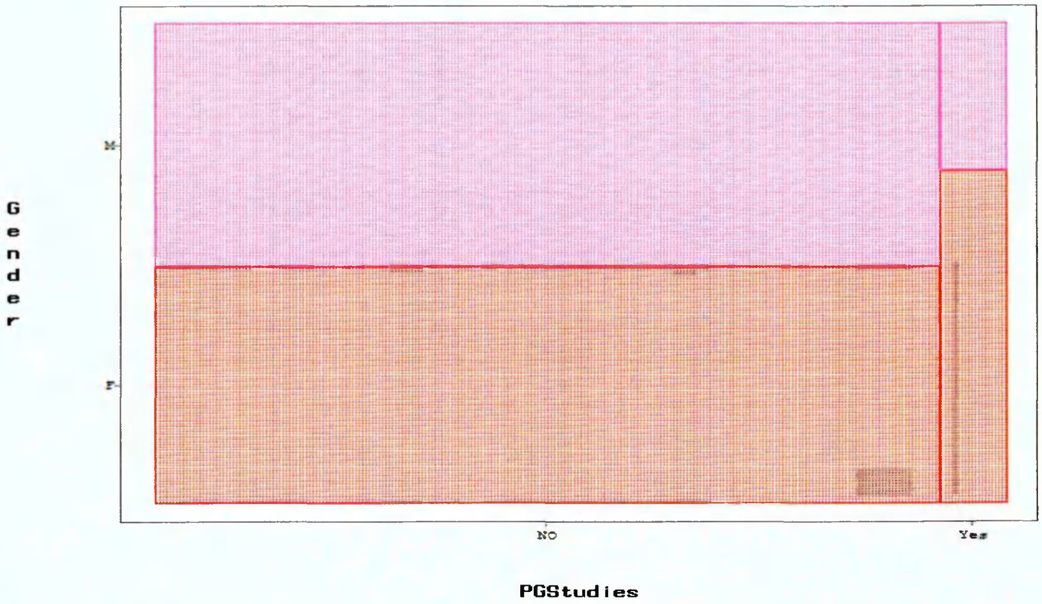
Adults with HE Qualifications by Postcode (QAHE) - Again the QAHE graph also suggests that there's a higher proportion of students who go on to study at postgrauate level from areas where QAHE is 1 (low).

Socioeconmic Group of parents - The mosaic plot of PGStudies against SocioEconmicGP highlights that more students from the managerial and technical group tended to go on to study a postgraduate degree. The group of students least likely to go on to study a postgraduate degree at SHU are from the professional SocioEconmicGP.

As stated above, all three of these variables are trying to measure the level of social deprivation in the student's background. Therefore, only one of these variables, SocioEconmicGP or QYPR and QAHE, should make it into the final model. The inclusion of all three variables (QYPR, QAHE, Socioeconmic Group) should be assessed at the DM stage.

Gender and Disability

The graphs, below, show Gender and Disability plotted against PGStudies.



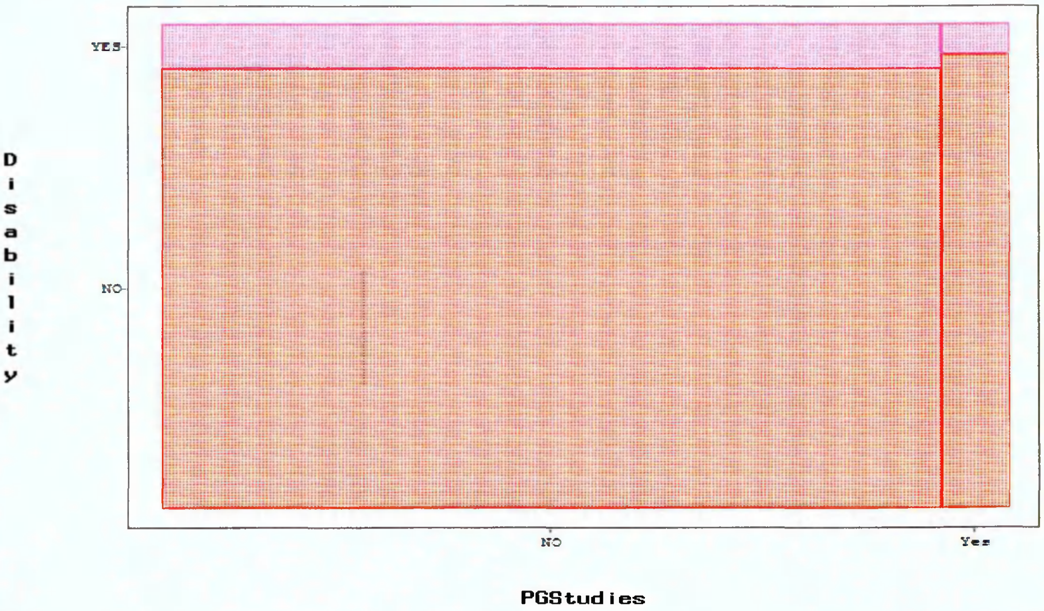


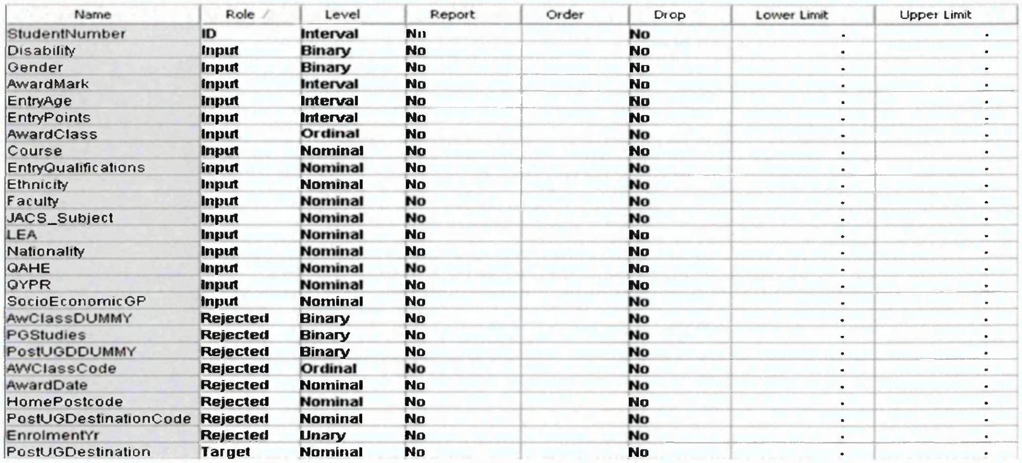
Figure 6.59 – Gender and Disability Mosiac Plots.

Gender - more females go on to study a postgraduate degree than males.

Disability - the mosaic plot above indicates that very few disabled students go on to take a postgraduate degree at SHU. The inclusion of Disability in the final model is questionable, given the small amount of disabled students in the dataset.

3. Student Employment Type

Using PostUGDestination as the target variable (rejecting the other two target variables (PGStudies and AwardClass) the list below shows the variables to be assessed.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
StudentNumber	ID	Interval	No		No	.	.
Disability	Input	Binary	No		No	.	.
Gender	Input	Binary	No		No	.	.
AwardMark	Input	Interval	No		No	.	.
EntryAge	Input	Interval	No		No	.	.
EntryPoints	Input	Interval	No		No	.	.
AwardClass	Input	Ordinal	No		No	.	.
Course	Input	Nominal	No		No	.	.
EntryQualifications	Input	Nominal	No		No	.	.
Ethnicity	Input	Nominal	No		No	.	.
Faculty	Input	Nominal	No		No	.	.
JACS_Subject	Input	Nominal	No		No	.	.
LEA	Input	Nominal	No		No	.	.
Nationality	Input	Nominal	No		No	.	.
QAHE	Input	Nominal	No		No	.	.
QYPR	Input	Nominal	No		No	.	.
SocioEconomicGP	Input	Nominal	No		No	.	.
AwClassDUMMY	Rejected	Binary	No		No	.	.
PGStudies	Rejected	Binary	No		No	.	.
PostUGDDUMMY	Rejected	Binary	No		No	.	.
AWClassCode	Rejected	Ordinal	No		No	.	.
AwardDate	Rejected	Nominal	No		No	.	.
HomePostcode	Rejected	Nominal	No		No	.	.
PostUGDestinationCode	Rejected	Nominal	No		No	.	.
EnrolmentYr	Rejected	Unary	No		No	.	.
PostUGDestination	Target	Nominal	No		No	.	.

Figure 6.60 – Student Employment Type Variables and Roles.

The role and levels of the variables are:

Role	Level	No. of variables
ID	Interval	1
Input	Nominal	10
Input	Ordinal	1
Input	Binary	2
Input	Interval	3
Rejected	Binary	3
Rejected	Nominal	3
Rejected	Ordinal	1
Rejected	Unary	1
Target	Nominal	1
Total No. Variables		**26**

Figure 6.61 – Student Employment Type Roles, Levels and Counts.

Rejected variables include:

- PGStudies - the target variable for another model;
- AWClassCode - relates to the different award classes for another model;
- AWClassDummy - used to determine potential groups for another model;
- AwardDate - the date that the student received the award this has no significance to the AwardClass received by the student;
- PostUGDUMMY - used to determine potential groups;
- PostUGDestinationCode - relates to the different postgraduate destinations;
- EnrolmentYr - a unary variable;
- HomePostcode - there are other location based variables in the dataset.

The PGStudies variable is given the role of target for the data understanding. However, PostUGDUMMY and PostUGDestinationCode may be used when assessing the variables/values, using variable selection, categorical input consolidation and interactive binning/grouping, in building the models section below.

The distribution of the data, below, is exactly the same as PGStudies and AwardClass distributions discussed previously.

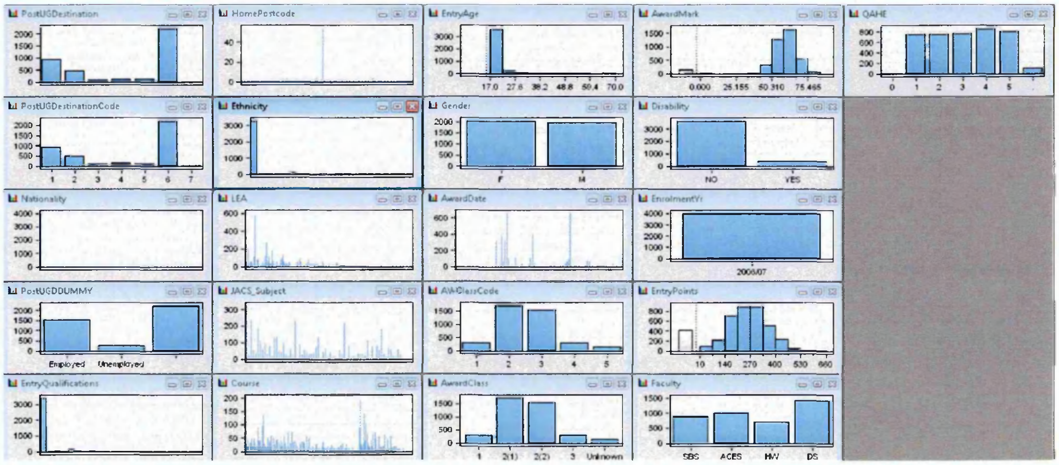
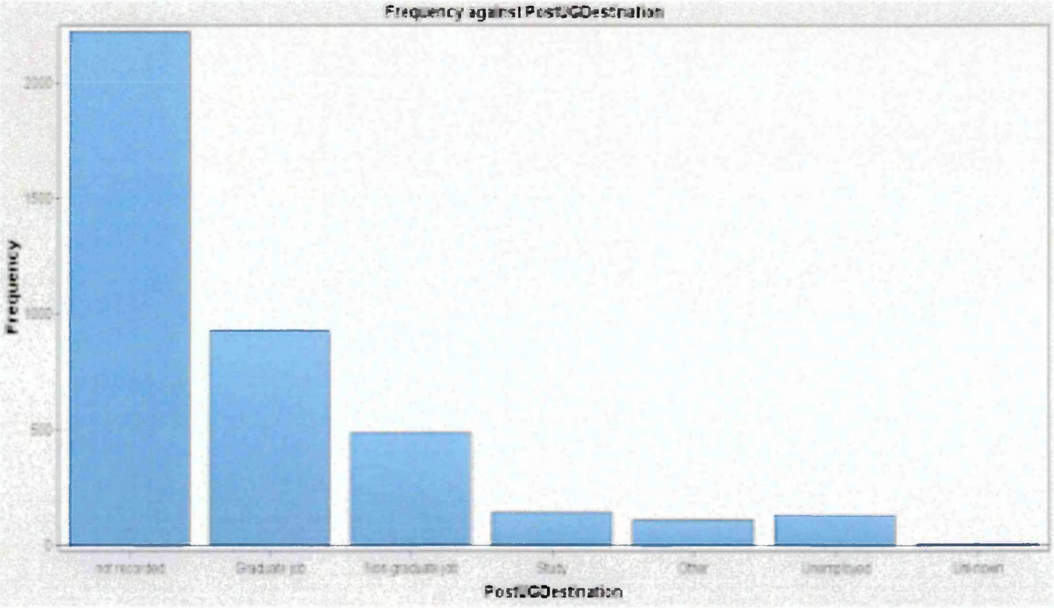


Figure 6.62 – Student Employment Type Initial Histograms.

Therefore, this section will consider the individual variables in relation to the PostUGDestination target variable.

POST UNDERGRADUATE DESTINATION

The frequency of PostUGDestination, PostUGDUMMY and PostUGDestinationCode are plotted below respectively.



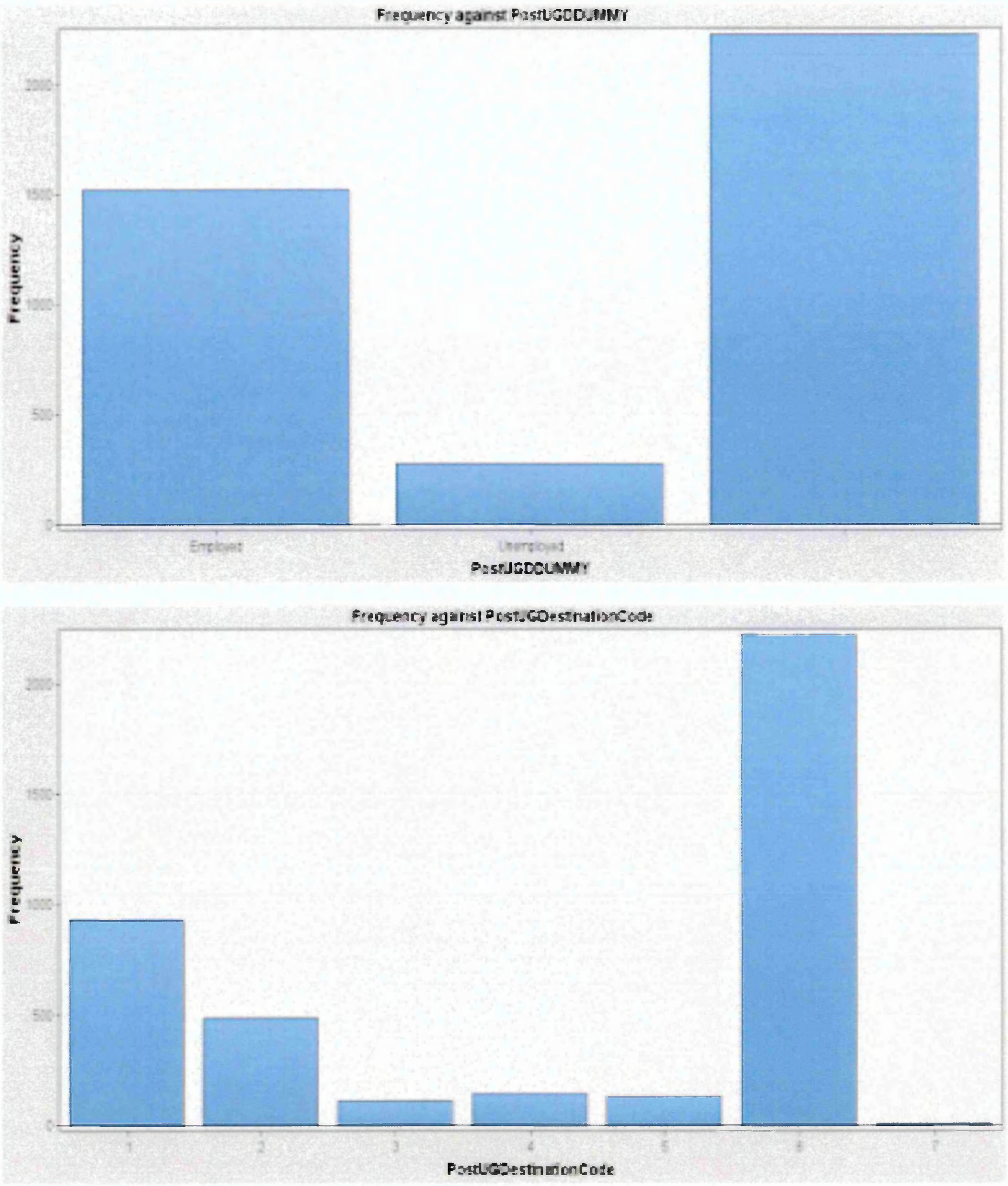


Figure 6.63 – Post Undergraduate Destination Bar Charts.

Of the 4023 students the university captured employment information about 1794 students, the type of information recorded by SHU is:

- Not Recorded
- Graduate Job
- Non-Graduate Job
- Study - not just further studies at SHU
- Other
- Unemployed
- Unknown

The PostUGDestination and PostUGDestinationCode graphs show that the majority of values are either not recorded or unknown. However, of the recorded values, SHU students appear to obtain more graduate jobs than any other category (non-graduate jobs, study, other, and unemployed). Ignoring the not recorded and unknown values, the PostUGDDUMMY variable contains more employed students than unemployed students.

Award Classification and Mark

The students AwardClass and AwardMark are plotted against PostUGDestination below.

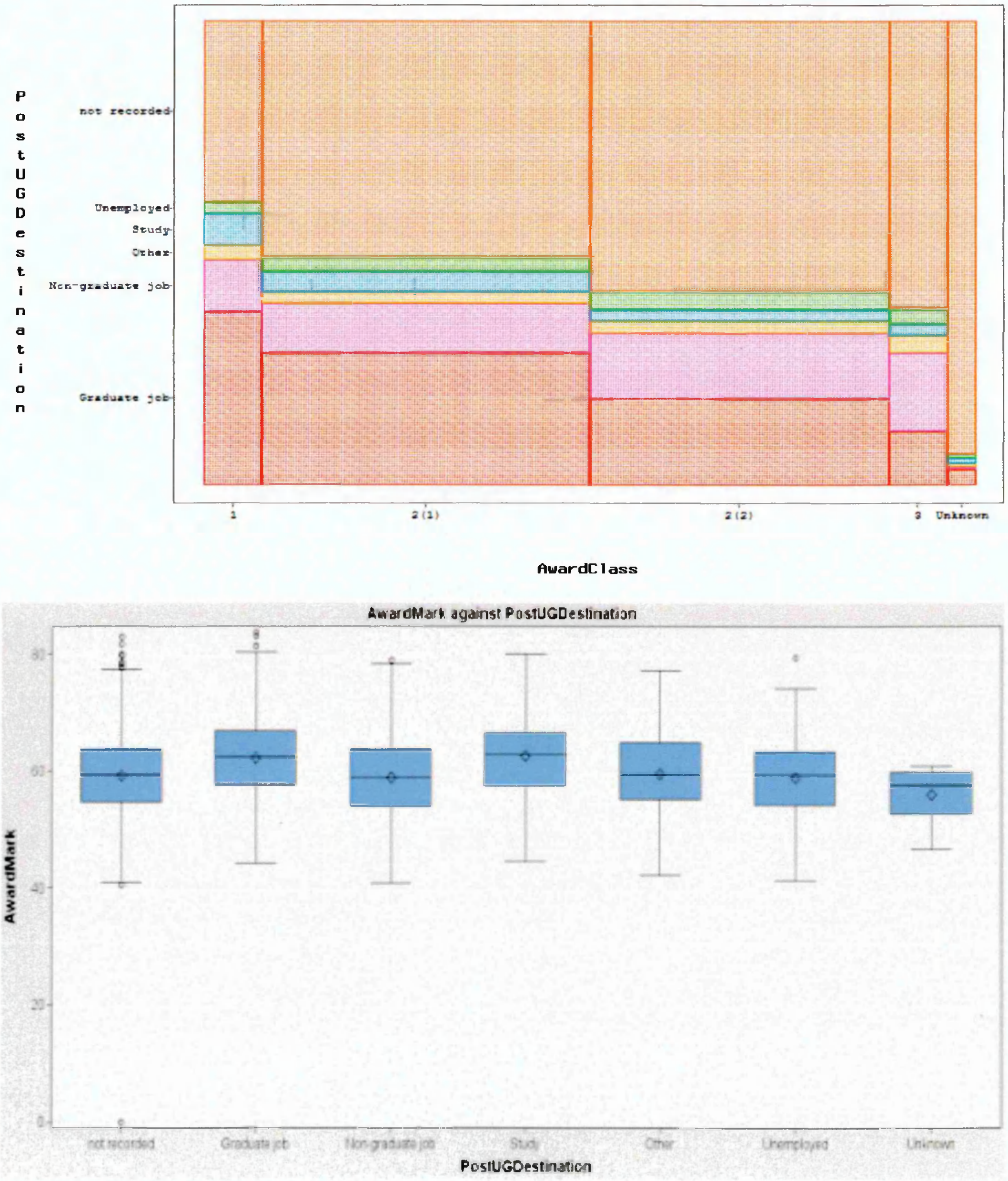


Figure 6.64 – Award Classification and Mark Mosaic Plot and Box Plot.

Award Classification - The mosaic plot of AwardClass against PostUGDestination shows that students who achieved a 1st classification are more likely to obtain a graduate job. This group also has the largest number of students who go on to take further studies. In addition to this, the largest number of students ending up in non-graduate jobs or unemployed seems to be those students who obtained either a 2:2 or a 3rd classification. Interestingly, the number of students who don't respond to the DOL survey seems to increase as the students AwardClassification is reduced.

Award Mark - The box plot of AwardMark against PostUGDestination indicates that the average mark of students obtaining a graduate job or going onto further study tend to be higher than those students who were unemployed, went on to work in non-graduate jobs or go on to do some other activity after their undergraduate degree. Interestingly, the lowest average award mark relates to those students who ended up working in non-graduate jobs.

As AwardClassification and AwardMark are measures of the student's success and AwardMark is used to calculate the AwardClassification, AwardMark will be used in favour of AwardClassification.

Entry Age

The box plot, below, shows student entry ages against PostUGDestination.

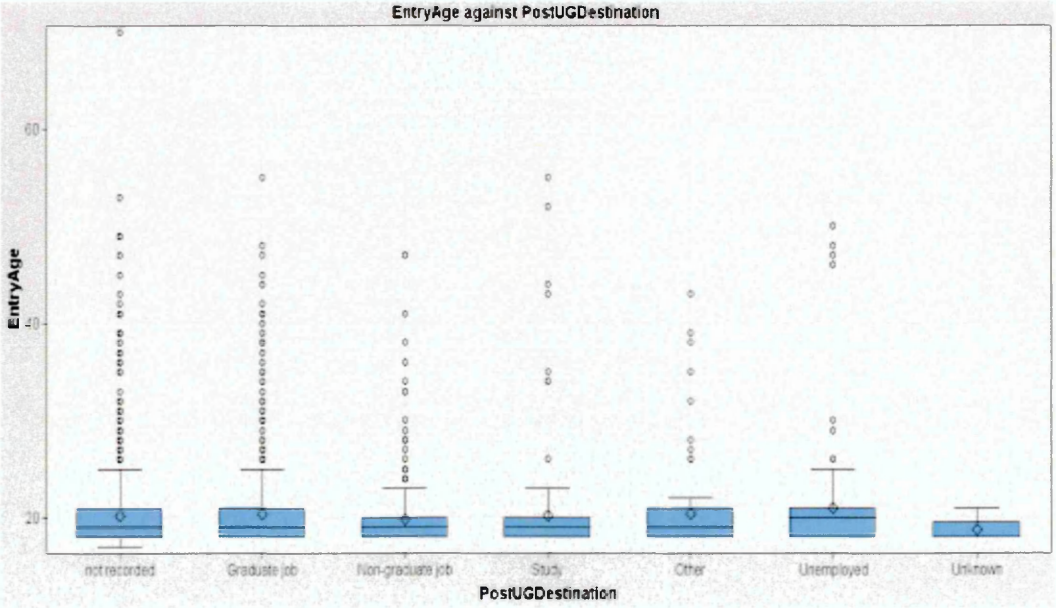


Figure 6.65 – Entry Age Box Plot.

The above graph shows that the average age of all students, at undergraduate entry, across all of the PostUGDestination values is below 20, with the exception of those students who end up unemployed. This appears to suggest that older students who enter full-time undergraduate degrees find it harder to get a job after graduation.

Nationality and Ethnicity

The students Nationality and Ethnicity are plotted against Frequency below.

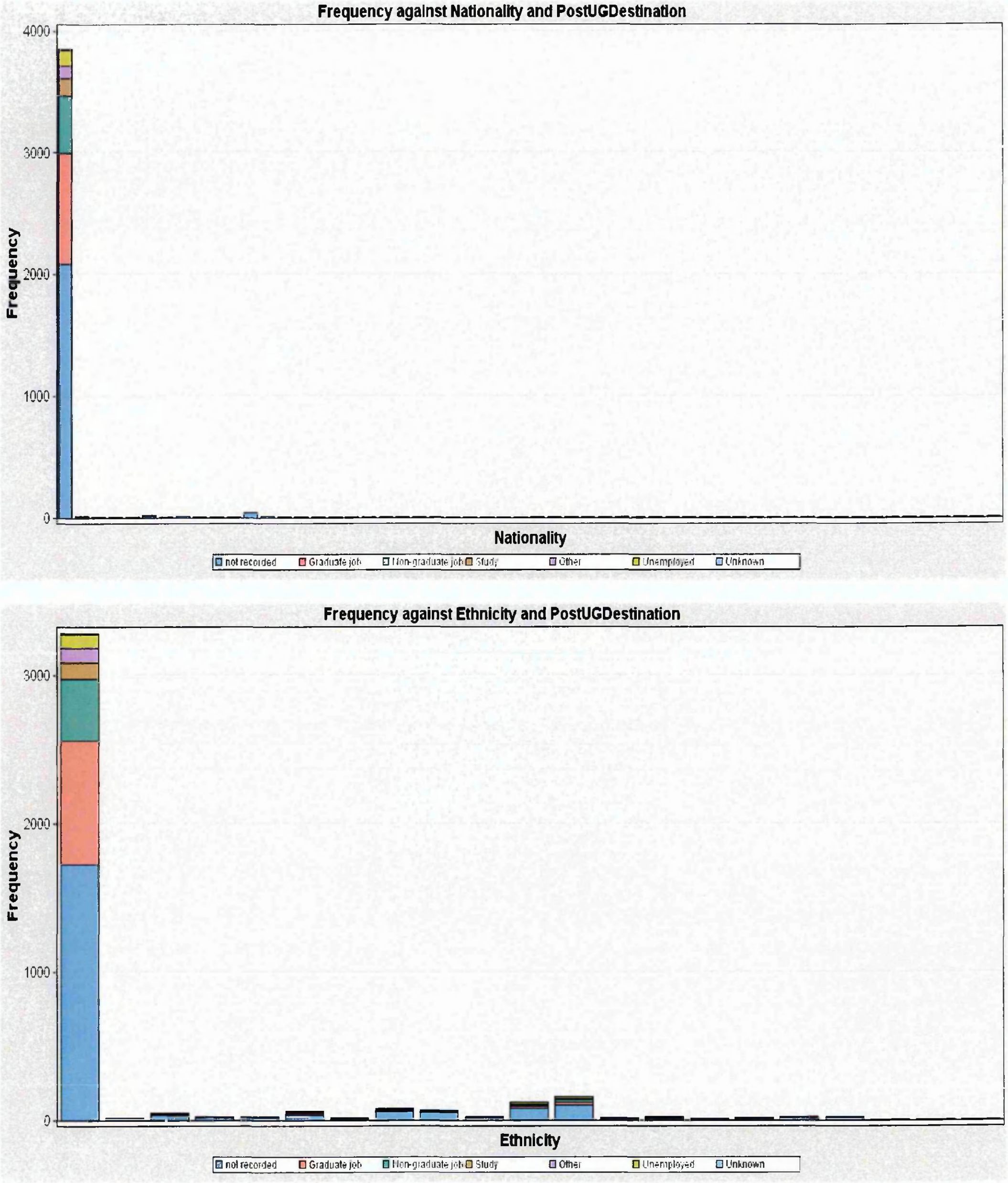


Figure 6.66 – Nationality and Ethnicity Bar Charts.

Nationality - as before the majority of students are from the UK and the largest proportion of these students, when ignoring the not recorded value, obtained a graduate job.

Ethnicity - again the largest proportion of students are White British and, of the recorded values, the majority of these students went on to obtain a graduate job.

The inclusion of Nationality and Ethnicity at the modelling stage will need to be assessed at the modelling stage. This is mainly due to the large number of UK and White British students in these variables.

Entry Qualifications and Entry Points

The students EntryQualifications and EntryPoints are plotted against the PostUGDestination variable below.

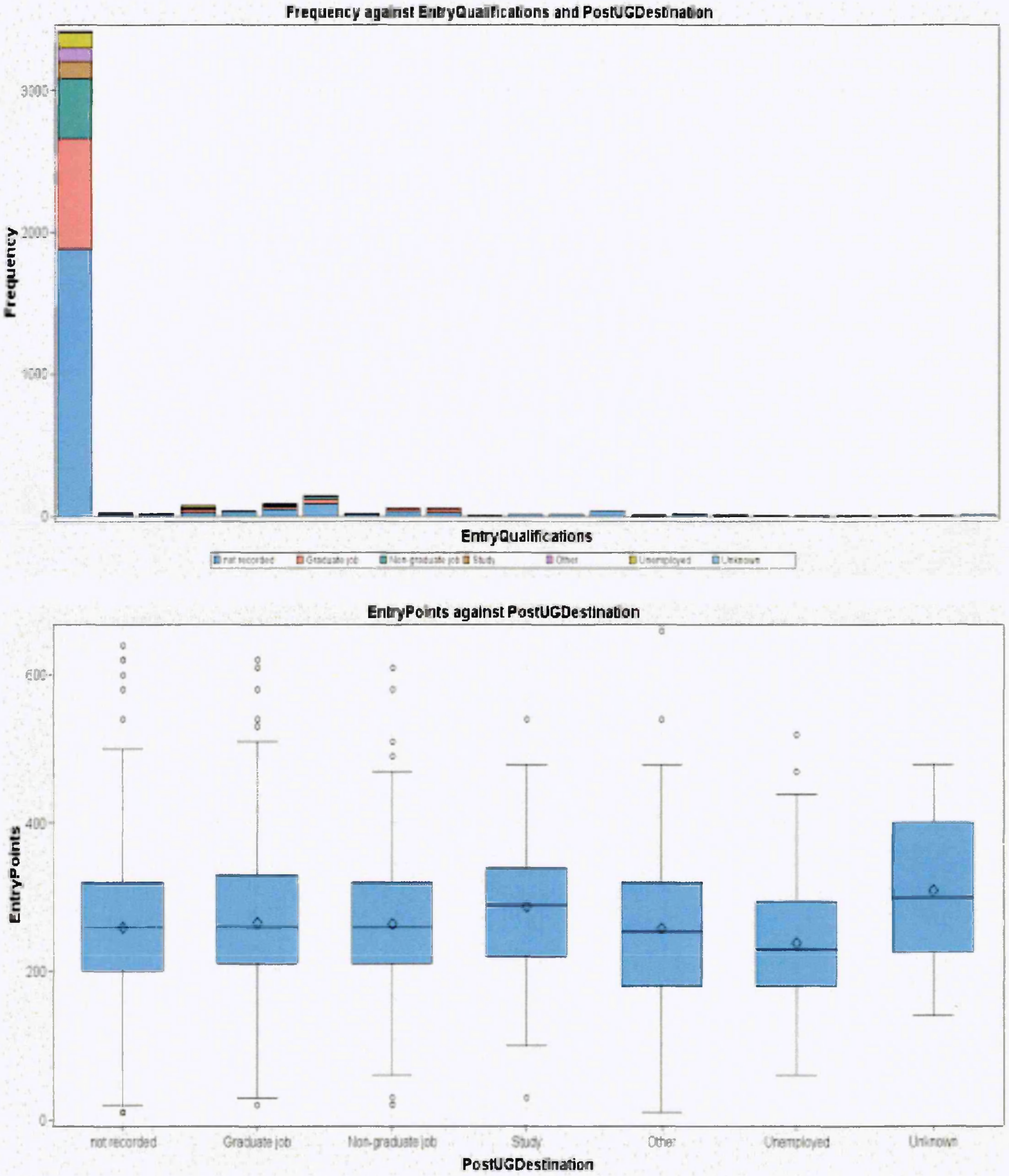


Figure 6.67 – Entry Qualifications and Points Bar Chart and Box Plot.

Entry Qualifications - as noted previously the majority of students enter university with A-Level/GNVQ3 qualifications and it is this group that secure the larger proportion of graduate jobs.

Entry Points - The box plot above shows that students who went on to take further study had the highest average undergraduate EntryPoints when they enrolled onto their original degree course. Conversely, students that ended up unemployed had the lowest average undergraduate EntryPoints when enrolling onto their original degree course.

Again, EntryQualifications should be rejected in favour of EntryPoints as this is a more comparable measure, going forward, of the students level 3 entry qualifications.

Course, Faculty and JACS Subject

This section considers Course, Faculty and JACS Subject in relation to the target variable of PostUGDestination.

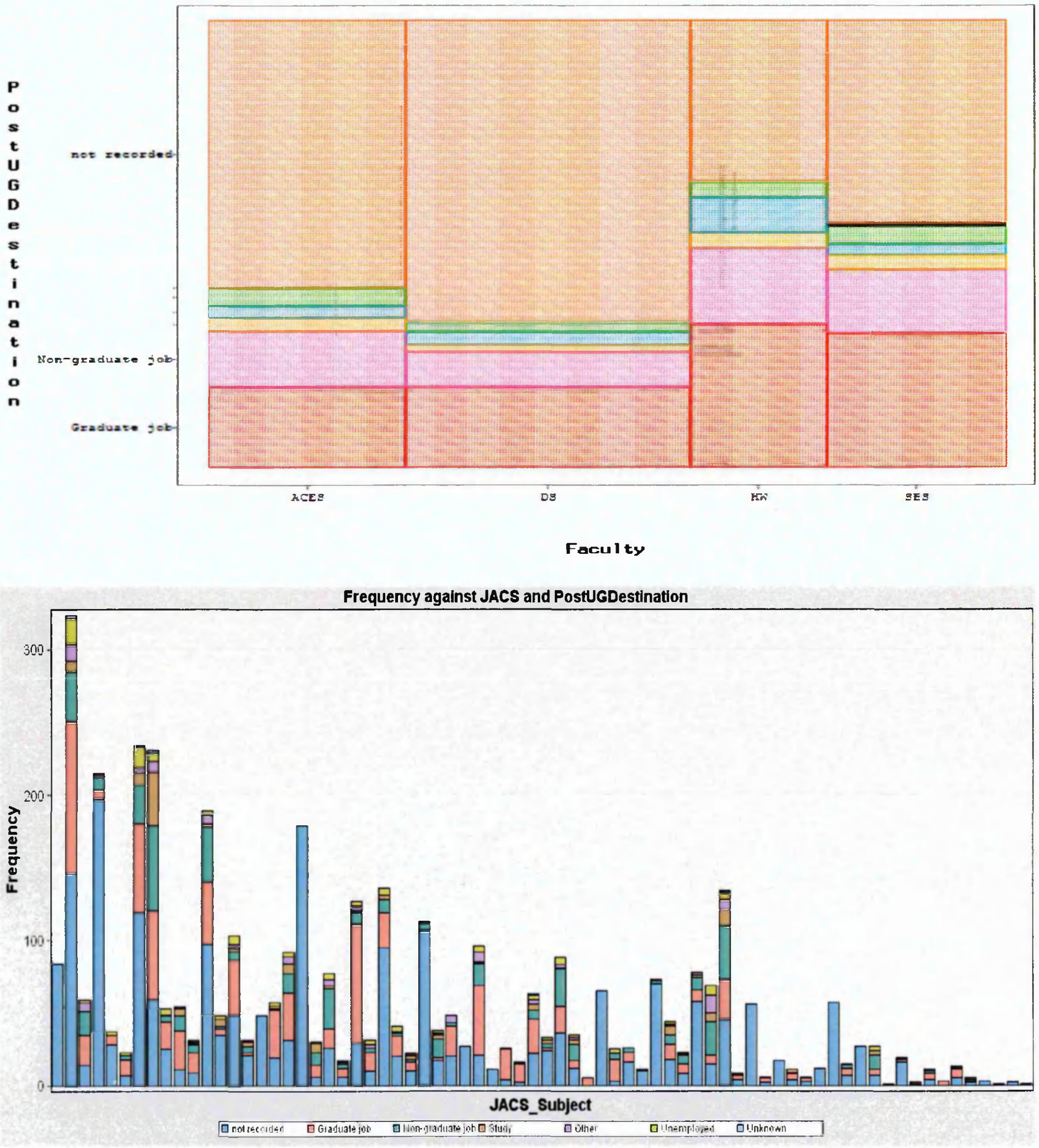


Figure 6.68 – Faculty and JACS Subject Mosaic Plot and Bar Chart.

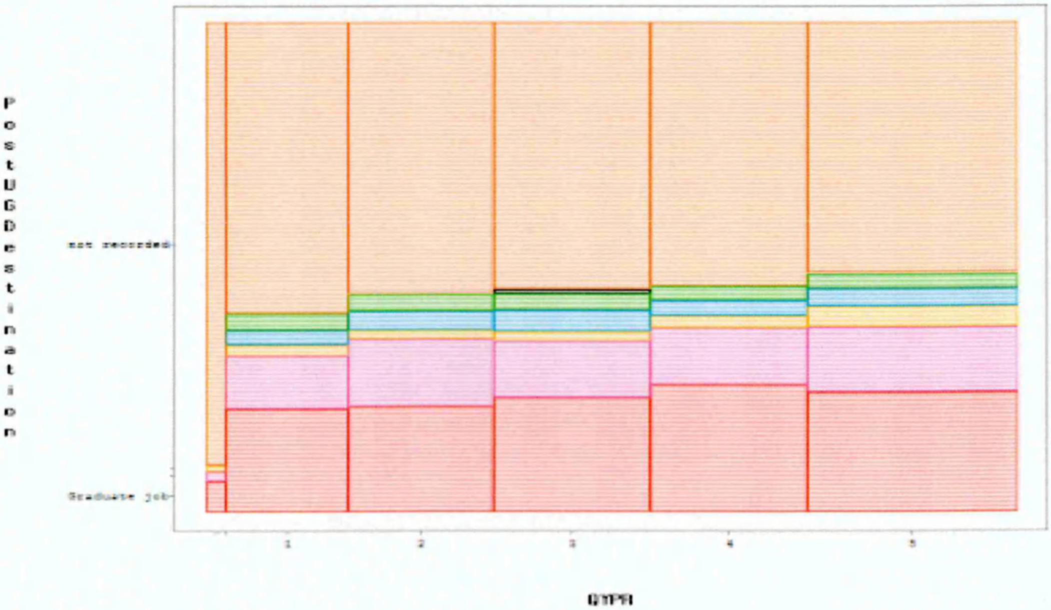
Course - Any graph of this variable would make it difficult to determine the different levels of students post undergraduate destinations.

Faculty - The mosaic plot of PostUGDestination against Faculty indicates that the faculty of Health and Wellbeing has more students who go on to obtain a graduate job. This faculty also has the largest proportion of students who went on to take further studies. The plot also indicates that the largest number of students who ended up unemployed were members of the faculty of Sheffield Business School and that the Faculty of Development and Society had less students responding to the DOL survey.

JACS Subject - Whilst it's difficult to determine the individual JACS subject groups, from the above graph, due to the number of categories. It does however show that some of the JACS Subjects have a larger proportion of the graduate job, unemployed, study and non-graduate job values than others. It is also perhaps worth pointing out that some of the subjects also had a 100% non-response.

Socioeconomic Groupings

QYPR, QAHE and SocioEconomicGP are plotted against PostUGDestination, below.



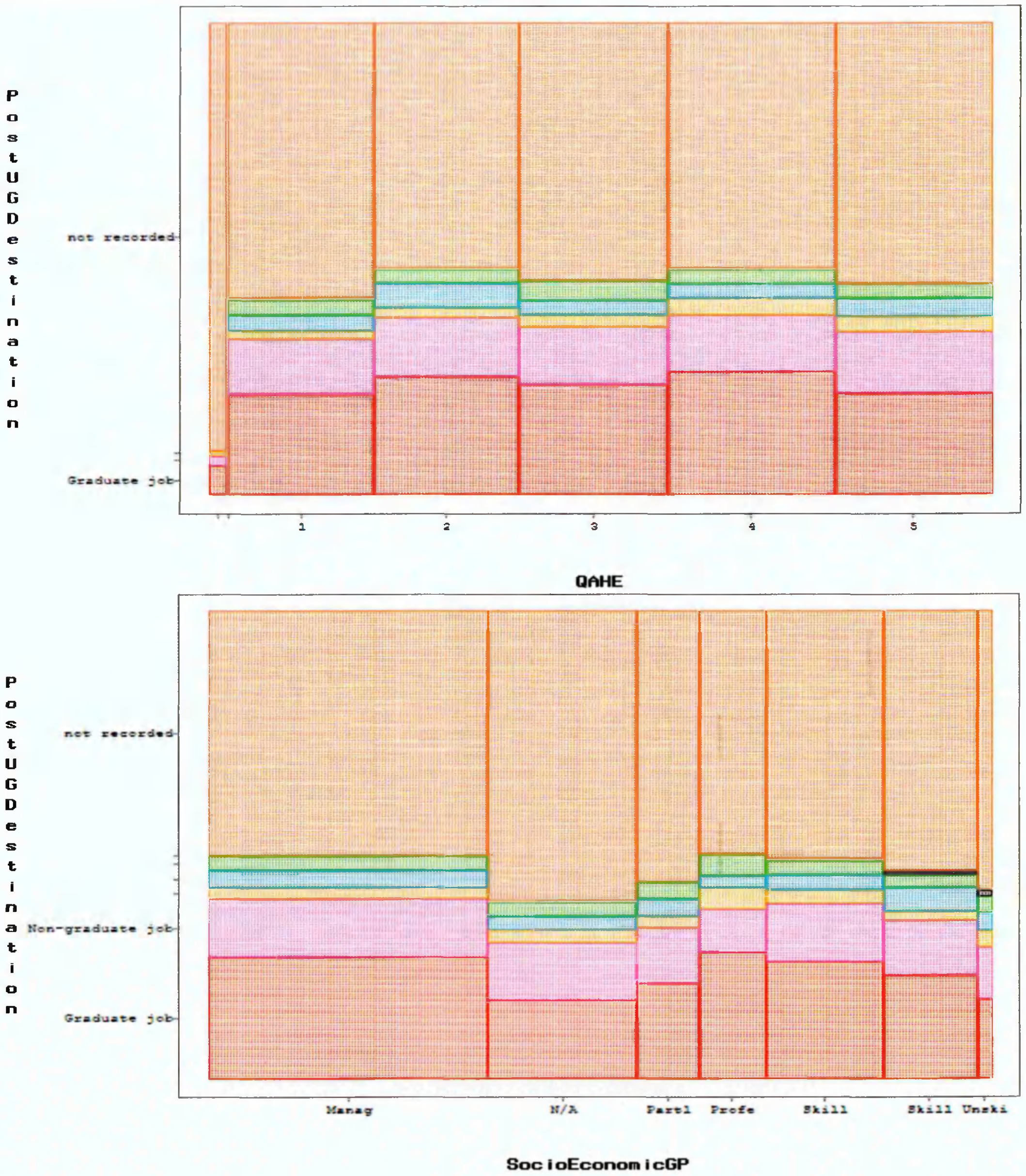


Figure 6.69 – Socioeconomic Groupings Mosaic Plots.

QYPR - the mosaic plot of PostUGDestination against QYPR shows that students from areas where QYPR is medium to high (4) are more likely to end up in a graduate job. QYPR 3 has more students who go on to further studies and QYPR 2 and 3 have more students who end up unemployed. Furthermore, it appears that students from areas where QYPR is high (5) tend to respond better to the DOL survey.

QAHE - the central mosaic plot indicates that more students from QAHE 2 and 4 tend to obtain a graduate job. More students from QAHE 2 (low to medium) tend to go on to further studies. Additionally, more students end up unemployed from areas where QAHE is medium (3).

Socioeconomic Group of parents - the mosaic plot of PostUGDestination against SocioEconomicGP highlights that students from managerial and technical, and professional groups tend to obtain more graduate jobs. More students from the skilled group tended to go on to undertake further studies. Moreover, there were more unemployed students in the professional group.

The inclusion of all three variables (QYPR, QAHE, SocioeconmicGP) should be assessed at the modelling stage.

GENDER AND DISABILITY

The graphs, below, show Gender and Disability plotted against PostUGDestination.

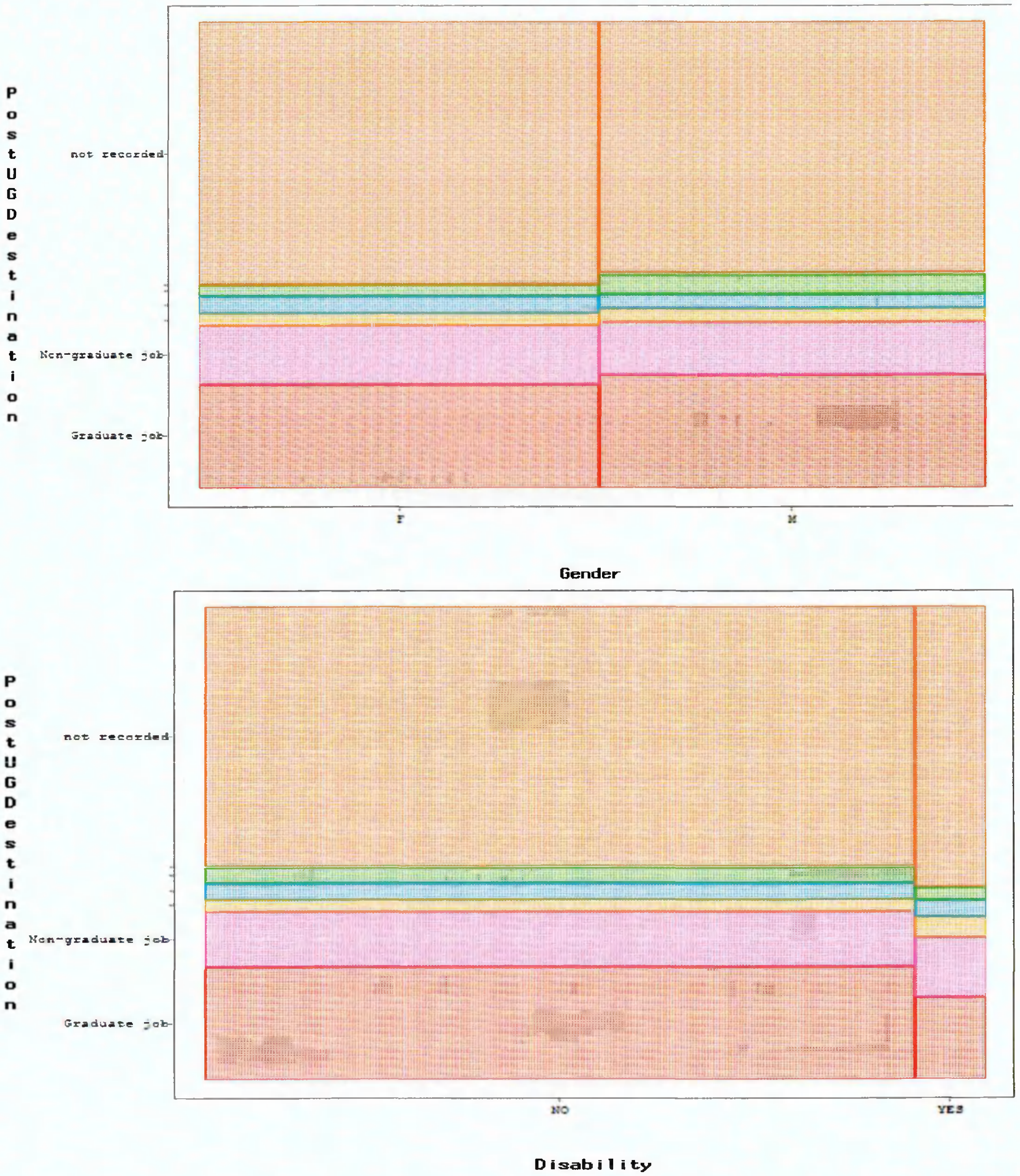


Figure 6.70 – Gender and Disability Mosaic Plots.

Gender - the mosaic plot above highlights that slightly more males end up getting graduate jobs than females, more males then females also end up unemployed. Males also seem to respond a little better to the DOL survey than females.

Disability - fewer disabled students end up working in a graduate job and less disabled students tend to respond to the DOL survey. The inclusion of Disability in the final model is questionable, given the small amount of disabled students in the dataset.

6.1.3.2.5 Number of Observations and Event Rates

In conclusion, the data understanding process outlined above identified that of the 4023 records:

- 3873 obtained an honours degree (96% event rate);
- 322 went onto postgraduate studies at SHU (8% event rate); and
- 1794 employment types were recorded (45% event rate).

It is difficult to assess the amount of data (observations and event rates) required to produce effective models. Berry and Linoff (2004) point out that DM is more effective with larger data sets, which usually contain more than 30,000 records. However, according to the research carried out by John and Langley (1996) and Oates and Jensen (1998), the sample size has little impact on the accuracy of the model provided that the sample size is greater than 300. They found that the accuracy of a sample size between 300 and 2180 was 2% as measured by the DM confidence factor.

Arguably some type of oversampling might have to be carried out when building the progression onto postgraduate studies model, due to the 8% event rate.

6.2 SUMMARY

This chapter details the process of creating three DM marts, using the BDLD introduced previously, which will be used in the following chapter. The data within each DM mart is explored and some discussion is presented around potential patterns and trends in the data. Through this process a number of additional variables, JACS Subject and Faculty were added to the DM marts. Additionally, as the demographics data was the same in each DM mart, a decision was made to condense the three DM marts into one single table (DM mart). The chapter concludes with a brief discussion around sample sizes and event rates, which highlights that some form of oversampling may be required (due to an 8% event rate) when carrying out the modelling of student progression onto postgraduate studies.

7 Data Mining

"Knowing how to exploit data effectively can help you to use available technologies to reveal the hidden patterns and trends contained therein." (Westphal and Claxton 1998, p xxi)

This chapter will discuss the development of three predictive models using the SAS® SEMMA, outlined in section 4.4.1.

7.1 Mining the Data

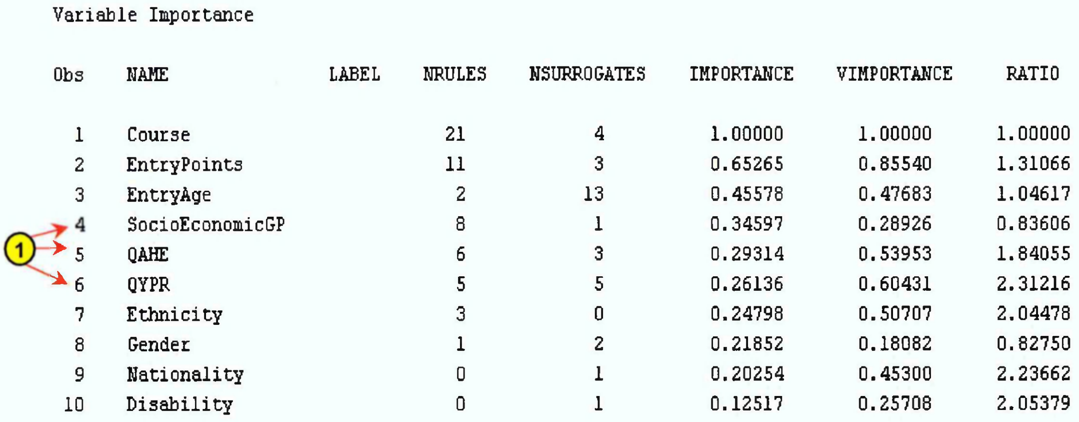
The models have been developed through numerous iterations, however, this section provides an overview of construction of the final award classification, progression onto postgraduate studies at SHU and employment DM models. It outlines the process followed in building all three models, using the SAS® SEMMA (Sample, Explore, Modify, Model, Assess) methodology, and each section will conclude with a brief overview of the final model.

7.1.1 Building the Award Classification Model (ACM)

This section details the process followed in building the award classification DM model using SAS® SEMMA.

7.1.1.1 ACM Sample

As stated previously the final DM mart contains 4023 observations of these 96% obtained an honours degree classification. The percentage of students obtaining a 1st classification was 7.43%, 42.65% obtained a 2:1, 38.70% achieved a 2:2 and 7.48% were awarded a 3rd classification. The remaining 3.73% failed to achieve an honours classification. Ultimately, it was decided not to sample the data set as the award classification percentages seemed representative of the population. In that a 2:1 and 2:2 classifications are more common than a 1st or a 3rd. A 80:20 data partition was also setup so that the data could be divided into training and validation respectively, an 80:20 split was favoured as this produced the better model. However, this resulted in a much smaller data set for validating the model. The variables within the data set were reduced through the authors own knowledge of the data/HE and by using a decision tree for variable selection. This resulted in the following inputs being selected for modelling:



```
Variable Importance

Obs    NAME               LABEL    NRULES    NSURROGATES    IMPORTANCE    VIMPORTANCE     RATIO

  1    Course                        21            4          1.00000       1.00000      1.00000
  2    EntryPoints                   11            3          0.65265       0.85540      1.31066
  3    EntryAge                       2           13          0.45578       0.47683      1.04617
  4    SocioEconomicGP                8            1          0.34597       0.28926      0.83606
  5    QAHE                           6            3          0.29314       0.53953      1.84055
  6    QYPR                           5            5          0.26136       0.60431      2.31216
  7    Ethnicity                      3            0          0.24798       0.50707      2.04478
  8    Gender                         1            2          0.21852       0.18082      0.82750
  9    Nationality                    0            1          0.20254       0.45300      2.23662
 10    Disability                     0            1          0.12517       0.25708      2.05379
```

Figure 7.1 – Variable Importance Table.

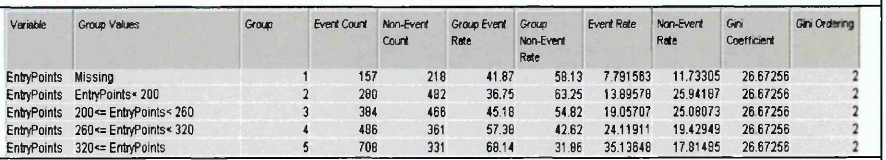
It is perhaps important to note that (1) are all measures of social deprivation, therefore if all of these appear in the final model then either <u>SocioEconomicGP</u> or <u>QAHE</u> and <u>QYPR</u> will have to be removed in favour of one of them, see 6.1.3.2.4.

7.1.1.2 ACM Explore

The data has already been thoroughly explored through the data understanding process outlined in Chapter 6, through this a number of graphs were produced and outliers were identified.

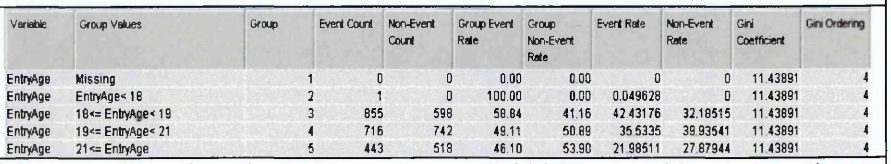
7.1.1.3 ACM Modify

This section will focus on using categorical variable consolidation (see glossary page vii) for categorical values, interactive binning/grouping (for interval values) and the authors own knowledge of the data/HE to determine suitable groups of values in relation to the target variable.

<table>
<tr><td colspan="2"><u>EntryPoints</u></td></tr>
<tr><td>Interactive Binning/ Grouping Results</td><td>
<table>
<tr><th>Variable</th><th>Group Values</th><th>Group</th><th>Event Count</th><th>Non-Event Count</th><th>Group Event Rate</th><th>Group Non-Event Rate</th><th>Event Rate</th><th>Non-Event Rate</th><th>Gini Coefficient</th><th>Gini Ordering</th></tr>
<tr><td>EntryPoints</td><td>Missing</td><td>1</td><td>157</td><td>218</td><td>41.87</td><td>58.13</td><td>7.791563</td><td>11.73305</td><td>26.67256</td><td>2</td></tr>
<tr><td>EntryPoints</td><td>EntryPoints< 200</td><td>2</td><td>280</td><td>482</td><td>36.75</td><td>63.25</td><td>13.89578</td><td>25.94187</td><td>26.67256</td><td>2</td></tr>
<tr><td>EntryPoints</td><td>200<= EntryPoints< 260</td><td>3</td><td>384</td><td>466</td><td>45.18</td><td>54.82</td><td>19.05707</td><td>25.08073</td><td>26.67256</td><td>2</td></tr>
<tr><td>EntryPoints</td><td>260<= EntryPoints< 320</td><td>4</td><td>486</td><td>361</td><td>57.38</td><td>42.62</td><td>24.11911</td><td>19.42949</td><td>26.67256</td><td>2</td></tr>
<tr><td>EntryPoints</td><td>320<= EntryPoints</td><td>5</td><td>708</td><td>331</td><td>68.14</td><td>31.86</td><td>35.13648</td><td>17.81485</td><td>26.67256</td><td>2</td></tr>
</table></td></tr>
<tr><td>Discussion</td><td>The interactive binning/grouping node suggests that the entry points can be grouped into 5 groups.</td></tr>
<tr><td>Inclusion in final model</td><td>Yes</td></tr>
<tr><td>Modify</td><td>No – Data Mining model will determine the best split</td></tr>
</table>

ENTRYAGE

INTERACTIVE BINNING/ GROUPING RESULTS



Variable	Group Values	Group	Event Count	Non-Event Count	Group Event Rate	Group Non-Event Rate	Event Rate	Non-Event Rate	Gini Coefficient	Gini Ordering
EntryAge	Missing	1	0	0	0.00	0.00	0	0	11.43891	4
EntryAge	EntryAge< 18	2	1	0	100.00	0.00	0.049628	0	11.43891	4
EntryAge	18<= EntryAge< 19	3	855	598	58.84	41.16	42.43176	32.18515	11.43891	4
EntryAge	19<= EntryAge< 21	4	716	742	49.11	50.89	35.5335	39.93541	11.43891	4
EntryAge	21<= EntryAge	5	443	518	46.10	53.90	21.98511	27.87944	11.43891	4

DISCUSSION

Initially, it was thought that two groups could be created to capture those students who entered their degree at an age that was less than or equal to 22 and those that entered after the age of 22. However, the interactive binning/grouping node suggests that four entry age groups could be created.

INCLUSION IN FINAL MODEL

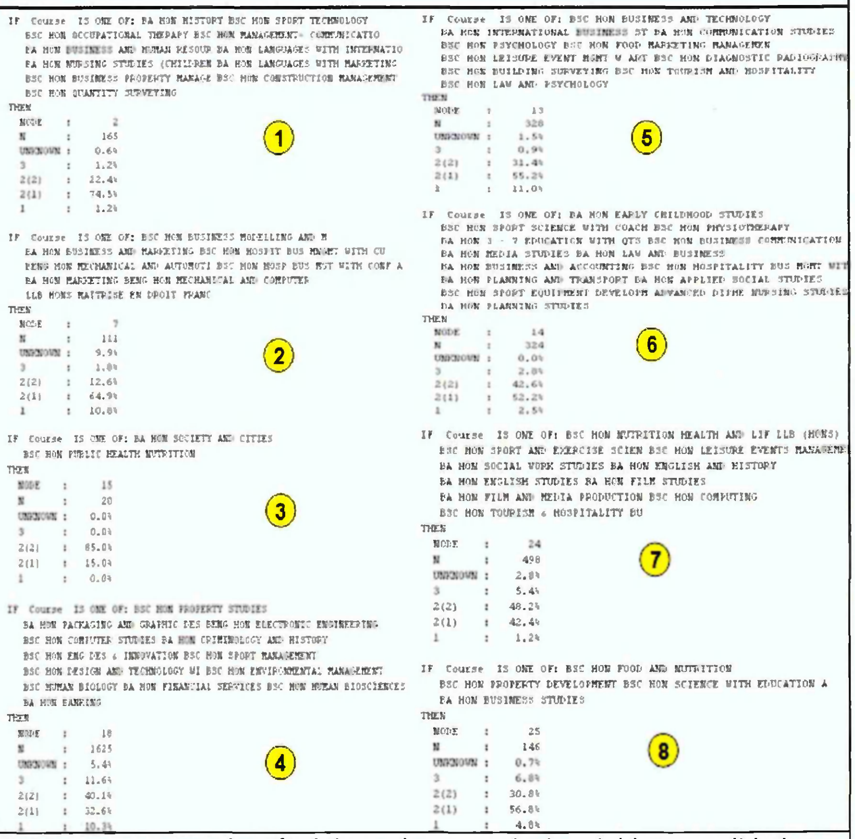
Yes

MODIFY

No – Data Mining model will determine the best split

COURSE

VARIABLE CONSOLIDATION



```
IF  Course  IS ONE OF: BA HON HISTORY BSC HON SPORT TECHNOLOGY
   BSC HON OCCUPATIONAL THERAPY BSC HON MANAGEMENT- COMMUNICATIO
   BA HON BUSINESS AND HUMAN RESOUR BA HON LANGUAGES WITH INTERNATIO
   BA HON NURSING STUDIES (CHILDREN BA HON LANGUAGES WITH MARKETING
   BSC HON BUSINESS PROPERTY MANAGE BSC HON CONSTRUCTION MANAGEMENT
   BSC HON QUANTITY SURVEYING
THEN
  NODE     :        2
  N        :      165
  UNKNOWN  :     0.6%
  3        :     1.2%
  2(2)     :    22.4%
  2(1)     :    74.5%
  1        :     1.2%

IF  Course  IS ONE OF: BSC HON BUSINESS MODELLING AND M
   BA HON BUSINESS AND MARKETING BSC HON HOSPIT BUS MNGMT WITH CU
   BENG HON MECHANICAL AND AUTOMOTI BSC HON HOSP BUS MGT WITH CONF A
   BA HON MARKETING BENG HON MECHANICAL AND COMPUTER
   LLB HONS MAITRISE EN DROIT FRANC
THEN
  NODE     :        7
  N        :      111
  UNKNOWN  :     9.9%
  3        :     1.8%
  2(2)     :    12.6%
  2(1)     :    64.9%
  1        :    10.8%

IF  Course  IS ONE OF: BA HON SOCIETY AND CITIES
   BSC HON PUBLIC HEALTH NUTRITION
THEN
  NODE     :       15
  N        :       20
  UNKNOWN  :     0.0%
  3        :     0.0%
  2(2)     :    85.0%
  2(1)     :    15.0%
  1        :     0.0%

IF  Course  IS ONE OF: BSC HON PROPERTY STUDIES
   BA HON PACKAGING AND GRAPHIC DES BENG HON ELECTRONIC ENGINEERING
   BSC HON COMPUTER STUDIES BA HON CRIMINOLOGY AND HISTORY
   BSC HON ENG DES & INNOVATION BSC HON SPORT MANAGEMENT
   BSC HON DESIGN AND TECHNOLOGY WI BSC HON ENVIRONMENTAL MANAGEMENT
   BSC HUMAN BIOLOGY BA HON FINANCIAL SERVICES BSC HON HUMAN BIOSCIENCES
   BA HON BANKING
THEN
  NODE     :       18
  N        :     1625
  UNKNOWN  :     5.4%
  3        :    11.6%
  2(2)     :    40.1%
  2(1)     :    32.6%
  1        :    10.3%

IF  Course  IS ONE OF: BSC HON BUSINESS AND TECHNOLOGY
   BA HON INTERNATIONAL BUSINESS ST BA HON COMMUNICATION STUDIES
   BSC HON PSYCHOLOGY BSC HON FOOD MARKETING MANAGEMEN
   BSC HON LEISURE EVENT MGMT W ART BSC HON DIAGNOSTIC RADIOGRAPHY
   BSC HON BUILDING SURVEYING BSC HON TOURISM AND HOSPITALITY
   BSC HON LAW AND PSYCHOLOGY
THEN
  NODE     :       13
  N        :      328
  UNKNOWN  :     1.5%
  3        :     0.9%
  2(2)     :    31.4%
  2(1)     :    55.2%
  1        :    11.0%

IF  Course  IS ONE OF: BA HON EARLY CHILDHOOD STUDIES
   BSC HON SPORT SCIENCE WITH COACH BSC HON PHYSIOTHERAPY
   BA HON 3 - 7 EDUCATION WITH QTS BSC HON BUSINESS COMMUNICATION
   BA HON MEDIA STUDIES BA HON LAW AND BUSINESS
   BA HON BUSINESS AND ACCOUNTING BSC HON HOSPITALITY BUS MGMT WIT
   BA HON PLANNING AND TRANSPORT BA HON APPLIED SOCIAL STUDIES
   BSC HON SPORT EQUIPMENT DEVELOPM ADVANCED DIPHE NURSING STUDIES
   BA HON PLANNING STUDIES
THEN
  NODE     :       14
  N        :      324
  UNKNOWN  :     0.0%
  3        :     2.8%
  2(2)     :    42.6%
  2(1)     :    52.2%
  1        :     2.5%

IF  Course  IS ONE OF: BSC HON NUTRITION HEALTH AND LIF LLB (HONS)
   BSC HON SPORT AND EXERCISE SCIEN BSC HON LEISURE EVENTS MANAGEME
   BA HON SOCIAL WORK STUDIES BA HON ENGLISH AND HISTORY
   BA HON ENGLISH STUDIES BA HON FILM STUDIES
   BA HON FILM AND MEDIA PRODUCTION BSC HON COMPUTING
   BSC HON TOURISM & HOSPITALITY BU
THEN
  NODE     :       24
  N        :      498
  UNKNOWN  :     2.8%
  3        :     5.4%
  2(2)     :    48.2%
  2(1)     :    42.4%
  1        :     1.2%

IF  Course  IS ONE OF: BSC HON FOOD AND NUTRITION
   BSC HON PROPERTY DEVELOPMENT BSC HON SCIENCE WITH EDUCATION A
   BA HON BUSINESS STUDIES
THEN
  NODE     :       25
  N        :      146
  UNKNOWN  :     0.7%
  3        :     6.8%
  2(2)     :    30.8%
  2(1)     :    56.8%
  1        :     4.8%
```

DISCUSSION

Eight groups were identified through categorical variable consolidation, these groups will be renamed as follows:

(1) CourseGP1	(2) CourseGP2
(3) CourseGP3	(4) CourseGP4
(5) CourseGP5	(6) CourseGP6
(7) CourseGP7	(8) CourseGP8

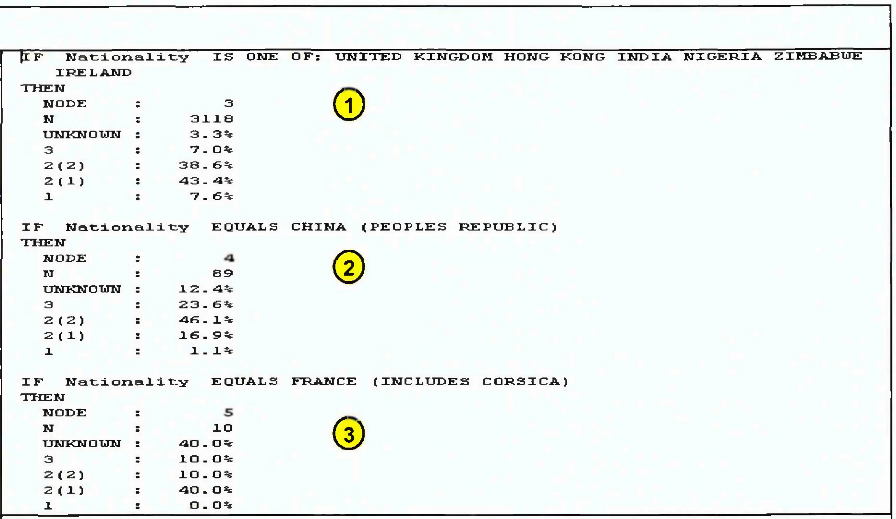
INCLUSION IN FINAL MODEL

Yes

MODIFY

Yes - A Decision Tree will be used to collapse the variable into a new variable called CourseGP. The Decision Tree produces a new variable called _Node_ this will be renamed, using a Transform Variable node, to CourseGP and the eight individual values will be replaced with the values above using a replacement node.

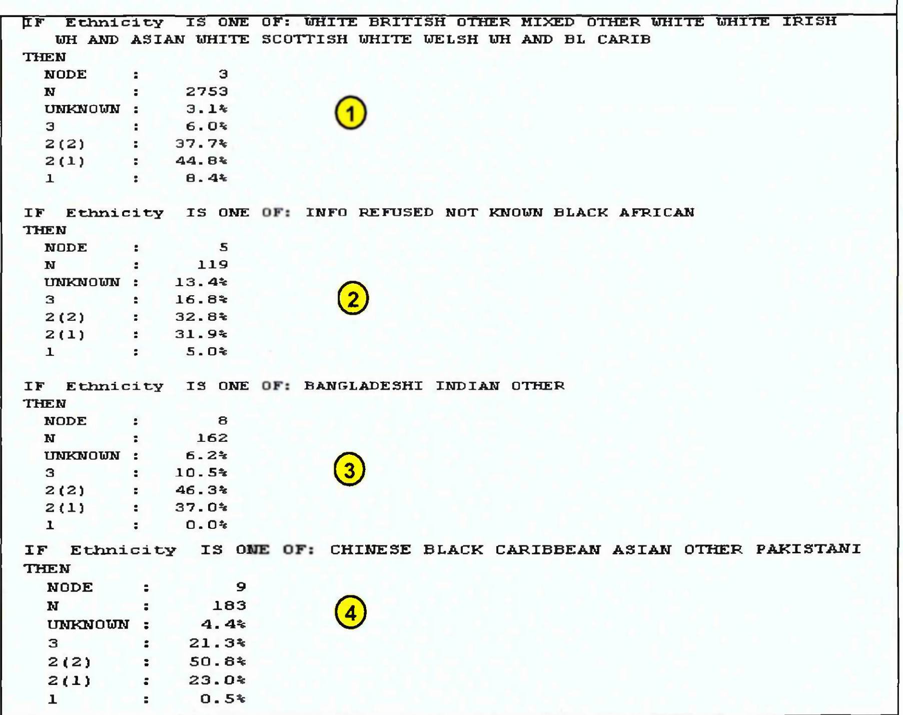
<table>
<tr><th colspan="2">NATIONALITY</th></tr>
<tr><td>VARIABLE CONSOLIDATION</td><td>



```
IF   Nationality   IS ONE OF: UNITED KINGDOM HONG KONG INDIA NIGERIA ZIMBABWE
   IRELAND
THEN
  NODE     :          3
  N        :       3118
  UNKNOWN  :       3.3%
  3        :       7.0%
  2(2)     :      38.6%
  2(1)     :      43.4%
  1        :       7.6%

IF   Nationality   EQUALS CHINA (PEOPLES REPUBLIC)
THEN
  NODE     :          4
  N        :         89
  UNKNOWN  :      12.4%
  3        :      23.6%
  2(2)     :      46.1%
  2(1)     :      16.9%
  1        :       1.1%

IF   Nationality   EQUALS FRANCE (INCLUDES CORSICA)
THEN
  NODE     :          5
  N        :         10
  UNKNOWN  :      40.0%
  3        :      10.0%
  2(2)     :      10.0%
  2(1)     :      40.0%
  1        :       0.0%
```

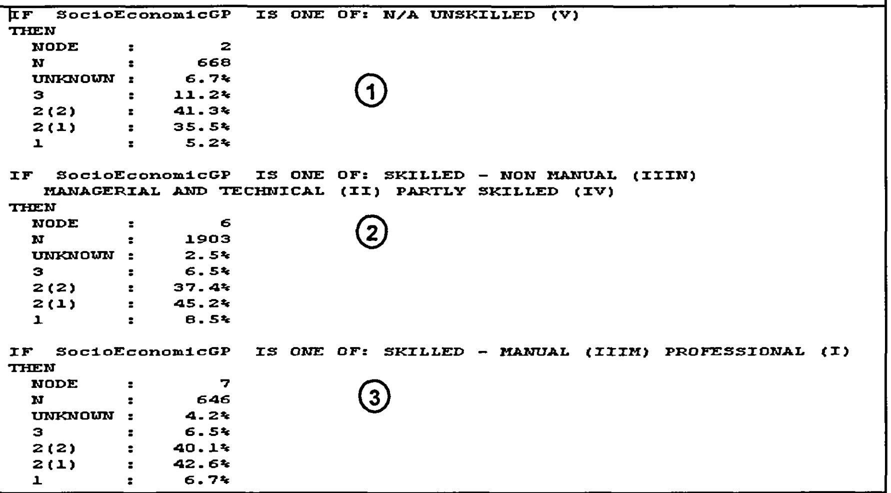
</td></tr>
<tr><td>DISCUSSION</td><td>Categorical variable consolidation identified three different groups. These will be renamed as follows: (1) UK & Others; (2) China; (3) France.</td></tr>
<tr><td>INCLUSION IN FINAL MODEL</td><td>Yes</td></tr>
<tr><td>MODIFY</td><td>Yes - the variables will be grouped using a decision tree and it will be renamed to NationalityGP, the three values will then be renamed as above.</td></tr>
</table>

<table>
<tr><th colspan="2">ETHNICITY</th></tr>
<tr><td>VARIABLE CONSOLIDATION</td><td>



```
IF   Ethnicity   IS ONE OF: WHITE BRITISH OTHER MIXED OTHER WHITE WHITE IRISH
   WH AND ASIAN WHITE SCOTTISH WHITE WELSH WH AND BL CARIB
THEN
  NODE     :          3
  N        :       2753
  UNKNOWN  :       3.1%
  3        :       6.0%
  2(2)     :      37.7%
  2(1)     :      44.8%
  1        :       8.4%

IF   Ethnicity   IS ONE OF: INFO REFUSED NOT KNOWN BLACK AFRICAN
THEN
  NODE     :          5
  N        :        119
  UNKNOWN  :      13.4%
  3        :      16.8%
  2(2)     :      32.8%
  2(1)     :      31.9%
  1        :       5.0%

IF   Ethnicity   IS ONE OF: BANGLADESHI INDIAN OTHER
THEN
  NODE     :          8
  N        :        162
  UNKNOWN  :       6.2%
  3        :      10.5%
  2(2)     :      46.3%
  2(1)     :      37.0%
  1        :       0.0%
IF   Ethnicity   IS ONE OF: CHINESE BLACK CARIBBEAN ASIAN OTHER PAKISTANI
THEN
  NODE     :          9
  N        :        183
  UNKNOWN  :       4.4%
  3        :      21.3%
  2(2)     :      50.8%
  2(1)     :      23.0%
  1        :       0.5%
```

</td></tr>
<tr><td>DISCUSSION</td><td>The results of the variable consolidation identified 4 groups, which will be renamed as follows: (1) White & Mixed; (2) Unknown & Black African; (3) Indian, Bangladeshi & Other (4) Asian, Black Caribbean & Pakistani.</td></tr>
<tr><td>INCLUSION IN FINAL MODEL</td><td>Yes</td></tr>
<tr><td>MODIFY</td><td>Yes - the variables will be grouped using a decision tree and it will be renamed to EthnicityGP, the four values will then be renamed as above.</td></tr>
</table>

<table>
<tr><th colspan="2">SOCIOECONOMICGP</th></tr>
<tr><td>VARIABLE CONSOLIDATION</td><td>


```
IF  SocioEconomicGP  IS ONE OF: N/A UNSKILLED (V)
THEN
  NODE     :        2
  N        :      668
  UNKNOWN  :     6.7%          ①
  3        :    11.2%
  2(2)     :    41.3%
  2(1)     :    35.5%
  1        :     5.2%

IF  SocioEconomicGP  IS ONE OF: SKILLED - NON MANUAL (IIIN)
   MANAGERIAL AND TECHNICAL (II) PARTLY SKILLED (IV)
THEN
  NODE     :        6
  N        :     1903          ②
  UNKNOWN  :     2.5%
  3        :     6.5%
  2(2)     :    37.4%
  2(1)     :    45.2%
  1        :     8.5%

IF  SocioEconomicGP  IS ONE OF: SKILLED - MANUAL (IIIM) PROFESSIONAL (I)
THEN
  NODE     :        7
  N        :      646          ③
  UNKNOWN  :     4.2%
  3        :     6.5%
  2(2)     :    40.1%
  2(1)     :    42.6%
  1        :     6.7%
```
</td></tr>
<tr><td>DISCUSSION</td><td>Variable consolidation found three groups these will be renamed as follows:
(1) Unknown/Unskilled; (2) Skilled(NM&P) & MT; (3) Skilled(M) & Prof</td></tr>
<tr><td>INCLUSION IN FINAL MODEL</td><td>Yes</td></tr>
<tr><td>MODIFY</td><td>Yes - the variables will be grouped using a decision tree and it will be renamed to SocioGP, the four values will then be renamed as above.</td></tr>
</table>

Ultimately, this resulted in the compilation of the following table below.

Variable	Rejected	Original Values	New Grouped Values	Missing Values	Replacement Missing Value
AwardClass	NO	YES	N/A	YES	UNKNOWN
EntryAge	NO	YES	N/A	NO	MISSING
CourseGP	NO	NO	(1) COURSEGP1 (2) COURSEGP2 (3) COURSEGP3 (4) COURSEGP4 (5) COURSEGP5 (6) COURSEGP6 (7) COURSEGP7 (8) COURSEGP8	YES	MISSING
Disability	NO	YES	N/A	NO	N/A
Gender	NO	YES	N/A	NO	N/A
EntryPoints	NO	YES	N/A	YES	MISSING
NationalityGP	NO	NO	(1) UK & OTHERS (2) CHINA (3) FRANCE	NO	N/A
EthnicityGP	NO	NO	(1) WHITE & MIXED (2) UNKNOWN & BLK AF (3) IND, BANGLA & OTHER (4) ASIAN, BLK CAR & PAK	NO	N/A
SocioGP	NO	NO	(1) UNKNOWN/UNSKILLED (2) SKILLED(NM,P) & MT (3) SKILLED(M) & PROF	NO	N/A
QYPR	NO	YES	N/A	YES	0
QAHE	NO	YES	N/A	YES	0
Entry Qualifications	YES	N/A	N/A	N/A	N/A
Faculty	YES	N/A	N/A	N/A	N/A
JACS_Subject	YES	N/A	N/A	N/A	N/A
AwardMark	YES	N/A	N/A	N/A	N/A
Home Postcode	YES	N/A	N/A	N/A	N/A
LEA	YES	N/A	N/A	N/A	N/A

Figure 7.2 - Final Data Values Used for Modelling Award Classification.

The table above, figure 7.2, is split into two areas. The green area reflects those variables that were identified through the data understanding and sampling process as being significant in predicting award classification. The red area relates to those variables that will not be included at the modelling stage. The table also highlights replacement values for grouping the data and dealing with missing values, which will be rectified in SAS® Enterprise Miner.

7.1.1.4 ACM MODEL

The modifications discussed previously were implemented and the new grouped values were replaced as per figure 7.2. Given that the target variable is defined, a number of supervised DM techniques (see sub-section 4.3.2) were applied. The selection of DM techniques, applied at the modelling stage, were based on the experiences of previous EDM studies (Superby *et al.* (2006), Romero *et al.* (2008), Dekker *et al.* (2009)). Therefore, three decision tree models (Entropy, Gini and ChiSquare) and three logistic regression models (using backwards, forwards and stepwise selection methods) were built. Figure 7.3, below shows the process up to and including the supervised DM techniques.

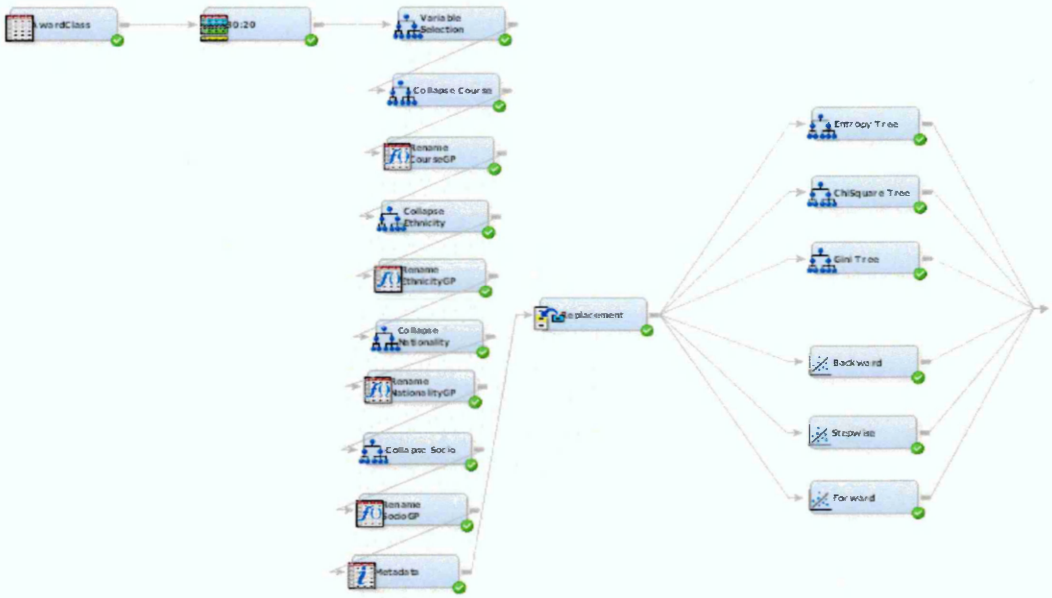
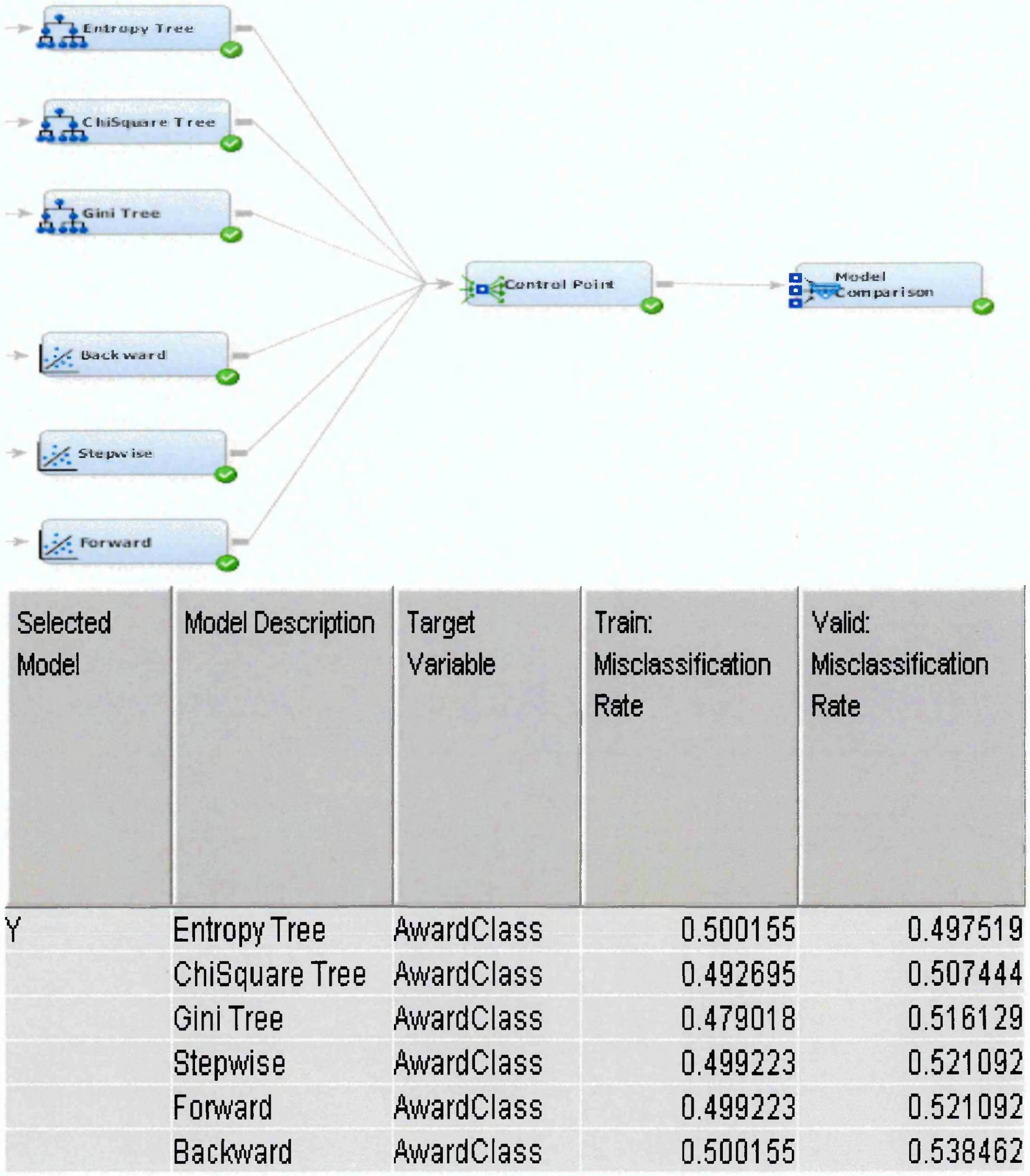


Figure 7.3 - SAS Process Flow.

7.1.1.5 ACM ASSESS

The modelling of the data went through numerous adjustments to determine the best model, which involved removing variables, changing sample sizes and model settings. Arguably, including details of all these different tests would undoubtedly affect the readability of this section. Therefore, this section presents the assessment of the best model. A control point and a model comparison node were then added to assess each model, see below.



Selected Model	Model Description	Target Variable	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Entropy Tree	AwardClass	0.500155	0.497519
	ChiSquare Tree	AwardClass	0.492695	0.507444
	Gini Tree	AwardClass	0.479018	0.516129
	Stepwise	AwardClass	0.499223	0.521092
	Forward	AwardClass	0.499223	0.521092
	Backward	AwardClass	0.500155	0.538462

Figure 7.4 – Model Comparison Node.

Whilst the model comparison node generates numerous results (above), the important results are the train and valid misclassification rates as these are measures of the amount of prediction error in the model. The model comparison node suggests that the Entropy Tree produces the best model. However, the assessment of misclassification rate (train and valid) indicates that this model would not be a good predictor of award classification as the model incorrectly predicted about 50% of the values in the training data and 49% of the validation data. However, high levels of inaccuracy are common in such research studies that have tried to predict student behaviour (Superby *et al.* 2006, Romero *et al.* 2008, Herzog 2006).

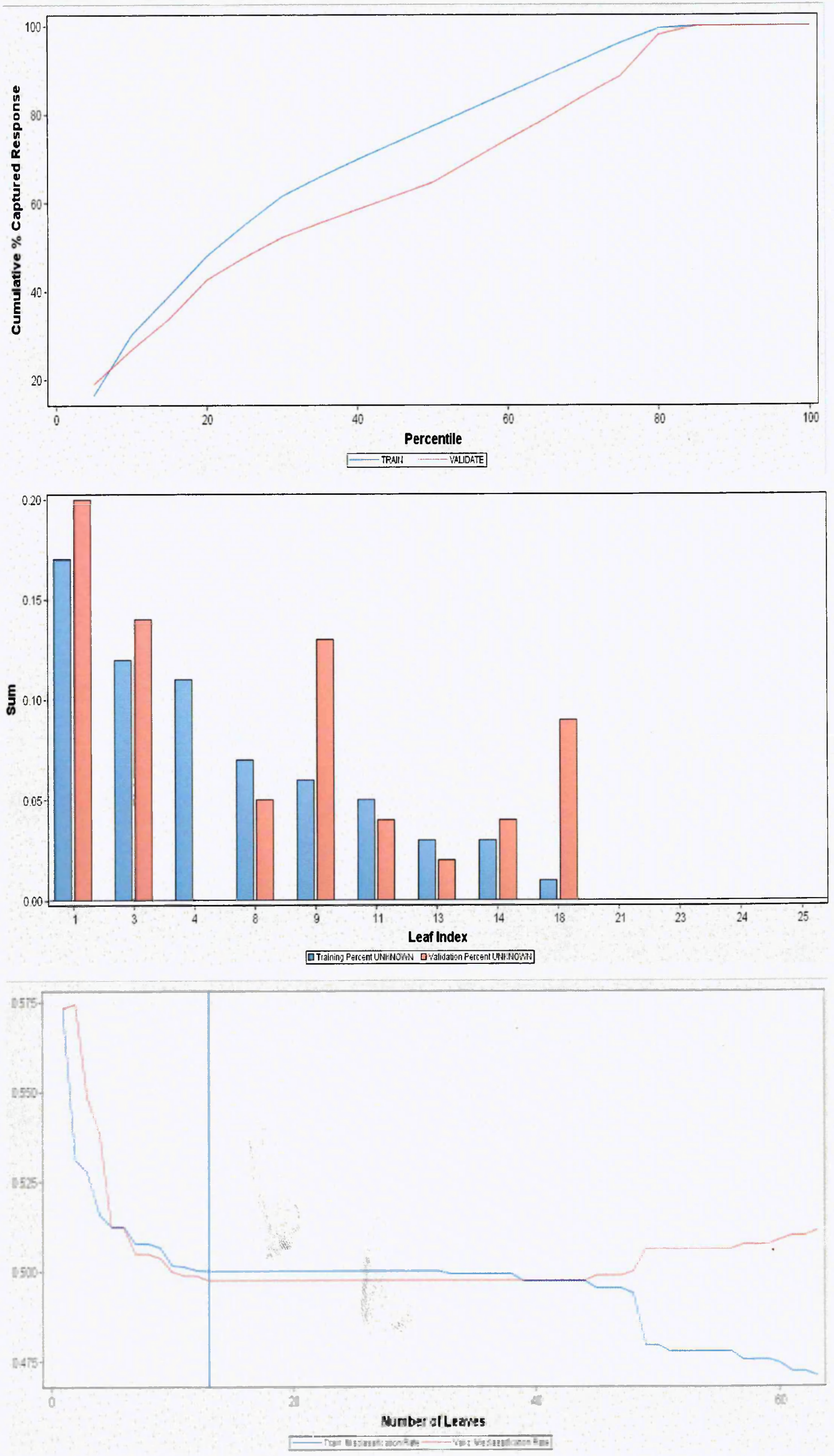
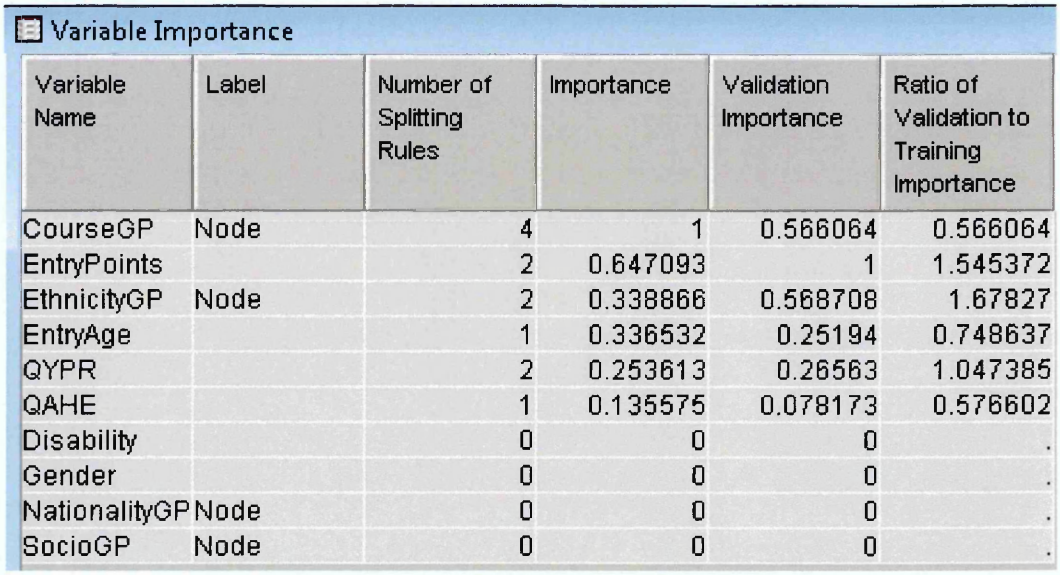


Figure 7.5 – Model Assessment.

The %captured response chart (top), shows that the top 20% of responses captures over 40% of the award classifications for the validation data set and over 50% for the train data set. Whilst these translate into reasonably high lift (see glossary page ix) rates, there is a large amount of variation between the train and validation lift rates. The leaf index bar chart (middle) shows the amount of training and validation data in each leaf of the final model. Ideally the model should produce an equal sum of training and validation in each leaf. Consequently, it is easy to see why the misclassification rates are so high in the misclassification chart (bottom). This could be improved by increasing the amount of training and validation data but a larger data set wasn't available. Henceforth, the quality of the model can be judged by looking at the variables selected by the tree (below) in conjunction with previous research carried out into student retention and EDM.



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
CourseGP	Node	4	1	0.566064	0.566064
EntryPoints		2	0.647093	1	1.545372
EthnicityGP	Node	2	0.338866	0.568708	1.67827
EntryAge		1	0.336532	0.25194	0.748637
QYPR		2	0.253613	0.26563	1.047385
QAHE		1	0.135575	0.078173	0.576602
Disability		0	0	0	.
Gender		0	0	0	.
NationalityGP	Node	0	0	0	.
SocioGP	Node	0	0	0	.

Figure 7.6 – Variable Importance Table.

The variables selected by the model (above), to predict award classification, are sensible. Indeed, some of the variables such as course, previous academic experience and entry age have all been identified as key variables in previous EDM studies (Superby *et al.* 2006, Dekker *et al.* 2009, Herzog 2006). The identification of additional variables (EthnicityGP, QYPR and QAHE) is hardly surprising, as "studies show that academic success is dependent on many factors [...]" (Dekker *et al.* 2009, p43).

7.1.1.6 THE FINAL ACM

Figure 7.7, below, provides an overview of the final Entropy tree model.

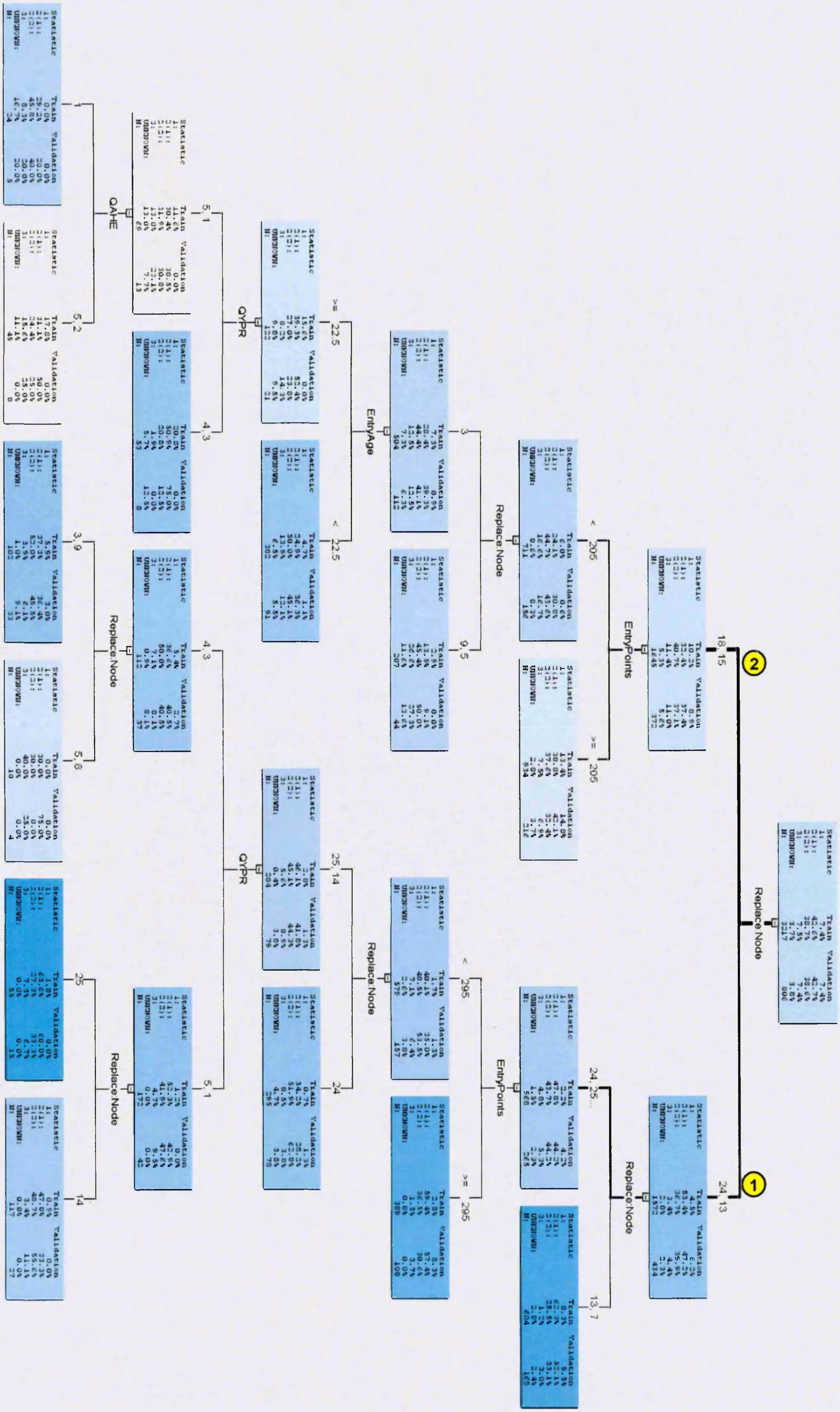
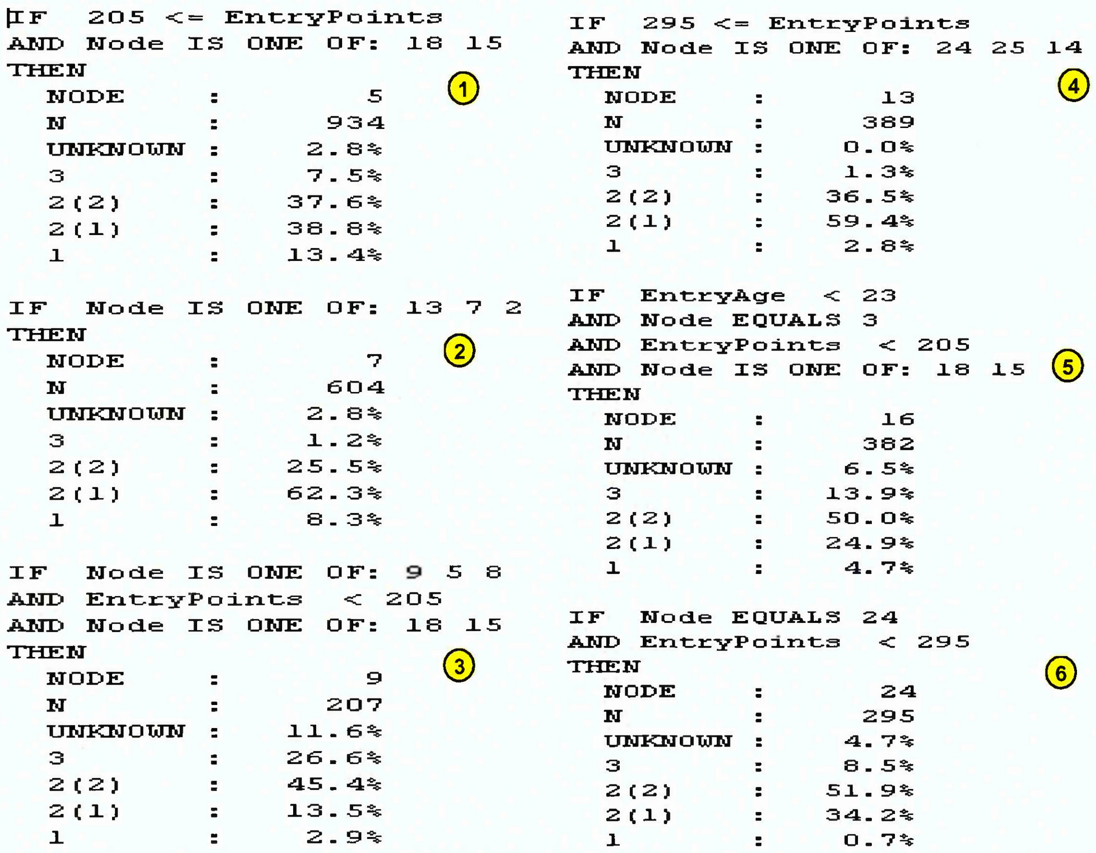


Figure 7.7 - Final Entropy Model.

Due to problems with the renaming of the label node, in SAS® Enterprise Miner, this section will provide a brief overview of the final tree, the individual rules of each leaf will then be discussed below. The tree starts by splitting CourseGP into two groups, moving down the right side of the tree (1), the CourseGP variable is split again into two groups. This is then split by EntryPoints and then CourseGP. One of the CourseGP strands are then split by QYPR, into two, these are then further split by EthnicityGP and CourseGP. Moving down the left side of the tree (2) the CourseGP variable is the then split by EntryPoints. This is split again by EthnicityGP then EntryAge and QYPR then QAHE. The tree outlined above produces the following rules:



```
IF  205 <= EntryPoints
AND Node IS ONE OF: 18 15
THEN
   NODE     :       5      ①
   N        :     934
   UNKNOWN  :    2.8%
   3        :    7.5%
   2(2)     :   37.6%
   2(1)     :   38.8%
   1        :   13.4%
```

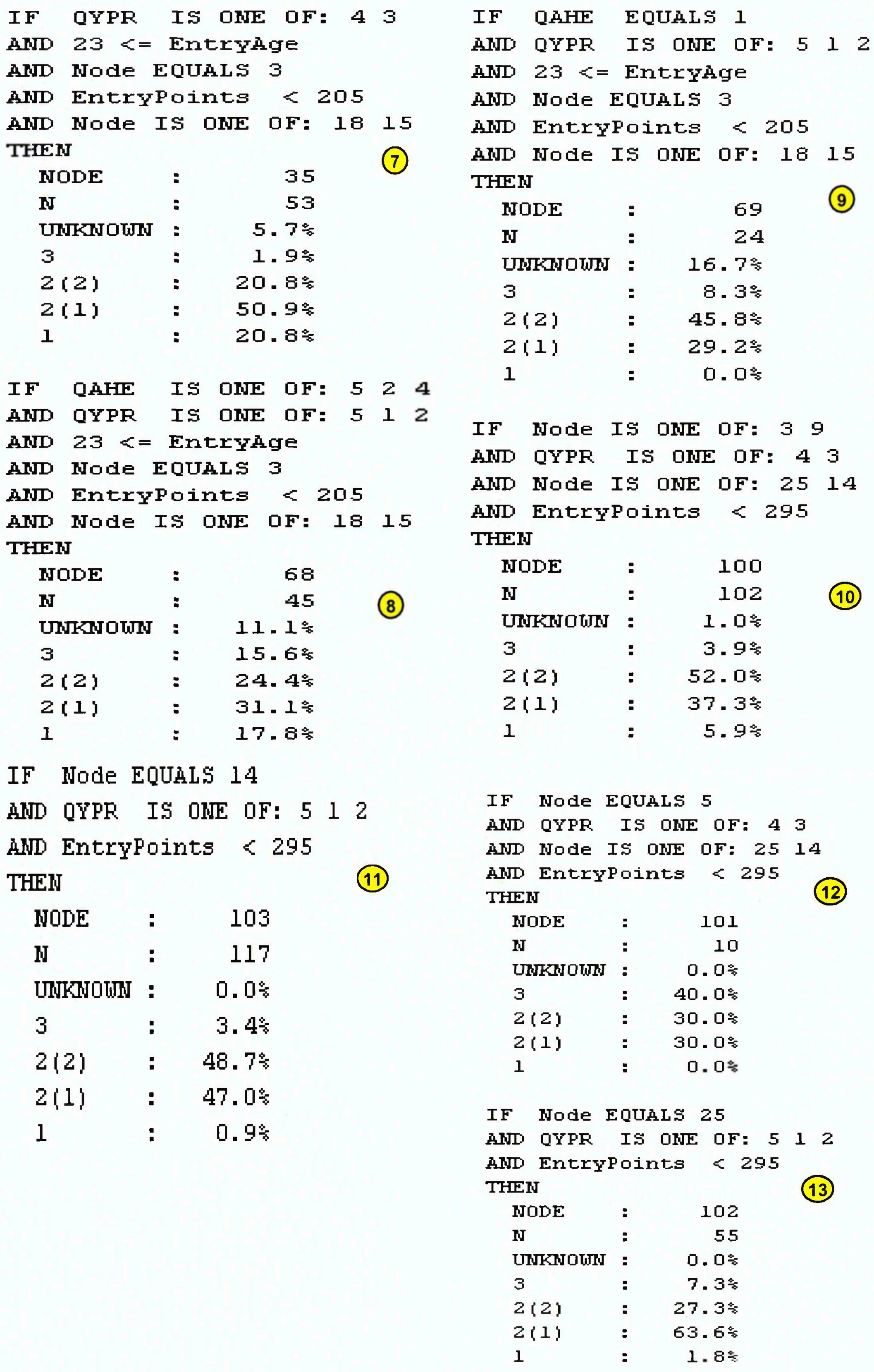
```
IF  Node IS ONE OF: 13 7 2
THEN
   NODE     :       7      ②
   N        :     604
   UNKNOWN  :    2.8%
   3        :    1.2%
   2(2)     :   25.5%
   2(1)     :   62.3%
   1        :    8.3%
```

```
IF  Node IS ONE OF: 9 5 8
AND EntryPoints  < 205
AND Node IS ONE OF: 18 15
THEN
   NODE     :       9      ③
   N        :     207
   UNKNOWN  :   11.6%
   3        :   26.6%
   2(2)     :   45.4%
   2(1)     :   13.5%
   1        :    2.9%
```

```
IF  295 <= EntryPoints
AND Node IS ONE OF: 24 25 14
THEN
   NODE     :      13      ④
   N        :     389
   UNKNOWN  :    0.0%
   3        :    1.3%
   2(2)     :   36.5%
   2(1)     :   59.4%
   1        :    2.8%
```

```
IF  EntryAge  < 23
AND Node EQUALS 3
AND EntryPoints  < 205
AND Node IS ONE OF: 18 15  ⑤
THEN
   NODE     :      16
   N        :     382
   UNKNOWN  :    6.5%
   3        :   13.9%
   2(2)     :   50.0%
   2(1)     :   24.9%
   1        :    4.7%
```

```
IF  Node EQUALS 24
AND EntryPoints  < 295
THEN
   NODE     :      24      ⑥
   N        :     295
   UNKNOWN  :    4.7%
   3        :    8.5%
   2(2)     :   51.9%
   2(1)     :   34.2%
   1        :    0.7%
```



```
IF  QYPR  IS ONE OF: 4 3
AND 23 <= EntryAge
AND Node EQUALS 3
AND EntryPoints  < 205
AND Node IS ONE OF: 18 15
THEN
  NODE     :       35
  N        :       53
  UNKNOWN  :     5.7%
  3        :     1.9%
  2(2)     :    20.8%
  2(1)     :    50.9%
  1        :    20.8%
```

```
IF  QAHE  IS ONE OF: 5 2 4
AND QYPR  IS ONE OF: 5 1 2
AND 23 <= EntryAge
AND Node EQUALS 3
AND EntryPoints  < 205
AND Node IS ONE OF: 18 15
THEN
  NODE     :       68
  N        :       45
  UNKNOWN  :    11.1%
  3        :    15.6%
  2(2)     :    24.4%
  2(1)     :    31.1%
  1        :    17.8%
```

```
IF  Node EQUALS 14
AND QYPR  IS ONE OF: 5 1 2
AND EntryPoints  < 295
THEN
  NODE     :      103
  N        :      117
  UNKNOWN  :     0.0%
  3        :     3.4%
  2(2)     :    48.7%
  2(1)     :    47.0%
  1        :     0.9%
```

```
IF  QAHE  EQUALS 1
AND QYPR  IS ONE OF: 5 1 2
AND 23 <= EntryAge
AND Node EQUALS 3
AND EntryPoints  < 205
AND Node IS ONE OF: 18 15
THEN
  NODE     :       69
  N        :       24
  UNKNOWN  :    16.7%
  3        :     8.3%
  2(2)     :    45.8%
  2(1)     :    29.2%
  1        :     0.0%
```

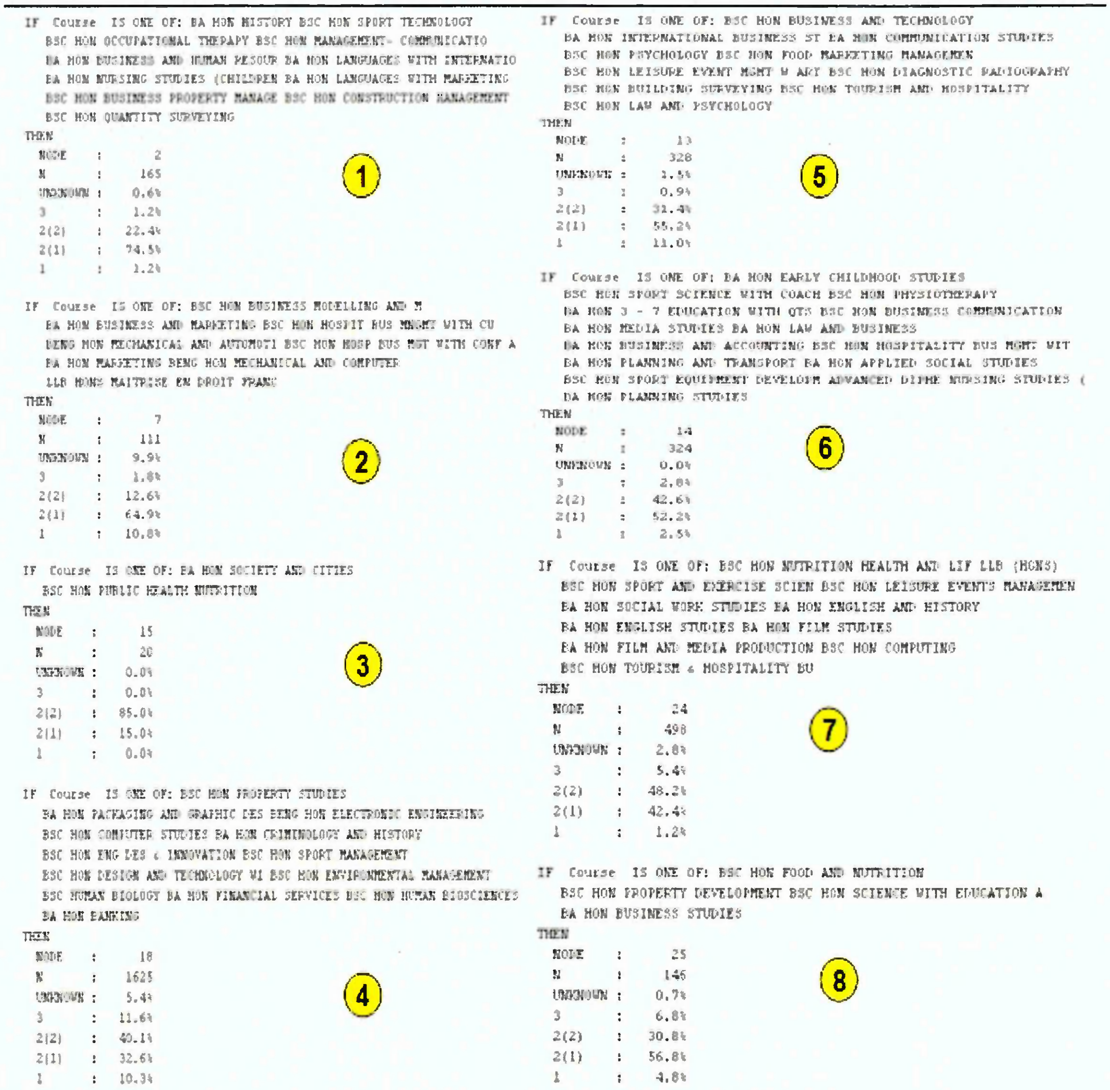
```
IF  Node IS ONE OF: 3 9
AND QYPR  IS ONE OF: 4 3
AND Node IS ONE OF: 25 14
AND EntryPoints  < 295
THEN
  NODE     :      100
  N        :      102
  UNKNOWN  :     1.0%
  3        :     3.9%
  2(2)     :    52.0%
  2(1)     :    37.3%
  1        :     5.9%
```

```
IF  Node EQUALS 5
AND QYPR  IS ONE OF: 4 3
AND Node IS ONE OF: 25 14
AND EntryPoints  < 295
THEN
  NODE     :      101
  N        :       10
  UNKNOWN  :     0.0%
  3        :    40.0%
  2(2)     :    30.0%
  2(1)     :    30.0%
  1        :     0.0%
```

```
IF  Node EQUALS 25
AND QYPR  IS ONE OF: 5 1 2
AND EntryPoints  < 295
THEN
  NODE     :      102
  N        :       55
  UNKNOWN  :     0.0%
  3        :     7.3%
  2(2)     :    27.3%
  2(1)     :    63.6%
  1        :     1.8%
```

Figure 7.8 - Final Rules.

In order to aid in the understanding of these rules, the course groupings have been repeated below from the modify section.



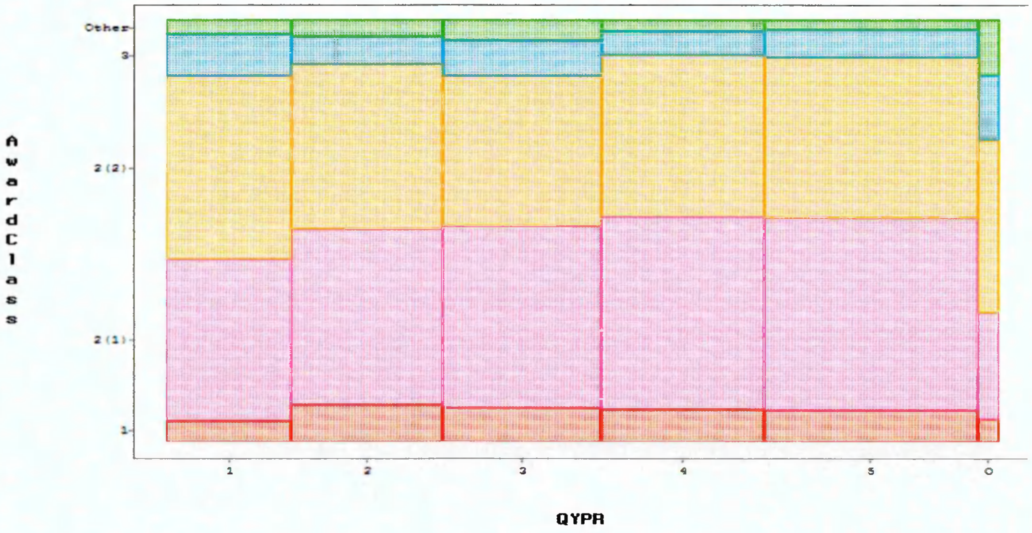
```
IF  Course  IS ONE OF: BA HON HISTORY BSC HON SPORT TECHNOLOGY
   BSC HON OCCUPATIONAL THEPAPY BSC HON MANAGEMENT- COMMUNICATIO
   BA HON BUSINESS AND HUMAN RESOUR BA HON LANGUAGES WITH INTERNATIO
   BA HON NURSING STUDIES (CHILDREN BA HON LANGUAGES WITH MARKETING
   BSC HON BUSINESS PROPERTY MANAGE BSC HON CONSTRUCTION MANAGEMENT
   BSC HON QUANTITY SURVEYING
THEN
  NODE     :       2
  N        :     165
  UNKNOWN  :    0.6%
  3        :    1.2%
  2(2)     :   22.4%
  2(1)     :   74.5%
  1        :    1.2%

IF  Course  IS ONE OF: BSC HON BUSINESS MODELLING AND M
   BA HON BUSINESS AND MARKETING BSC HON HOSPIT BUS MNGMT WITH CU
   BENG HON MECHANICAL AND AUTOMOTI BSC HON HOSP BUS MGT WITH CONF A
   BA HON MARKETING BENG HON MECHANICAL AND COMPUTER
   LLB HONS MAITRISE EN DROIT FRANC
THEN
  NODE     :       7
  N        :     111
  UNKNOWN  :    9.9%
  3        :    1.8%
  2(2)     :   12.6%
  2(1)     :   64.9%
  1        :   10.8%

IF  Course  IS ONE OF: BA HON SOCIETY AND CITIES
   BSC HON PUBLIC HEALTH NUTRITION
THEN
  NODE     :      15
  N        :      20
  UNKNOWN  :    0.0%
  3        :    0.0%
  2(2)     :   85.0%
  2(1)     :   15.0%
  1        :    0.0%

IF  Course  IS ONE OF: BSC HON PROPERTY STUDIES
   BA HON PACKAGING AND GRAPHIC DES BENG HON ELECTRONIC ENGINEERING
   BSC HON COMPUTER STUDIES BA HON CRIMINOLOGY AND HISTORY
   BSC HON ENG DES & INNOVATION BSC HON SPORT MANAGEMENT
   BSC HON DESIGN AND TECHNOLOGY WI BSC HON ENVIRONMENTAL MANAGEMENT
   BSC HUMAN BIOLOGY BA HON FINANCIAL SERVICES BSC HON HUMAN BIOSCIENCES
   BA HON BANKING
THEN
  NODE     :      18
  N        :    1625
  UNKNOWN  :    5.4%
  3        :   11.6%
  2(2)     :   40.1%
  2(1)     :   32.6%
  1        :   10.3%

IF  Course  IS ONE OF: BSC HON BUSINESS AND TECHNOLOGY
   BA HON INTERNATIONAL BUSINESS ST BA HON COMMUNICATION STUDIES
   BSC HON PSYCHOLOGY BSC HON FOOD MARKETING MANAGEMEN
   BSC HON LEISURE EVENT MGMT W ART BSC HON DIAGNOSTIC RADIOGRAPHY
   BSC HON BUILDING SURVEYING BSC HON TOURISM AND HOSPITALITY
   BSC HON LAW AND PSYCHOLOGY
THEN
  NODE     :      13
  N        :     328
  UNKNOWN  :    1.5%
  3        :    0.9%
  2(2)     :   31.4%
  2(1)     :   55.2%
  1        :   11.0%

IF  Course  IS ONE OF: BA HON EARLY CHILDHOOD STUDIES
   BSC HON SPORT SCIENCE WITH COACH BSC HON PHYSIOTHERAPY
   BA HON 3 - 7 EDUCATION WITH QTS BSC HON BUSINESS COMMUNICATION
   BA HON MEDIA STUDIES BA HON LAW AND BUSINESS
   BA HON BUSINESS AND ACCOUNTING BSC HON HOSPITALITY BUS MGMT WIT
   BA HON PLANNING AND TRANSPORT BA HON APPLIED SOCIAL STUDIES
   BSC HON SPORT EQUIPMENT DEVELOPM ADVANCED DIPHE NURSING STUDIES (
   BA HON PLANNING STUDIES
THEN
  NODE     :      14
  N        :     324
  UNKNOWN  :    0.0%
  3        :    2.8%
  2(2)     :   42.6%
  2(1)     :   52.2%
  1        :    2.5%

IF  Course  IS ONE OF: BSC HON NUTRITION HEALTH AND LIF LLB (HONS)
   BSC HON SPORT AND EXERCISE SCIEN BSC HON LEISURE EVENTS MANAGEMEN
   BA HON SOCIAL WORK STUDIES BA HON ENGLISH AND HISTORY
   BA HON ENGLISH STUDIES BA HON FILM STUDIES
   BA HON FILM AND MEDIA PRODUCTION BSC HON COMPUTING
   BSC HON TOURISM & HOSPITALITY BU
THEN
  NODE     :      24
  N        :     498
  UNKNOWN  :    2.8%
  3        :    5.4%
  2(2)     :   48.2%
  2(1)     :   42.4%
  1        :    1.2%

IF  Course  IS ONE OF: BSC HON FOOD AND NUTRITION
   BSC HON PROPERTY DEVELOPMENT BSC HON SCIENCE WITH EDUCATION A
   BA HON BUSINESS STUDIES
THEN
  NODE     :      25
  N        :     146
  UNKNOWN  :    0.7%
  3        :    6.8%
  2(2)     :   30.8%
  2(1)     :   56.8%
  1        :    4.8%
```

These rules can be interpreted as follows:

Rule (1)	If EntryPoints is greater than 205 and CourseGP is CourseGP4 or CourseGP3 then students are more likely to achieve a 2:1 or a 2:2. Also 13.4% of students in this group obtain a 1st.
Rule (2)	If CourseGP is one of CourseGP5, CourseGP2 or CourseGP1 then students are more likely to obtain a 2:1.
Rule (3)	If EthnicityGP is Unknown & Black African or Indian, Bangladeshi & Other or Asian, Black Caribbean & Pakistani and EntryPoints is less than 205 and CourseGP is CourseGP3 or CourseGP4 then students are more likely to obtain a 2:2.
Rule (4)	If EntryPoints is greater than or equal to 295 and CourseGP is one of CourseGP6, CourseGP7 or CourseGP8 then the award classification is more like to be 2:1.
Rule (5)	If EntryAge is less than 23 and EthnicityGP is White & Mixed and EntryPoints is less than 205 and CourseGP is one of CourseGP4 or CourseGP3 then students are more likely to obtain a 2:2.
Rule (6)	If CourseGP is CourseGP7 and EntryPoints are less than 295 then a 2:2 is more likely.
Rule (7)	If QYPR is Medium-High or Medium and EntryAge is greater than or equal to 23 and EthnicityGP is White & Mixed and EntryPoints is greater than 205 and CourseGP is one of CourseGP3 or CourseGP4 then students are more likely to obtain a 2:1. This group also has the highest number of students obtaining a 1st (20.8%).

Rule (8)	If QAHE is one of High, Medium-High or Low-Medium and QYPR is High, Low-Medium or Low and EntryAge is greater than or equal to 23 and EthnicityGP is White & Mixed and EntryPoints in less than 205 and CourseGP is one of CourseGP3 or CourseGP4. Then students are more likely to obtain a 2:1. In addition to this 17.8% of students, in this group, obtain a 1st.
Rule (9)	If QAHE is Low and QYPR is High, Low, or Low-Medium and EntryAge is greater than or equal to 23 and EthnicityGP is White & Mixed and EntryPoints is less than 205 and CourseGP is CourseGP3 or CourseGP4 then the likely award classification is a 2:2.
Rule (10)	If EthnicityGP is White & Mixed or Asian, Black Caribbean & Pakistani and QYPR is one of Medium-High or Medium and CourseGP is CourseGP6 or CourseGP8 and EntryPoints are less than 295. Then students are more likely to achieve a 2:2.
Rule (11)	If CourseGP is CourseGP6 and QYPR is one of High, Low, or Low-Medium and EntryPoints are less than 295. Then students are almost equally likely to obtain a 2:1 (47%) or a 2:2 (48.7%).
Rule (12)	If EthnicityGP is Unknown & Black African and QYPR is Medium-High or Medium and CourseGP is CourseGP6 or CourseGP8 and EntryPoints are less than 295. The students are more likely to obtain a 3rd
Rule (13)	If CourseGP is CourseGP8 and QYPR is High, Low, or Low-Medium and EntryPoints is less than 295 than students are more likely to obtain a 2:1.

The rules above suggest that non-white ethnic groups are less likely to achieve a high award classification. Therefore, it is possible that there might be a dependency between student's ethnicity and award mark. Overall, these rules are sensible when compared to exploration of the data (outlined above). Indeed, assessing rule 7 as an example (as this has the largest number of 1st classifications), in relation to the data exploration, it is apparent from the graphs (represented below) that QYPR 3 and 4 have the second highest number of 1st classifications, however, these groups have the largest number of students. The EntryAge graph indicates that the majority of students obtaining a 1st are around 20.



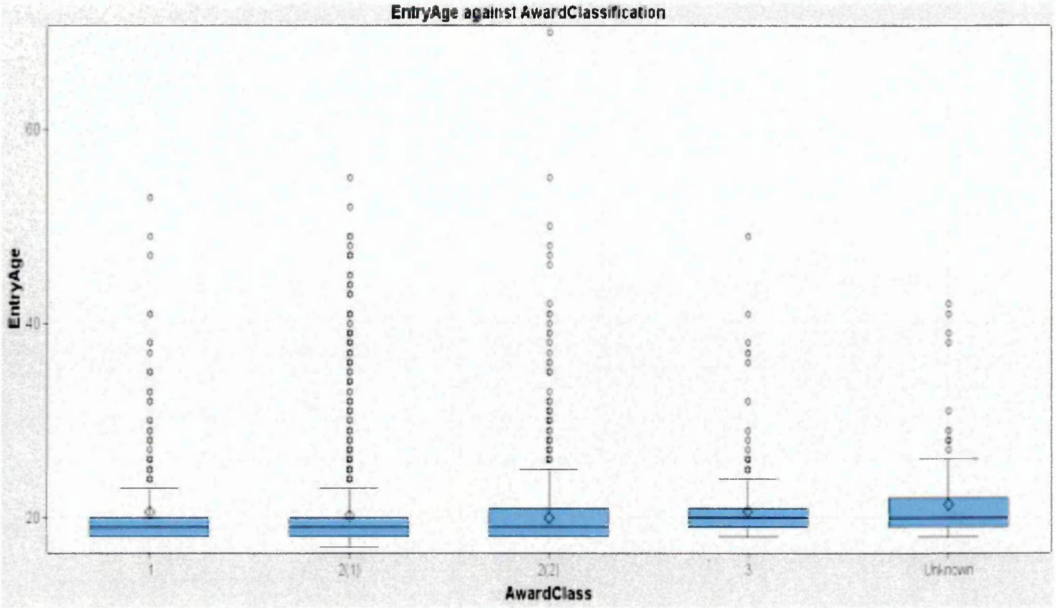
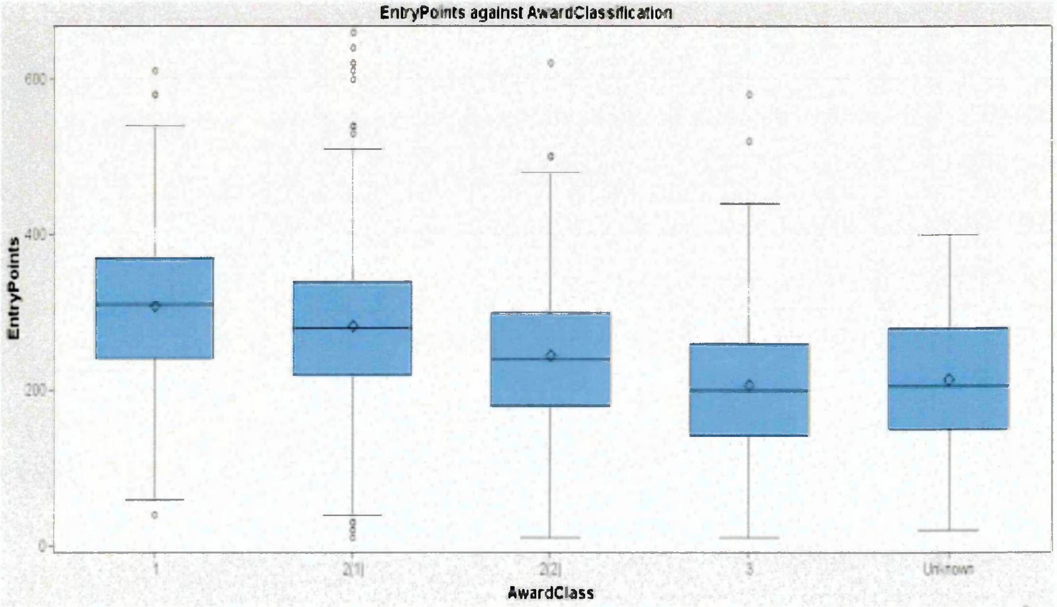


Figure 7.9 – Assessing Rule 7 Plots.

The EntryPoints box plot below also confirms that the majority of students obtaining a 1st entered onto their undergraduate degree with around 300 points and the majority of white students are also in EthnicityGP 3, including White British.



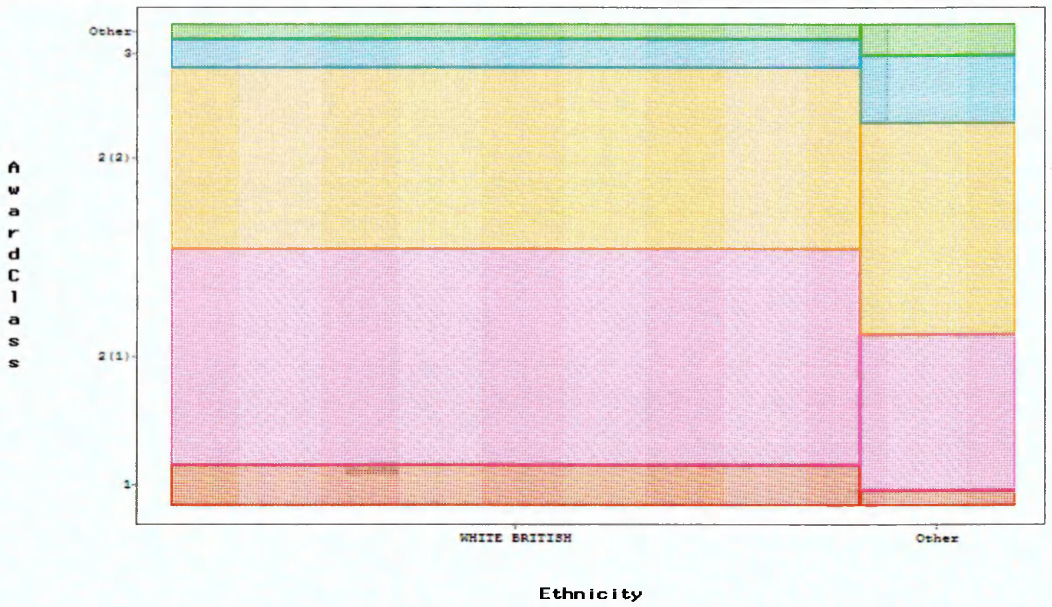


Figure 7.10 – Assessment other Rule 7 Plots.

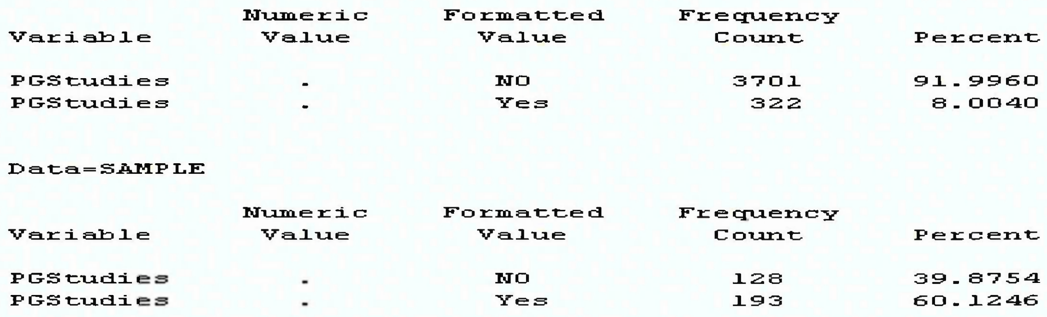
Therefore, the final model provides a good representation of the underlying data.

7.1.2 Building the Postgraduate Studies Model (PSM)

This section details the process followed in building the postgraduate studies DM model using SAS® SEMMA.

7.1.2.1 PSM Sample

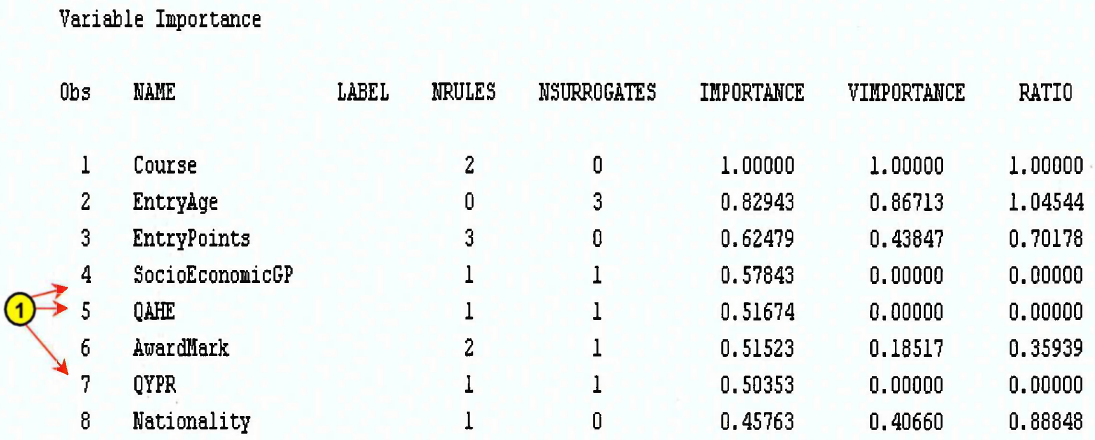
Of the 4023 students in the data set 322 students went on to study a postgraduate degree at SHU, which is an 8% event rate. Therefore, prior probabilities were created and oversampling was carried out on the data as this created a data set where there were 193 "Yes" and 128 "No" events - a total of 321 students.



```
               Numeric     Formatted     Frequency
Variable        Value        Value         Count        Percent

PGStudies         .           NO           3701         91.9960
PGStudies         .           Yes           322          8.0040

Data=SAMPLE

               Numeric     Formatted     Frequency
Variable        Value        Value         Count        Percent

PGStudies         .           NO            128         39.8754
PGStudies         .           Yes           193         60.1246
```

Figure 7.11 – Oversampling the Rare Event.

A 80:20 data partition was also setup so that the data could be divided into training and validation respectively. Again the variables within the data set were reduced through the authors own knowledge of the data/HE and by using a decision tree for variable selection. This resulted in the following inputs being selected for modelling:



```
Variable Importance

Obs   NAME              LABEL   NRULES   NSURROGATES   IMPORTANCE   VIMPORTANCE   RATIO

 1    Course                       2          0          1.00000      1.00000     1.00000
 2    EntryAge                     0          3          0.82943      0.86713     1.04544
 3    EntryPoints                  3          0          0.62479      0.43847     0.70178
 4    SocioEconomicGP              1          1          0.57843      0.00000     0.00000
 5    QAHE                         1          1          0.51674      0.00000     0.00000
 6    AwardMark                    2          1          0.51523      0.18517     0.35939
 7    QYPR                         1          1          0.50353      0.00000     0.00000
 8    Nationality                  1          0          0.45763      0.40660     0.88848
```

Figure 7.12 – Variable Importance Table.

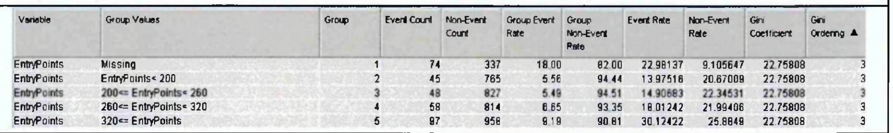
As previously discussed (1) refers to measures of social deprivation and one of the variables either SocioEconomicGP or QAHE and QYPR will have to be removed in favour of the other.

7.1.2.2 PSM Explore

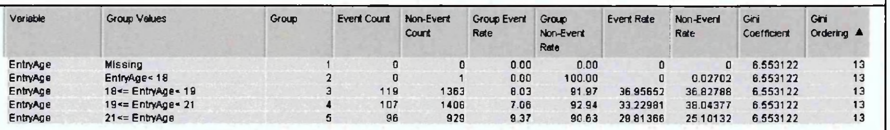
The data has already been thoroughly explored through the data understanding process outlined in Chapter 6, through this a number of graphs were produced, outliers were identified.

7.1.2.3 PSM Modify

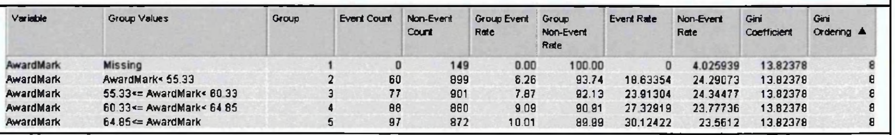
This section will focus on using categorical variable consolidation (see glossary page vii) for categorical values, interactive binning/grouping (for interval values) and the authors own knowledge of the data/HE to determine suitable groups of values in relation to the target variable.

ENTRYPOINTS	
INTERACTIVE BINNING/ GROUPING RESULTS	
DISCUSSION	Again the same five groups were identified by SAS® for EntryPoints.
INCLUSION IN FINAL MODEL	Yes
MODIFY	No – Data Mining model will determine the best split

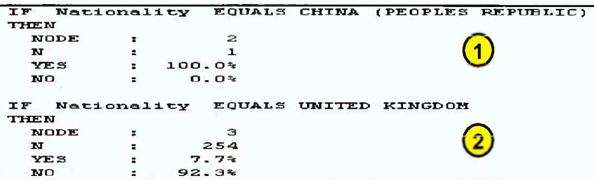
Variable	Group Values	Group	Event Count	Non-Event Count	Group Event Rate	Group Non-Event Rate	Event Rate	Non-Event Rate	Gini Coefficient	Gini Ordering ▲
EntryPoints	Missing	1	74	337	18.00	82.00	22.98137	9.105647	22.75808	3
EntryPoints	EntryPoints< 200	2	45	765	5.56	94.44	13.97516	20.67009	22.75808	3
EntryPoints	200<= EntryPoints< 260	3	48	827	5.49	94.51	14.90683	22.34531	22.75808	3
EntryPoints	260<= EntryPoints< 320	4	58	814	6.65	93.35	18.01242	21.99406	22.75808	3
EntryPoints	320<= EntryPoints	5	97	958	9.19	90.81	30.12422	25.8849	22.75808	3

ENTRYAGE	
INTERACTIVE BINNING/ GROUPING RESULTS	
DISCUSSION	The EntryAge groupings identified by the SAS® Interactive Binning/Grouping node is the same as the groups identified when the target variable was AWClassDUMMY.
INCLUSION IN FINAL MODEL	Yes
MODIFY	No – Data Mining model will determine the best split

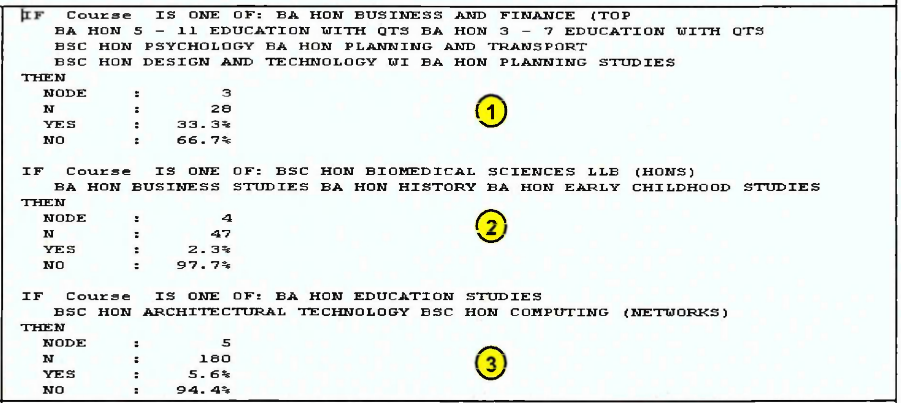
Variable	Group Values	Group	Event Count	Non-Event Count	Group Event Rate	Group Non-Event Rate	Event Rate	Non-Event Rate	Gini Coefficient	Gini Ordering ▲
EntryAge	Missing	1	0	0	0.00	0.00	0	0	6.553122	13
EntryAge	EntryAge< 18	2	0	1	0.00	100.00	0	0.02702	6.553122	13
EntryAge	18<= EntryAge< 19	3	119	1363	8.03	91.97	36.95652	36.82788	6.553122	13
EntryAge	19<= EntryAge< 21	4	107	1408	7.06	92.94	33.22981	38.04377	6.553122	13
EntryAge	21<= EntryAge	5	96	929	9.37	90.63	29.81366	25.10132	6.553122	13

AWARDMARK	
INTERACTIVE BINNING/ GROUPING RESULTS	
DISCUSSION	SAS® suggests that AwardMark could be grouped into 5 groups
INCLUSION IN FINAL MODEL	Yes
MODIFY	No – Data Mining model will determine the best split

Variable	Group Values	Group	Event Count	Non-Event Count	Group Event Rate	Group Non-Event Rate	Event Rate	Non-Event Rate	Gini Coefficient	Gini Ordering ▲
AwardMark	Missing	1	0	149	0.00	100.00	0	4.025939	13.82378	8
AwardMark	AwardMark< 55.33	2	60	899	6.26	93.74	18.63354	24.29073	13.82378	8
AwardMark	55.33<= AwardMark< 60.33	3	77	901	7.87	92.13	23.91304	24.34477	13.82378	8
AwardMark	60.33<= AwardMark< 64.85	4	88	880	9.09	90.91	27.32919	23.77736	13.82378	8
AwardMark	64.85<= AwardMark	5	97	872	10.01	89.99	30.12422	23.5612	13.82378	8

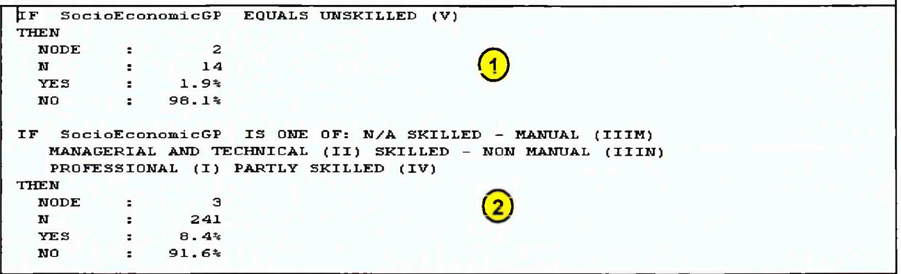
NATIONALITY	
VARIABLE CONSOLIDATION	
DISCUSSION	The consolidation of the Nationality variable determined the following two groups: (1) China; (2) UK.
INCLUSION IN FINAL MODEL	Yes
MODIFY	Yes - the variables will be grouped using a decision tree and it will be renamed to NationalityGP, the two values will then be renamed as above.

```
IF  Nationality  EQUALS  CHINA  (PEOPLES  REPUBLIC)
THEN
   NODE     :         2
   N        :         1                ①
   YES      :    100.0%
   NO       :      0.0%

IF  Nationality  EQUALS  UNITED  KINGDOM
THEN
   NODE     :         3
   N        :       254                ②
   YES      :      7.7%
   NO       :     92.3%
```

<table>
<tr><td colspan="2">Course</td></tr>
<tr><td>Variable Consolidation</td><td>

<pre>
IF Course IS ONE OF: BA HON BUSINESS AND FINANCE (TOP
   BA HON 5 - 11 EDUCATION WITH QTS BA HON 3 - 7 EDUCATION WITH QTS
   BSC HON PSYCHOLOGY BA HON PLANNING AND TRANSPORT
   BSC HON DESIGN AND TECHNOLOGY WI BA HON PLANNING STUDIES
THEN
  NODE   :   3
  N   :   28
  YES   :   33.3%
  NO   :   66.7%

IF Course IS ONE OF: BSC HON BIOMEDICAL SCIENCES LLB (HONS)
   BA HON BUSINESS STUDIES BA HON HISTORY BA HON EARLY CHILDHOOD STUDIES
THEN
  NODE   :   4
  N   :   47
  YES   :   2.3%
  NO   :   97.7%

IF Course IS ONE OF: BA HON EDUCATION STUDIES
   BSC HON ARCHITECTURAL TECHNOLOGY BSC HON COMPUTING (NETWORKS)
THEN
  NODE   :   5
  N   :   180
  YES   :   5.6%
  NO   :   94.4%
</pre></td></tr>
<tr><td>Discussion</td><td>Through variable consolidation three groups were identified these will be renamed to: (1) CourseGP1; (2) CourseGP2; (3) CourseGP3.</td></tr>
<tr><td>Inclusion in Final Model</td><td>Yes</td></tr>
<tr><td>Modify</td><td>Yes - A Decision Tree will be used to collapse the variable into a new variable called CourseGP. The Decision Tree produces a new variable called _Node_ this will be renamed, using a Transform Variable node, to CourseGP and the three individual values will be replaced with the values above using a replacement node.</td></tr>
<tr><td colspan="2">SocioEconomicGP</td></tr>
<tr><td>Variable Consolidation</td><td>

<pre>
IF SocioEconomicGP EQUALS UNSKILLED (V)
THEN
  NODE   :   2
  N   :   14
  YES   :   1.9%
  NO   :   98.1%

IF SocioEconomicGP IS ONE OF: N/A SKILLED - MANUAL (IIIM)
   MANAGERIAL AND TECHNICAL (II) SKILLED - NON MANUAL (IIIN)
   PROFESSIONAL (I) PARTLY SKILLED (IV)
THEN
  NODE   :   3
  N   :   241
  YES   :   8.4%
  NO   :   91.6%
</pre></td></tr>
<tr><td>Discussion</td><td>The consolidation process identified two value groups these will be renamed as follows: (1) Unskilled; (2) Skilled(M,NM,P),Unknown &amp; MTP.</td></tr>
<tr><td>Inclusion in Final Model</td><td>Yes</td></tr>
<tr><td>Modify</td><td>Yes - the variables will be grouped using a decision tree and it will be renamed to SocioGP, the two values will then be renamed as above.</td></tr>
</table>

Ultimately, this resulted in the compilation of the following table below.

Variable	Rejected	Original Values	New Grouped Values	Missing Values	Replacement Missing Value
PGStudies	NO	YES	N/A	NO	N/A
CourseGP	NO	NO	(1) COURSEGP1 (2) COURSEGP2 (3) COURSEGP3	YES	MISSING
EntryAge	NO	YES	N/A	YES	MISSING
EntryPoints	NO	YES	N/A	YES	MISSING
SocioGP	NO	NO	(1) UNSKILLED (2) SKILLED(M,NM,P), UNKNOWN & MTP	NO	N/A
QAHE	NO	YES	N/A	YES	0
AwardMark	NO	YES	N/A	YES	MISSING
QYPR	NO	YES	N/A	YES	0
NationalityGP	NO	NO	(1) CHINA (2) UK	YES	MISSING
Gender	YES	N/A	N/A	N/A	N/A
Ethnicity	YES	N/A	N/A	N/A	N/A
AwardClass	YES	N/A	N/A	N/A	N/A
Entry Qualifications	YES	N/A	N/A	N/A	N/A
Faculty	YES	N/A	N/A	N/A	N/A
JACS_Subject	YES	N/A	N/A	N/A	N/A
Disability	YES	N/A	N/A	N/A	N/A
HomePostcode	YES	N/A	N/A	N/A	N/A
LEA	YES	N/A	N/A	N/A	N/A

Figure 7.13 - Final Data Values Used for Modelling Postgraduate Studies.

The table above, figure 7.13, is split into two areas. The green area reflects those variables that were identified through data understanding and sampling process as being significant in predicting postgraduate studies, the red area relates to those variables that will not be included at the modelling stage. The table also highlights replacement values for grouping the data and dealing with missing values, which will be rectified in SAS® Enterprise Miner.

7.1.2.4 PSM Model

The modifications outlined above were implemented and the new grouped values were replaced as per figure 7.13. Given that the target variable is defined a number of supervised DM techniques (see sub-section 4.3.2) were applied. The selection of DM techniques, applied at the modelling stage, were based on the experiences of previous EDM studies (Superby *et al.* 2006, Romero *et al.* 2008, Dekker *et al.* 2009). Therefore three decision tree models (Entropy, Gini and ChiSquare) and three logistic regression models (using backwards, forwards and stepwise selection methods) were built. Figure 7.14, below shows the process up to and including the supervised DM techniques.

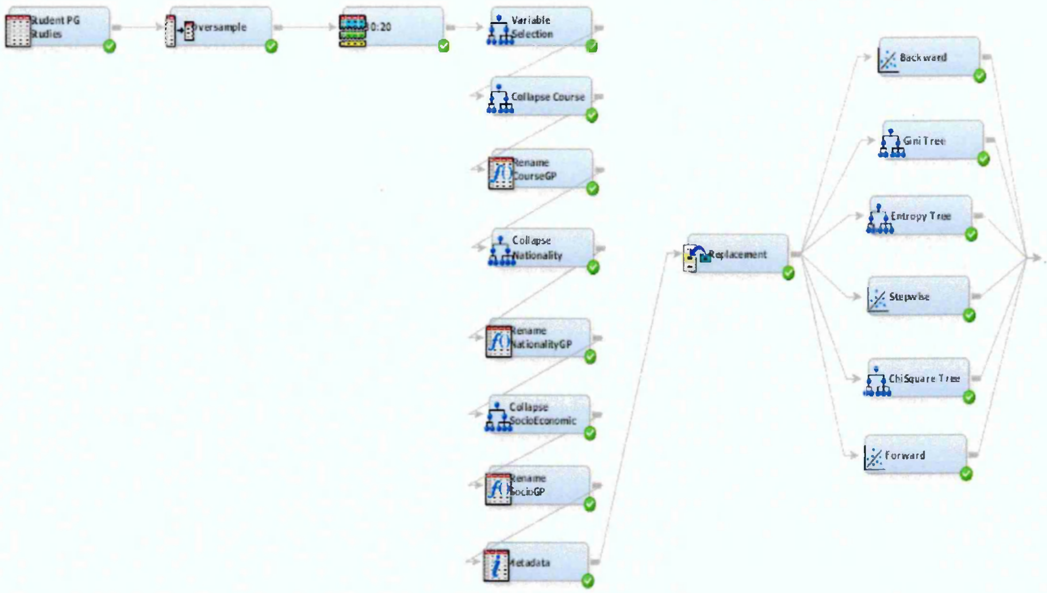
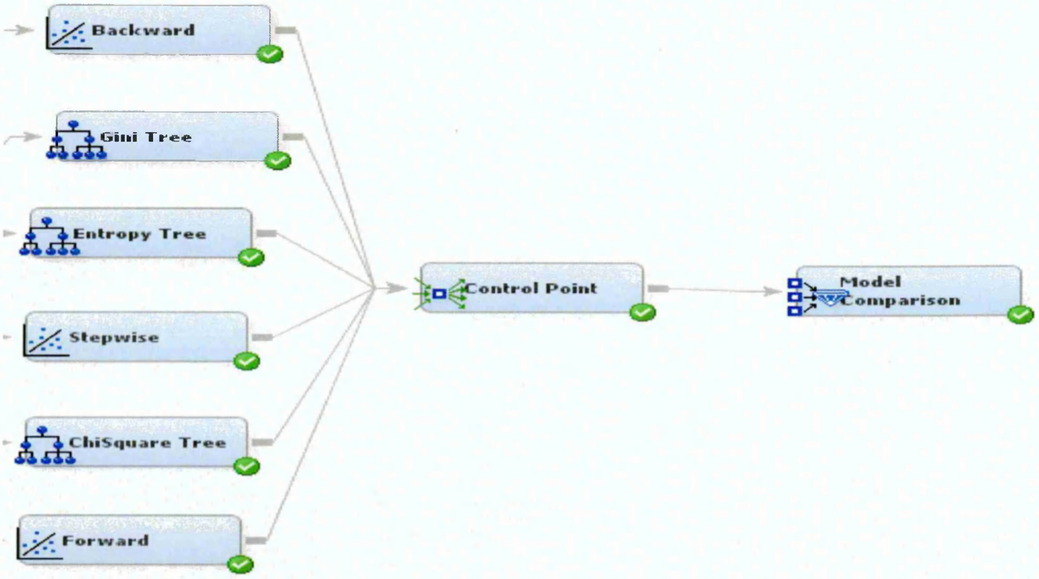
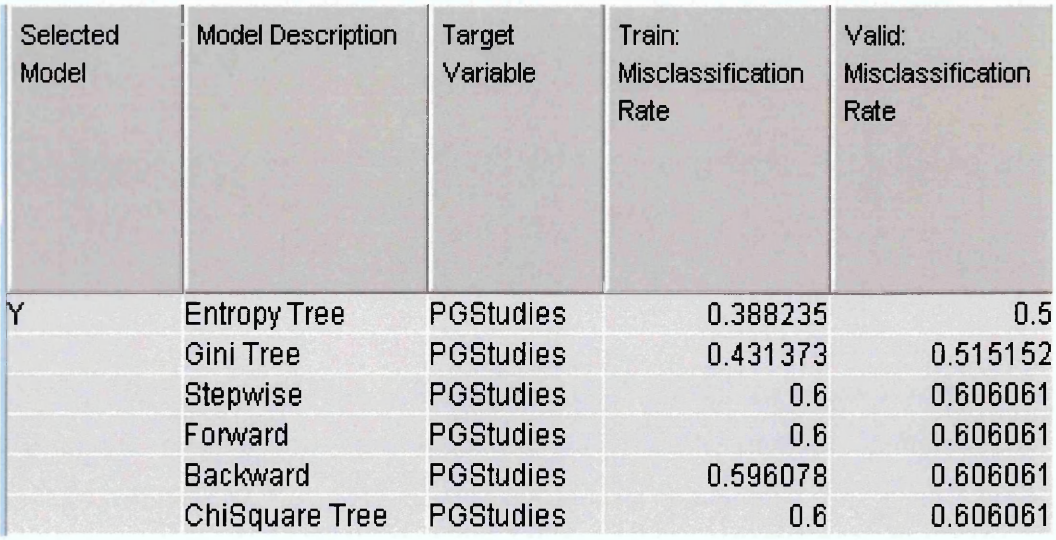


Figure 7.14 - SAS Process Flow.

7.1.2.5 PSM Assess

The modelling of the data went through numerous adjustments to determine the best model, which involved removing variables, changing sample sizes and model settings. Arguably, including details of all these different tests would undoubtedly affect the readability of this section. Therefore, this section presents the assessment of the best model. A control point and a model comparison node were then added to assess each model, see below.

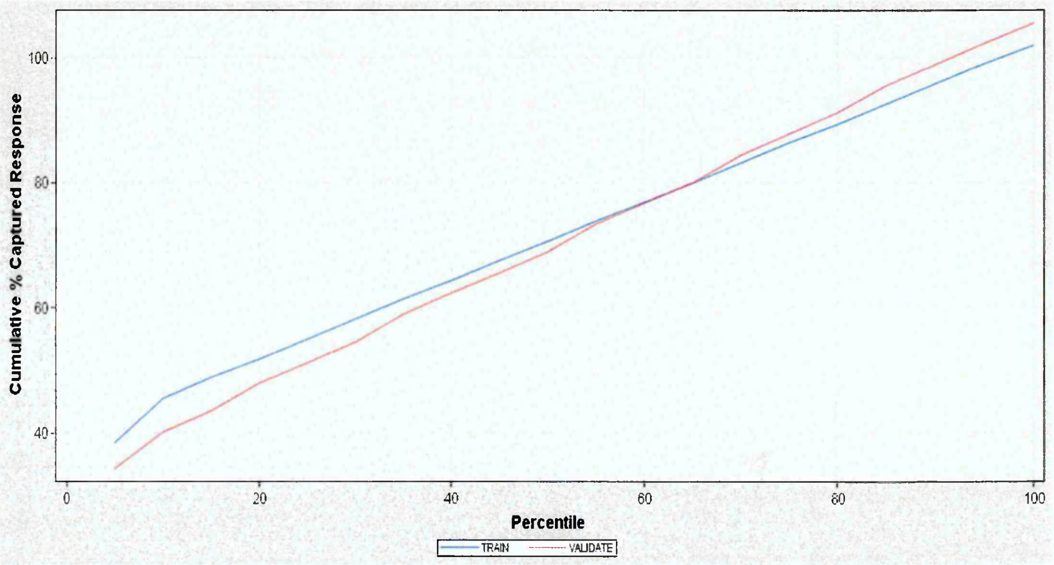




Selected Model	Model Description	Target Variable	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Entropy Tree	PGStudies	0.388235	0.5
	Gini Tree	PGStudies	0.431373	0.515152
	Stepwise	PGStudies	0.6	0.606061
	Forward	PGStudies	0.6	0.606061
	Backward	PGStudies	0.596078	0.606061
	ChiSquare Tree	PGStudies	0.6	0.606061

Figure 7.15 – Model Comparison Node.

Whilst the model comparison node generates numerous results (above), the important results are the train and valid misclassification rates as these are measures of the amount of prediction error in the model. The model comparison node suggests that the Entropy Tree produces the best model. An initial assessment of the misclassification rate (train and valid) highlights a large difference between the two misclassification rates (approximately 38% and 50%), this is due in part to the small amount of overall data (322 observations) and the 80:20 split of the data. Ultimately it is questionable whether this model would be a good predictor of postgraduate studies. Although, as noted previously, high levels of inaccuracy are common in such research studies that have tried to predict student behaviour (Superby *et al.* 2006, Romero *et al.* 2008, Herzog 2006).



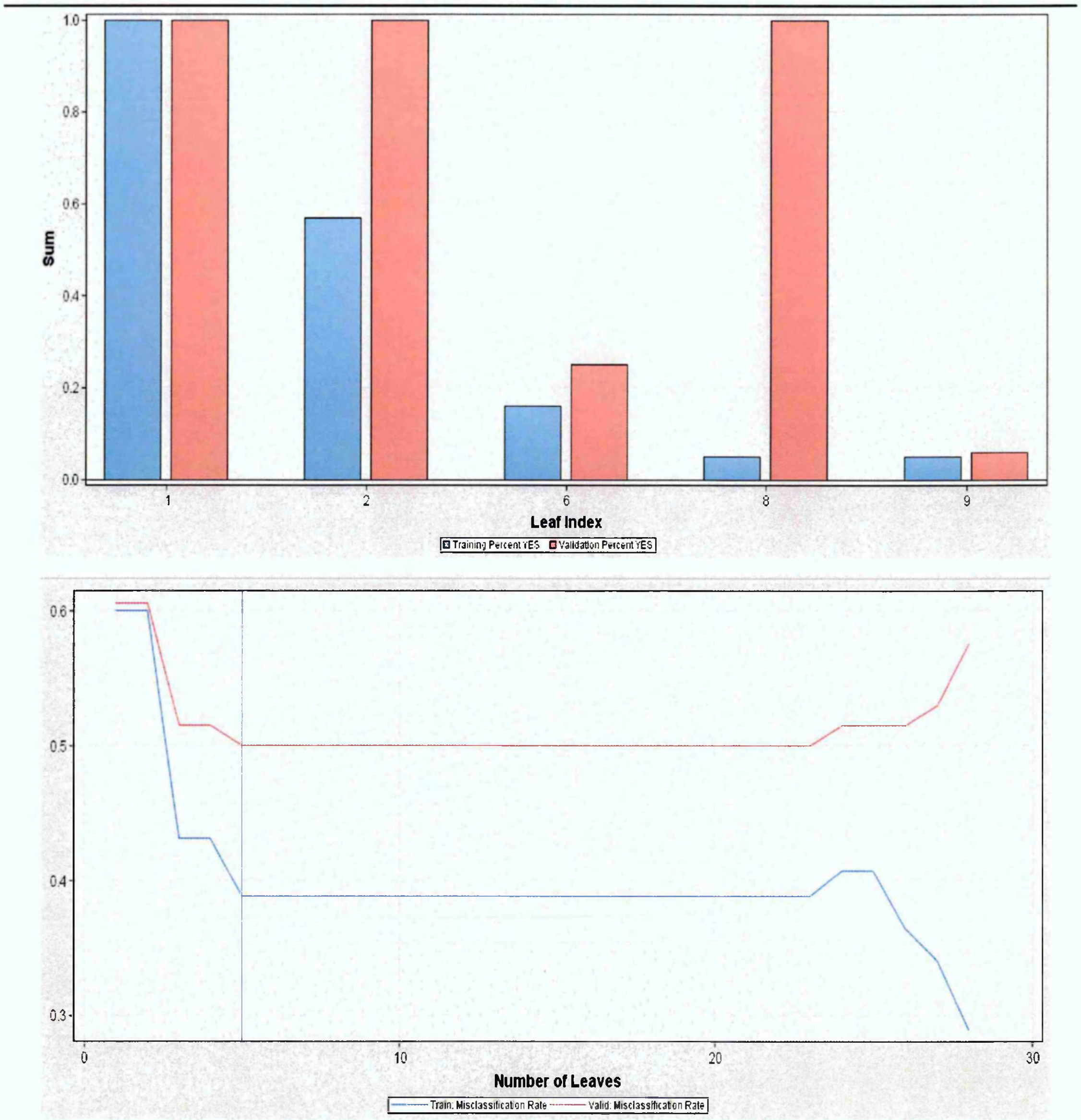
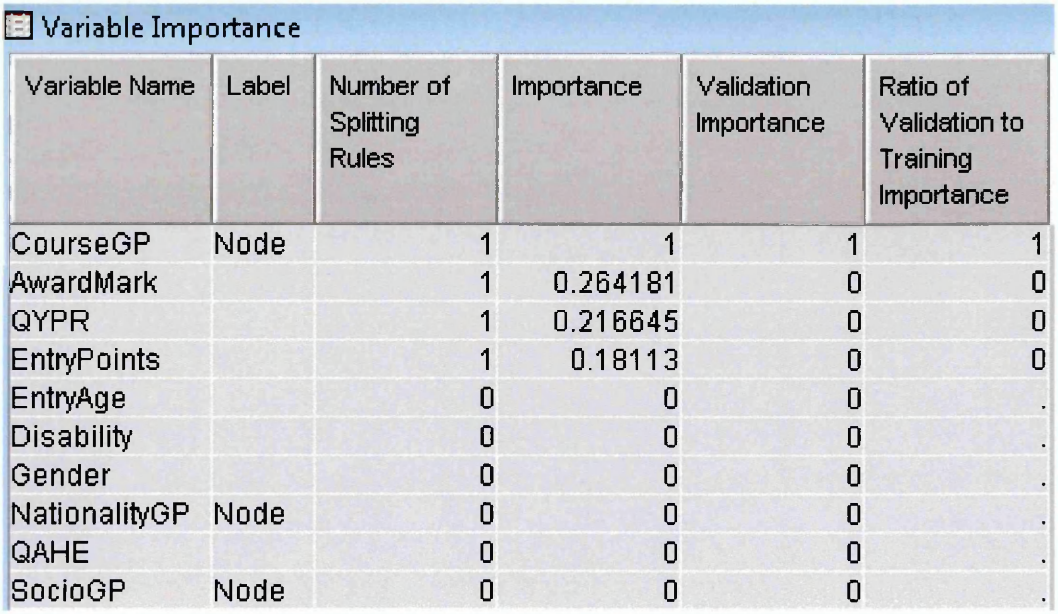


Figure 7.16 – Model Assessment.

The %captured response chart (top), shows that the top 20% of responses captures over 50% of the progression onto postgraduate studies data for the validation data set and over 55% for the train data set. Again whilst these translate into reasonably high lift (see glossary page ix) rates, there is a large amount of variation between the train and validation lift rates. The leaf index bar chart (middle) shows the amount of training and validation data in each leaf of the final model. Unlike the previous model the leaf index chart highlights that all leafs have some training and validation data but, the proportion of this is significantly different in some of the leafs. The differences in the misclassification rates can be seen in the misclassification chart (bottom), this also highlights that a tree with five leafs is the optimal solution. Arguably, this could be improved by increasing the amount of training and validation data but a larger data set wasn't available. Henceforth, the quality of the model can be judged by looking at the variables selected by the tree (below) in conjunction with previous research carried out into student retention and EDM.



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
CourseGP	Node	1	1	1	1
AwardMark		1	0.264181	0	0
QYPR		1	0.216645	0	0
EntryPoints		1	0.18113	0	0
EntryAge		0	0	0	.
Disability		0	0	0	.
Gender		0	0	0	.
NationalityGP	Node	0	0	0	.
QAHE		0	0	0	.
SocioGP	Node	0	0	0	.

Figure 7.17 – Variable Importance Table.

The variables selected by the Entropy tree also seem reasonable given the target variable of postgraduate studies. Indeed, the student's previous academic experience (in terms of entry points), the undergraduate course studied and the final award mark are plausible given the target, these variables are also pertinent in other EDM studies (Dekker *et al.* 2009, p43, Superby *et al.* 2006). The inclusion of QYPR in the final model also seems practical as this is a measure of young people's participation in HE by postcode.

7.1.2.6 The Final PSM

Figure 7.18, below, provides an overview of the final Entropy tree model.

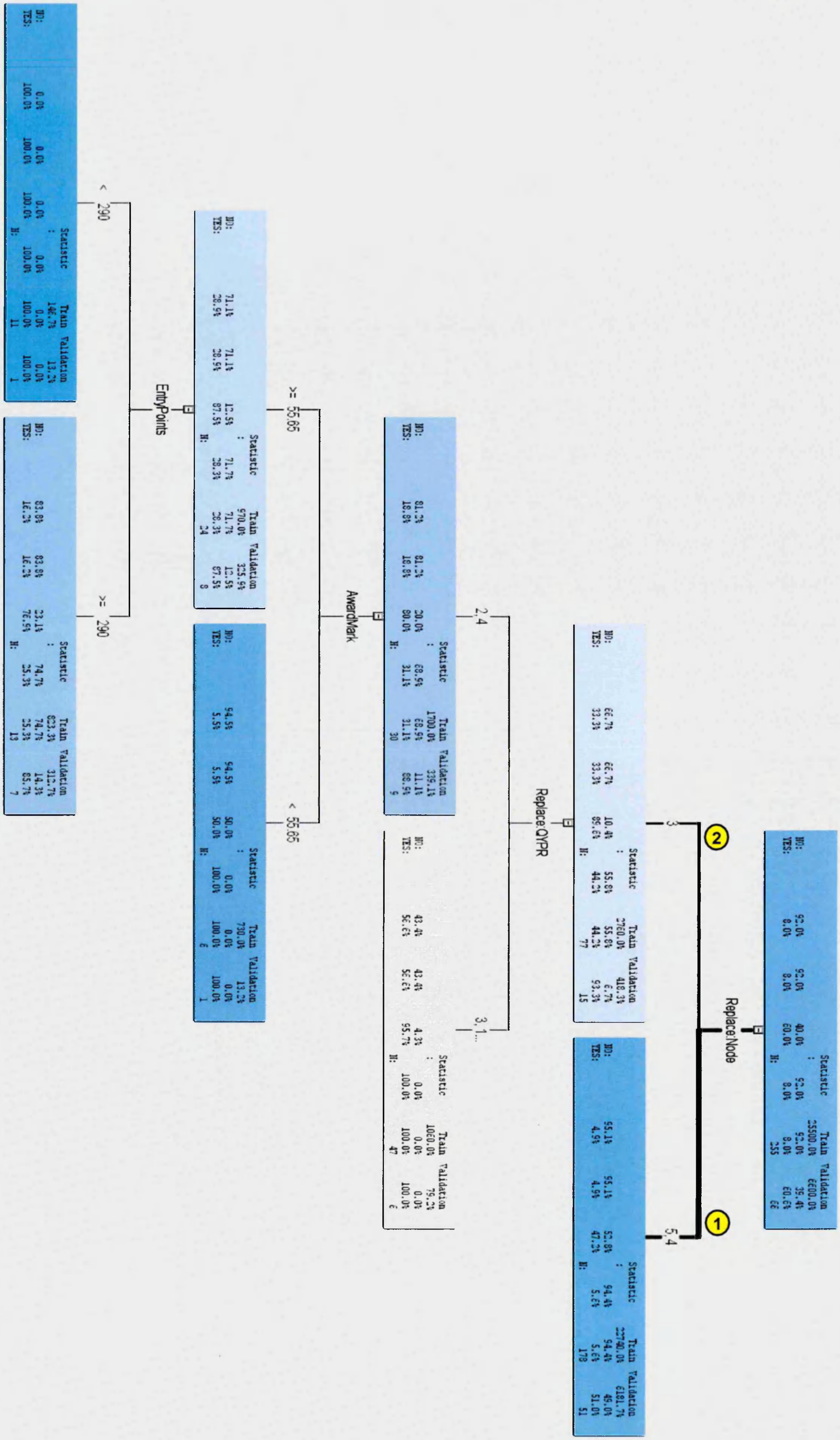
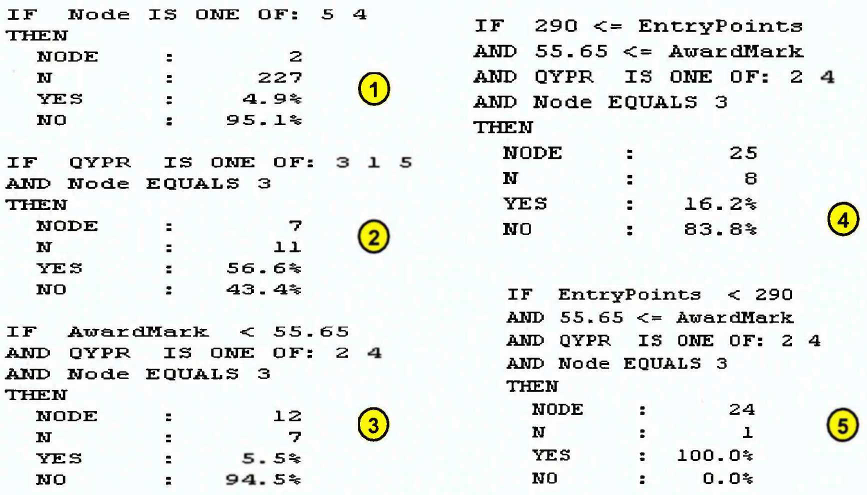


Figure 7.18 - Final Entropy Tree.

The tree initiates by splitting the CourseGP node into two groups. Following the left side of the tree (2), the data is then split by QYPR then AwardMark and EntryPoints. Again these variables are sensible given the target variable (progression onto postgraduate studies), this tree produces the following rules.

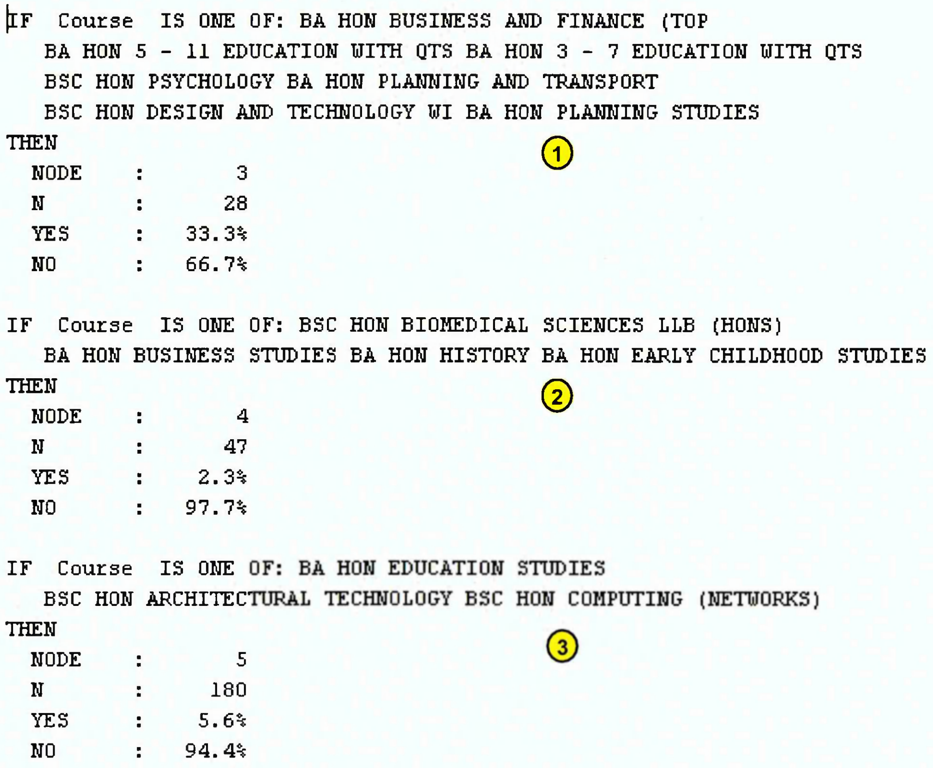


```
IF   Node IS ONE OF:  5  4
THEN
   NODE      :          2
   N         :        227
   YES       :       4.9%
   NO        :      95.1%

IF   QYPR   IS ONE OF:  3  1  5
AND Node EQUALS  3
THEN
   NODE      :          7
   N         :         11
   YES       :      56.6%
   NO        :      43.4%

IF   AwardMark   <  55.65
AND QYPR   IS ONE OF:  2  4
AND Node EQUALS  3
THEN
   NODE      :         12
   N         :          7
   YES       :       5.5%
   NO        :      94.5%
```

```
IF   290 <= EntryPoints
AND 55.65 <= AwardMark
AND QYPR   IS ONE OF:  2  4
AND Node EQUALS  3
THEN
   NODE      :        25
   N         :         8
   YES       :     16.2%
   NO        :     83.8%

IF  EntryPoints   < 290
AND 55.65 <= AwardMark
AND QYPR   IS ONE OF:  2  4
AND Node EQUALS  3
THEN
   NODE      :        24
   N         :         1
   YES       :    100.0%
   NO        :      0.0%
```

Figure 7.19 - Final Rules.

Again to aid in the understanding of these rules, the course groupings have been repeated below from the modify section.

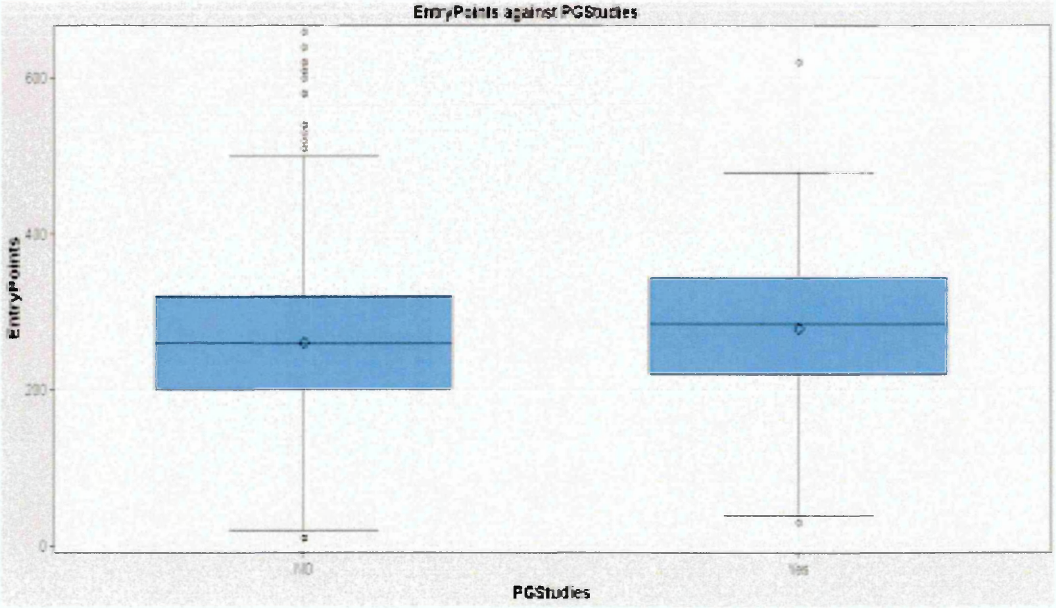


```
IF  Course  IS ONE OF: BA HON BUSINESS AND FINANCE (TOP
   BA HON 5 - 11 EDUCATION WITH QTS BA HON 3 - 7 EDUCATION WITH QTS
   BSC HON PSYCHOLOGY BA HON PLANNING AND TRANSPORT
   BSC HON DESIGN AND TECHNOLOGY WI BA HON PLANNING STUDIES
THEN
  NODE     :        3
  N        :       28
  YES      :    33.3%
  NO       :    66.7%

IF  Course  IS ONE OF: BSC HON BIOMEDICAL SCIENCES LLB (HONS)
   BA HON BUSINESS STUDIES BA HON HISTORY BA HON EARLY CHILDHOOD STUDIES
THEN
  NODE     :        4
  N        :       47
  YES      :     2.3%
  NO       :    97.7%

IF  Course  IS ONE OF: BA HON EDUCATION STUDIES
   BSC HON ARCHITECTURAL TECHNOLOGY BSC HON COMPUTING (NETWORKS)
THEN
  NODE     :        5
  N        :      180
  YES      :     5.6%
  NO       :    94.4%
```

These rules can be interpreted as follows:

Rule (1)	If CourseGP is either CourseGP2 or CourseGP3 then students or unlikely to progress onto postgraduate studies.
Rule (2)	If QYPR is Medium, Low or High and CourseGP is CourseGP1 then students are more likely to go onto postgraduate studies.
Rule (3)	If AwardMark is less than 55.65 and QYPR is one of Low-Medium or Medium-High and CourseGP is CourseGP1 then students are less likely to go onto postgraduate studies.
Rule (4)	If EntryPoints are greater than or equal to 290 and AwardMark is greater than 55.65 and QYPR is Low-Medium or Medium-High and CourseGP is CourseGP1 then student are less likely to progress onto postgraduate studies.
Rule (5)	If EntryPoints are less than or equal to 290 and AwardMark is greater than 55.65 and QYPR is Low-Medium or Medium-High and CourseGP is CourseGP1 then student are more likely to progress onto postgraduate studies.

These rules are sensible given the exploration of the data, carried out previously. However, the models ability to successfully predict progression onto postgraduate studies is questionable due to the small amount of data used to build the model. An assessment of the exploration graphs (below), for rule 5, confirms this doubt.



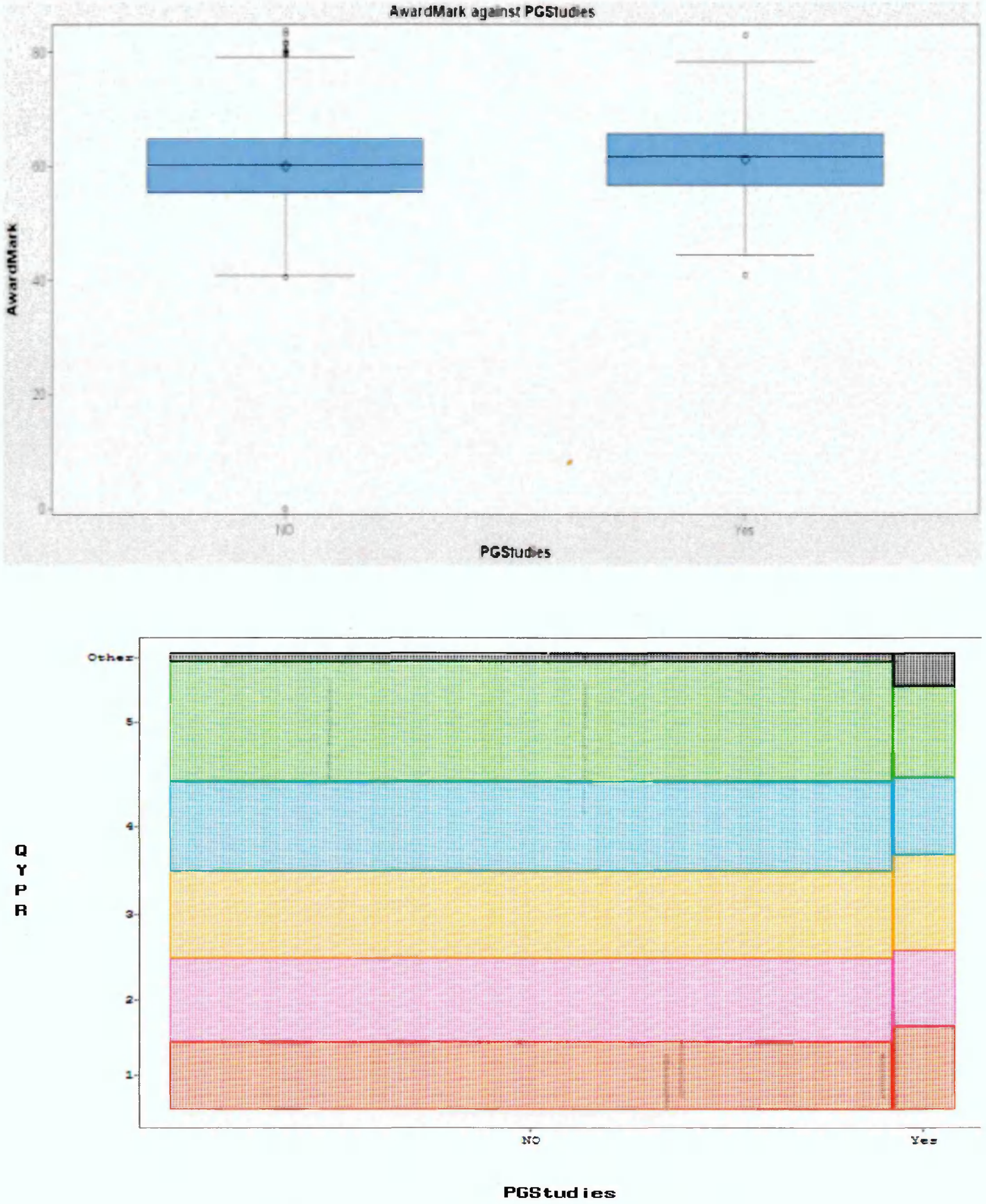


Figure 7.20 – Assessing Rule 5 Plots.

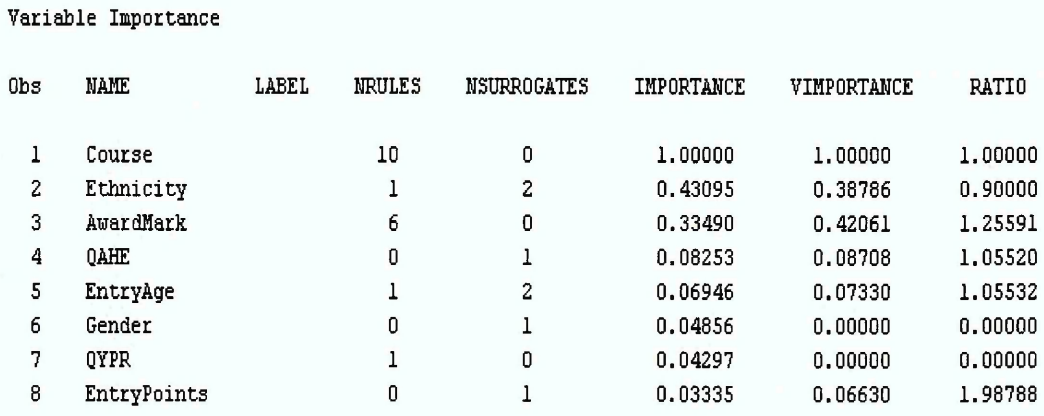
Indeed, the first box plot of EntryPoint against PGstudies indicates that the majority of students who go on to take postgraduate studies enter their undergraduate degree with around 300 points, the final model suggest less than or equal to 290. The middle box plot of AwardMark also shows that majority of students, progressing onto postgraduate studies obtain an average award mark of above 60, the final model suggests greater than 55.65. The mosaic plot of QYPR also highlights that more students in QYPR 3 progressed onto postgraduate studies. The discrepancies between the graphs and the final rules are due to the oversampling of the data and help to explain the large variation between the train (38%) and valid (50%) misclassification rates.

7.1.3 Building the Employment Model (EM)

This section details the process followed in building the employment DM model using SAS® SEMMA.

7.1.3.1 EM Sample

The employment DM mart contains 1749 recorded employment types, the remaining are not recorded/unknown, which is a 45% event rate. The recorded values breakdown as follows: 23.14% of student obtained a graduate job, 12.01% non-graduate jobs, 2.66% other, 3.55% study, 3.23% unemployed and 55.41% not recorded/unknown. The not recorded/unknown values were left in the data set as this highlighted some interesting relationships between certain variables such as ethnicity. A 70:30 data partition (provided the best results) was setup so that the data could be divided into training and validation respectively. Again the variables within the data set were reduced through the authors own knowledge of the data/HE and by using a decision tree for variable selection. This resulted in the following inputs being selected for modelling:



```
Variable Importance

Obs   NAME          LABEL   NRULES   NSURROGATES   IMPORTANCE   VIMPORTANCE    RATIO

 1    Course                  10         0           1.00000      1.00000     1.00000
 2    Ethnicity                1         2           0.43095      0.38786     0.90000
 3    AwardMark                6         0           0.33490      0.42061     1.25591
 4    QAHE                     0         1           0.08253      0.08708     1.05520
 5    EntryAge                 1         2           0.06946      0.07330     1.05532
 6    Gender                   0         1           0.04856      0.00000     0.00000
 7    QYPR                     1         0           0.04297      0.00000     0.00000
 8    EntryPoints              0         1           0.03335      0.06630     1.98788
```

Figure 7.21 – Variable Importance Table.

7.1.3.2 EM Explore

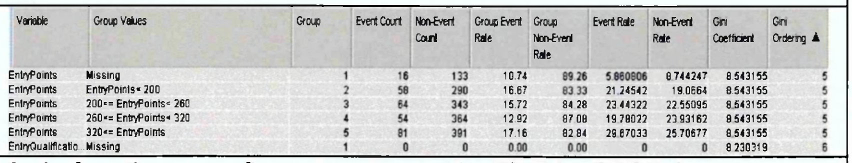
The data has already been thoroughly explored through the data understanding process outlined in Chapter 6, through this a number of graphs were produced, outliers were identified.

7.1.3.3 EM Modify

This section will focus on using categorical variable consolidation (see glossary page vii) for categorical values, interactive binning/grouping (for interval values) and the authors own knowledge of the data/HE to determine suitable groups of values in relation to the target variable.

ENTRYPOINTS

INTERACTIVE BINNING/ GROUPING RESULTS

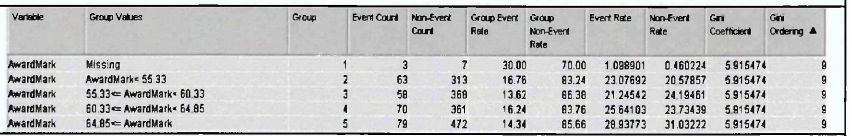


Variable	Group Values	Group	Event Count	Non-Event Count	Group Event Rate	Group Non-Event Rate	Event Rate	Non-Event Rate	Gini Coefficient	Gini Ordering ▲
EntryPoints	Missing	1	16	133	10.74	89.26	5.860806	8.744247	8.543155	5
EntryPoints	EntryPoints< 200	2	58	290	16.67	83.33	21.24542	19.0664	8.543155	5
EntryPoints	200<= EntryPoints< 260	3	64	343	15.72	84.28	23.44322	22.55095	8.543155	5
EntryPoints	260<= EntryPoints< 320	4	54	364	12.92	87.08	19.78022	23.93162	8.543155	5
EntryPoints	320<= EntryPoints	5	81	391	17.16	82.84	29.67033	25.70677	8.543155	5
EntryQualificatio...	Missing	1	0	0	0.00	0.00	0	0	8.230319	6

DISCUSSION	As before the same five groups were identified by SAS® for EntryPoints.
INCLUSION IN FINAL MODEL	Yes
MODIFY	No – Data Mining model will determine the best split

AWARDMARK

INTERACTIVE BINNING/ GROUPING RESULTS

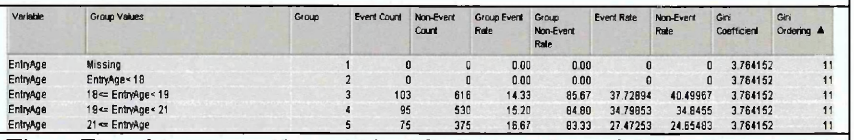


Variable	Group Values	Group	Event Count	Non-Event Count	Group Event Rate	Group Non-Event Rate	Event Rate	Non-Event Rate	Gini Coefficient	Gini Ordering ▲
AwardMark	Missing	1	3	7	30.00	70.00	1.098901	0.460224	5.915474	9
AwardMark	AwardMark< 55.33	2	63	313	16.76	83.24	23.07692	20.57857	5.915474	9
AwardMark	55.33<= AwardMark< 60.33	3	58	368	13.62	86.38	21.24542	24.19461	5.915474	9
AwardMark	60.33<= AwardMark< 64.85	4	70	361	16.24	83.76	25.64103	23.73439	5.915474	9
AwardMark	64.85<= AwardMark	5	79	472	14.34	85.66	28.93773	31.03222	5.915474	9

DISCUSSION	SAS® suggests that AwardMark could be grouped into 5 groups
INCLUSION IN FINAL MODEL	Yes
MODIFY	No – Data Mining model will determine the best split

ENTRYAGE

INTERACTIVE BINNING/ GROUPING RESULTS



Variable	Group Values	Group	Event Count	Non-Event Count	Group Event Rate	Group Non-Event Rate	Event Rate	Non-Event Rate	Gini Coefficient	Gini Ordering ▲
EntryAge	Missing	1	0	0	0.00	0.00	0	0	3.764152	11
EntryAge	EntryAge< 18	2	0	0	0.00	0.00	0	0	3.764152	11
EntryAge	18<= EntryAge< 19	3	103	616	14.33	85.67	37.72894	40.49967	3.764152	11
EntryAge	19<= EntryAge< 21	4	95	530	15.20	84.80	34.79853	34.8455	3.764152	11
EntryAge	21<= EntryAge	5	75	375	16.67	83.33	27.47253	24.65483	3.764152	11

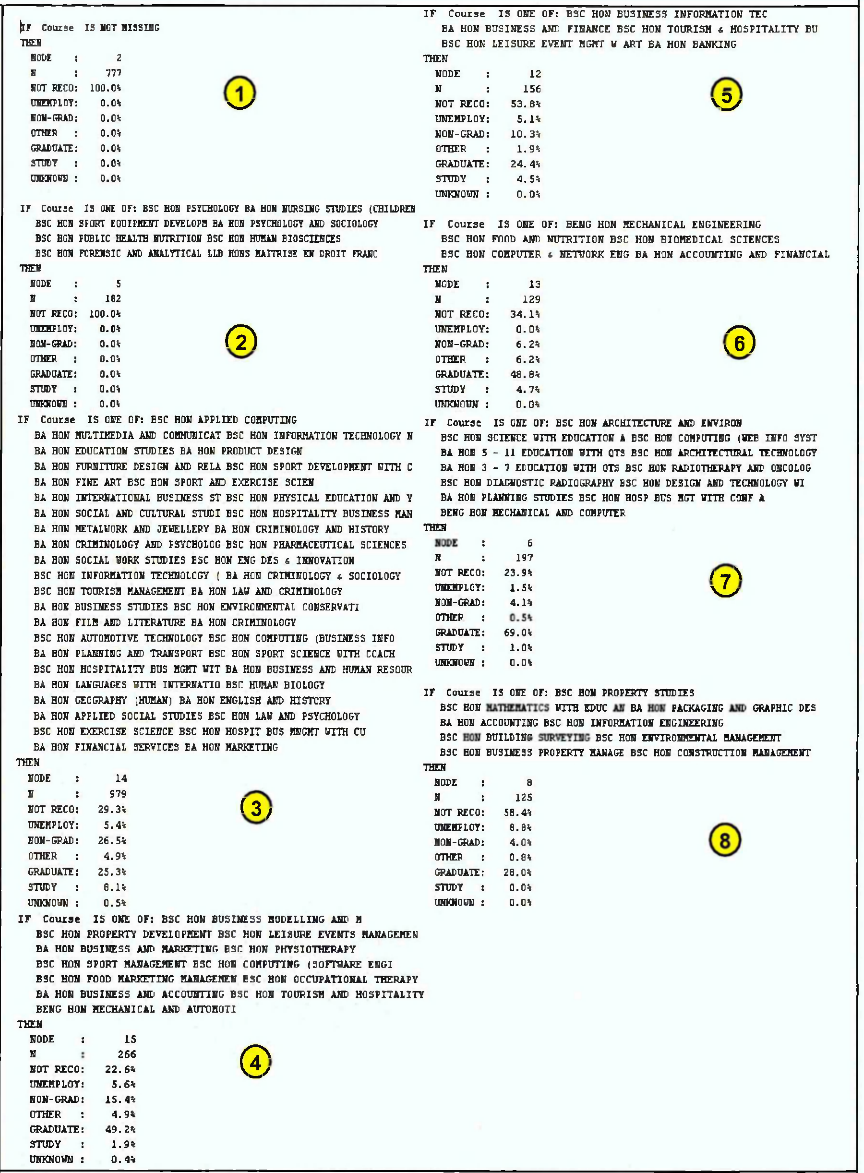
DISCUSSION	The EntryAge groupings identified by the SAS® Interactive Binning/Grouping node is the same as the groups identified when the target variable was AWClassDUMMY and PGStudies.
INCLUSION IN FINAL MODEL	Yes
MODIFY	No – Data Mining model will determine the best split

ETHNICITY

VARIABLE CONSOLIDATION

```
IF  Ethnicity  IS ONE OF: CHINESE BLACK CARIBBEAN INFO REFUSED ASIAN OTHER
    PAKISTANI NOT KNOWN INDIAN OTHER WHITE WHITE IRISH WH AND BL CARIB
THEN
  NODE      :       2
  N         :     381
  NOT RECO:   71.1%
  UNEMPLOY:    4.7%                    1
  NON-GRAD:    9.2%
  OTHER    :    1.0%
  GRADUATE:    9.7%
  STUDY    :    3.9%
  UNKNOWN  :    0.3%

IF  Ethnicity  IS ONE OF: WHITE BRITISH BANGLADESHI OTHER MIXED
    BLACK AFRICAN OTHER WH AND ASIAN WHITE WELSH WHITE SCOTTISH
THEN
  NODE      :       3
  N         :    2430
  NOT RECO:   52.8%
  UNEMPLOY:    3.0%                    2
  NON-GRAD:   12.4%
  OTHER    :    2.9%
  GRADUATE:   25.3%
  STUDY    :    3.5%
  UNKNOWN  :    0.2%
```

DISCUSSION	The results of the variable consolidation identified two groups, which will be renamed as follows: (1)Mixed Group; (2) Mainly White;
INCLUSION IN FINAL MODEL	Yes
MODIFY	Yes - the variables will be grouped using a decision tree and it will be renamed to EthnicityGP, the four values will then be renamed as above.

COURSE	
VARIABLE CONSOLIDATION	

```
IF  Course  IS NOT MISSING
THEN
  NODE     :       2
  N        :     777
  NOT RECO:  100.0%
  UNEMPLOY:    0.0%
  NON-GRAD:    0.0%
  OTHER    :   0.0%
  GRADUATE:    0.0%
  STUDY    :   0.0%
  UNKNOWN :    0.0%

IF  Course  IS ONE OF: BSC HON PSYCHOLOGY BA HON NURSING STUDIES (CHILDREN
   BSC HON SPORT EQUIPMENT DEVELOPM BA HON PSYCHOLOGY AND SOCIOLOGY
   BSC HON PUBLIC HEALTH NUTRITION BSC HON HUMAN BIOSCIENCES
   BSC HON FORENSIC AND ANALYTICAL LLB HONS MAITRISE EN DROIT FRANC
THEN
  NODE     :       5
  N        :     182
  NOT RECO:  100.0%
  UNEMPLOY:    0.0%
  NON-GRAD:    0.0%
  OTHER    :   0.0%
  GRADUATE:    0.0%
  STUDY    :   0.0%
  UNKNOWN :    0.0%
IF  Course  IS ONE OF: BSC HON APPLIED COMPUTING
   BA HON MULTIMEDIA AND COMMUNICAT BSC HON INFORMATION TECHNOLOGY N
   BA HON EDUCATION STUDIES BA HON PRODUCT DESIGN
   BA HON FURNITURE DESIGN AND RELA BSC HON SPORT DEVELOPMENT WITH C
   BA HON FINE ART BSC HON SPORT AND EXERCISE SCIEN
   BA HON INTERNATIONAL BUSINESS ST BSC HON PHYSICAL EDUCATION AND Y
   BA HON SOCIAL AND CULTURAL STUDI BSC HON HOSPITALITY BUSINESS MAN
   BA HON METALWORK AND JEWELLERY BA HON CRIMINOLOGY AND HISTORY
   BA HON CRIMINOLOGY AND PSYCHOLOG BSC HON PHARMACEUTICAL SCIENCES
   BA HON SOCIAL WORK STUDIES BSC HON ENG DES & INNOVATION
   BSC HON INFORMATION TECHNOLOGY ( BA HON CRIMINOLOGY & SOCIOLOGY
   BSC HON TOURISM MANAGEMENT BA HON LAW AND CRIMINOLOGY
   BA HON BUSINESS STUDIES BSC HON ENVIRONMENTAL CONSERVATI
   BA HON FILM AND LITERATURE BA HON CRIMINOLOGY
   BSC HON AUTOMOTIVE TECHNOLOGY BSC HON COMPUTING (BUSINESS INFO
   BA HON PLANNING AND TRANSPORT BSC HON SPORT SCIENCE WITH COACH
   BSC HON HOSPITALITY BUS MGMT WIT BA HON BUSINESS AND HUMAN RESOUR
   BA HON LANGUAGES WITH INTERNATIO BSC HUMAN BIOLOGY
   BA HON GEOGRAPHY (HUMAN) BA HON ENGLISH AND HISTORY
   BA HON APPLIED SOCIAL STUDIES BSC HON LAW AND PSYCHOLOGY
   BSC HON EXERCISE SCIENCE BSC HON HOSPIT BUS MNGMT WITH CU
   BA HON FINANCIAL SERVICES BA HON MARKETING
THEN
  NODE     :      14
  N        :     979
  NOT RECO:   29.3%
  UNEMPLOY:    5.4%
  NON-GRAD:   26.5%
  OTHER    :   4.9%
  GRADUATE:   25.3%
  STUDY    :   8.1%
  UNKNOWN :    0.5%
IF  Course  IS ONE OF: BSC HON BUSINESS MODELLING AND M
   BSC HON PROPERTY DEVELOPMENT BSC HON LEISURE EVENTS MANAGEMEN
   BA HON BUSINESS AND MARKETING BSC HON PHYSIOTHERAPY
   BSC HON SPORT MANAGEMENT BSC HON COMPUTING (SOFTWARE ENGI
   BSC HON FOOD MARKETING MANAGEMEN BSC HON OCCUPATIONAL THERAPY
   BA HON BUSINESS AND ACCOUNTING BSC HON TOURISM AND HOSPITALITY
   BENG HON MECHANICAL AND AUTOMOTI
THEN
  NODE     :      15
  N        :     266
  NOT RECO:   22.6%
  UNEMPLOY:    5.6%
  NON-GRAD:   15.4%
  OTHER    :   4.9%
  GRADUATE:   49.2%
  STUDY    :   1.9%
  UNKNOWN :    0.4%
```

```
IF  Course  IS ONE OF: BSC HON BUSINESS INFORMATION TEC
   BA HON BUSINESS AND FINANCE BSC HON TOURISM & HOSPITALITY BU
   BSC HON LEISURE EVENT MGMT W ART BA HON BANKING
THEN
  NODE     :      12
  N        :     156
  NOT RECO:   53.8%
  UNEMPLOY:    5.1%
  NON-GRAD:   10.3%
  OTHER    :   1.9%
  GRADUATE:   24.4%
  STUDY    :   4.5%
  UNKNOWN :    0.0%

IF  Course  IS ONE OF: BENG HON MECHANICAL ENGINEERING
   BSC HON FOOD AND NUTRITION BSC HON BIOMEDICAL SCIENCES
   BSC HON COMPUTER & NETWORK ENG BA HON ACCOUNTING AND FINANCIAL
THEN
  NODE     :      13
  N        :     129
  NOT RECO:   34.1%
  UNEMPLOY:    0.0%
  NON-GRAD:    6.2%
  OTHER    :   6.2%
  GRADUATE:   48.8%
  STUDY    :   4.7%
  UNKNOWN :    0.0%
IF  Course  IS ONE OF: BSC HON ARCHITECTURE AND ENVIRON
   BSC HON SCIENCE WITH EDUCATION A BSC HON COMPUTING (WEB INFO SYST
   BA HON 5 - 11 EDUCATION WITH QTS BSC HON ARCHITECTURAL TECHNOLOGY
   BA HON 3 - 7 EDUCATION WITH QTS BSC HON RADIOTHERAPY AND ONCOLOG
   BSC HON DIAGNOSTIC RADIOGRAPHY BSC HON DESIGN AND TECHNOLOGY WI
   BA HON PLANNING STUDIES BSC HON HOSP BUS MGT WITH CONF A
   BENG HON MECHANICAL AND COMPUTER
THEN
  NODE     :       6
  N        :     197
  NOT RECO:   23.9%
  UNEMPLOY:    1.5%
  NON-GRAD:    4.1%
  OTHER    :   0.5%
  GRADUATE:   69.0%
  STUDY    :   1.0%
  UNKNOWN :    0.0%

IF  Course  IS ONE OF: BSC HON PROPERTY STUDIES
   BSC HON MATHEMATICS WITH EDUC AN BA HON PACKAGING AND GRAPHIC DES
   BA HON ACCOUNTING BSC HON INFORMATION ENGINEERING
   BSC HON BUILDING SURVEYING BSC HON ENVIRONMENTAL MANAGEMENT
   BSC HON BUSINESS PROPERTY MANAGE BSC HON CONSTRUCTION MANAGEMENT
THEN
  NODE     :       8
  N        :     125
  NOT RECO:   58.4%
  UNEMPLOY:    8.8%
  NON-GRAD:    4.0%
  OTHER    :   0.8%
  GRADUATE:   28.0%
  STUDY    :   0.0%
  UNKNOWN :    0.0%
```

DISCUSSION	Eight groups were identified through categorical variable consolidation, these groups will be renamed as follows: (1) CourseGP1 (2) CourseGP2 (3) CourseGP3 (4) CourseGP4 (5) CourseGP5 (6) CourseGP6 (7) CourseGP7 (8) CourseGP8
INCLUSION IN FINAL MODEL	Yes
MODIFY	Yes - A Decision Tree will be used to collapse the variable into a new variable called CourseGP. The Decision Tree produces a new variable called _Node_ this will be renamed, using a Transform Variable node, to CourseGP and the eight individual values will be replaced with the values above using a replacement node.

Ultimately, this resulted in the compilation of the following table below.

Variable	Rejected	Original Values	New Grouped Values	Missing Values	Replacement Missing Value
PostUG Destination	NO	YES	N/A	NO	N/A
CourseGP	NO	NO	(1) COURSEGP1 (2) COURSEGP2 (3) COURSEGP3 (4) COURSEGP4 (5) COURSEGP5 (6) COURSEGP6 (7) COURSEGP7 (8) COURSEGP8	YES	MISSING
EthnicityGP	NO	NO	(1) MIXED GROUP (2) MAINLY WHITE	N/A	N/A
AwardMark	NO	YES	N/A	YES	MISSING
QAHE	NO	YES	N/A	YES	0
EntryAge	NO	YES	N/A	NO	MISSING
Gender	NO	YES	N/A	N/A	N/A
QYPR	NO	YES	N/A	YES	0
EntryPoints	NO	NO	N/A	YES	MISSING
JACS_Subject	YES	N/A	N/A	N/A	N/A
Socio EconomicGP	YES	N/A	N/A	N/A	N/A
Nationality	YES	N/A	N/A	N/A	N/A
AwardClass	YES	N/A	N/A	N/A	N/A
Entry Qualifications	YES	N/A	N/A	N/A	N/A
Faculty	YES	N/A	N/A	N/A	N/A
Course	YES	N/A	N/A	N/A	N/A
Disability	YES	N/A	N/A	N/A	N/A
Home Postcode	YES	N/A	N/A	N/A	N/A
LEA	YES	N/A	N/A	N/A	N/A

Figure 7.22 - Final Data Values Used for Modelling Employment Type.

The table above, figure 7.22, is split into two areas. The green area reflects those variables that were identified through data understanding and sampling process as being significant in predicting award classification, the red area relates to those variables that will not be included at the modelling stage. The table also highlights replacement values for grouping the data and dealing with missing values, which will be rectified in SAS® Enterprise Miner.

7.1.3.4 EM Model

The modifications outlined above were implemented and the new grouped values were replaced as per figure 7.22. Given that the target variable is defined, a number of supervised DM techniques (see sub-section 4.3.2) were applied. The selection of DM techniques, applied at the modelling stage, were based on the experiences of previous EDM studies (Superby *et al.* 2006, Romero *et al.* 2008, Dekker *et al.* 2009). Therefore three decision tree models (Entropy, Gini and ChiSquare) and three logistic regression models (using backwards, forwards and stepwise selection methods) were built. Figure 7.23, below shows the process up to and including the supervised DM techniques.

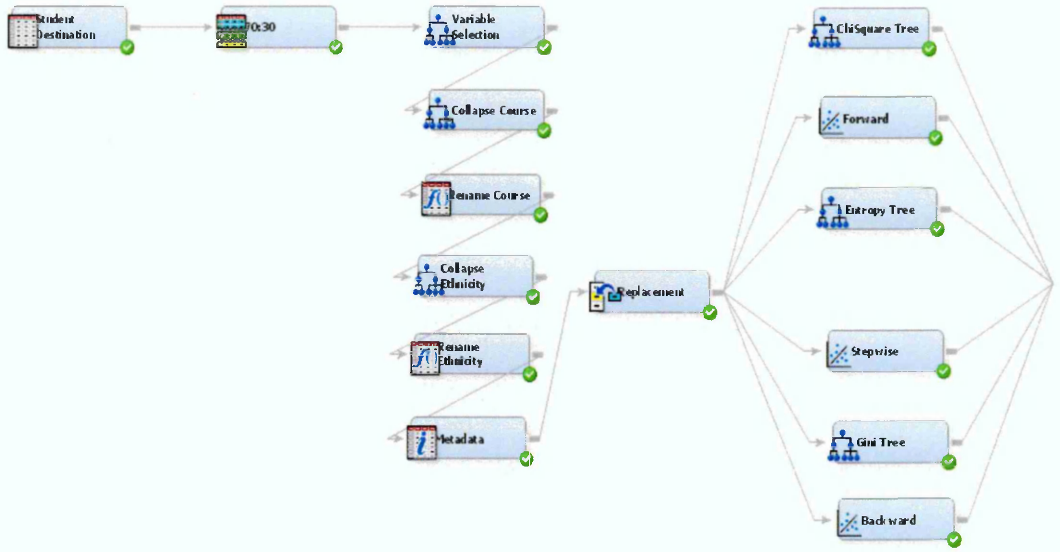
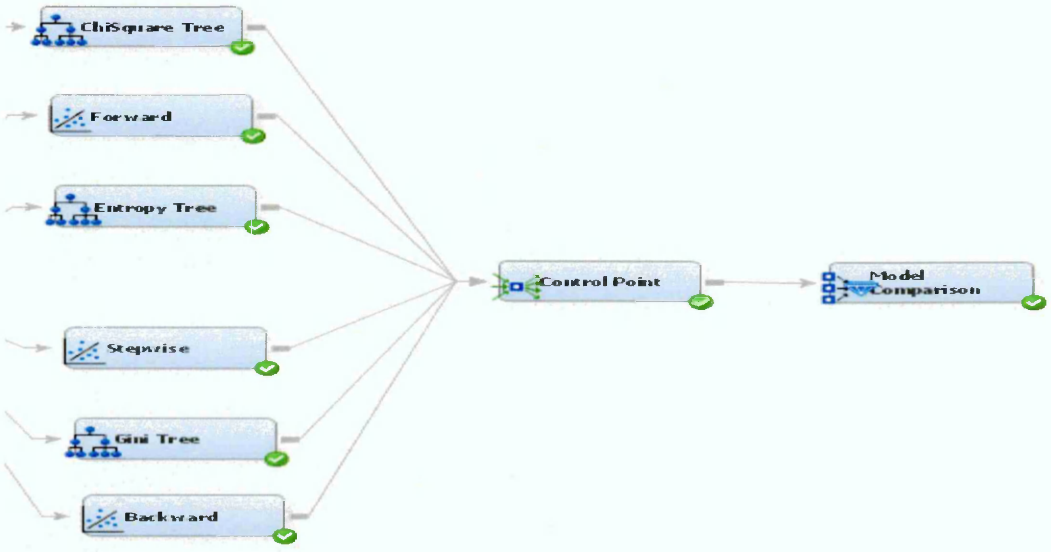
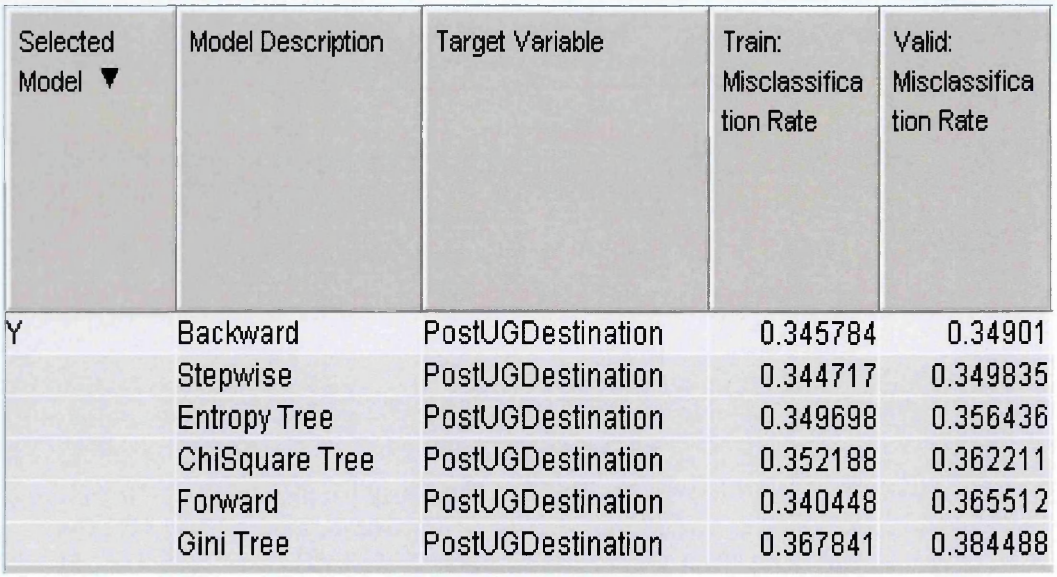


Figure 7.23 - SAS Process Flow.

7.1.3.5 EM Assess

The modelling of the data went through numerous adjustments to determine the best model, which involved removing variables, changing sample sizes and model settings. Arguably, including details of all these different tests would undoubtedly affect the readability of this section. Therefore, this section presents the assessment of the best model. A control point and a model comparison node were then added to assess each model, see below.

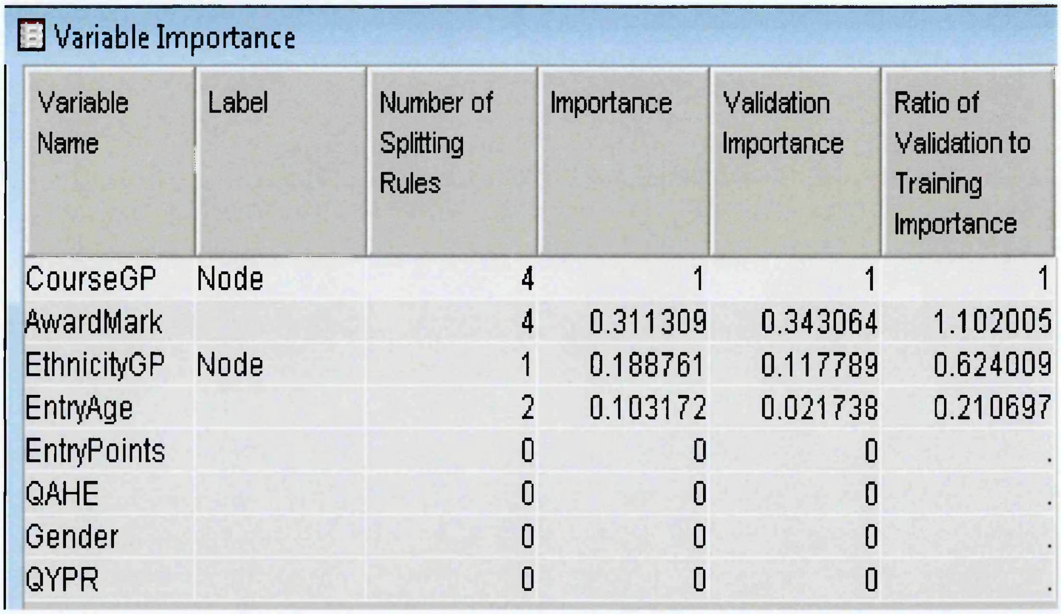




Selected Model ▼	Model Description	Target Variable	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Backward	PostUGDestination	0.345784	0.34901
	Stepwise	PostUGDestination	0.344717	0.349835
	Entropy Tree	PostUGDestination	0.349698	0.356436
	ChiSquare Tree	PostUGDestination	0.352188	0.362211
	Forward	PostUGDestination	0.340448	0.365512
	Gini Tree	PostUGDestination	0.367841	0.384488

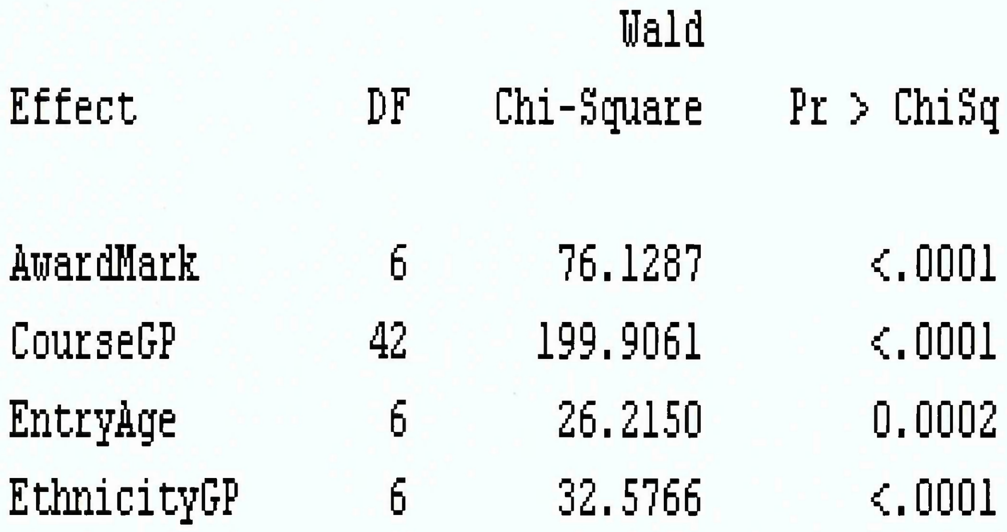
Figure 7.24 – Model Comparison Node.

The results of the modelling (above), indicates that the backward elimination model produces the best model. However, the differences in misclassification rates between the regression models and the Entropy Tree are small. In addition to this, backward elimination is well known for not creating the best model as removed variables aren't reconsidered. Therefore, the stepwise elimination would produce a better model as this uses a combination of both backward and forward elimination so that removed variables are reconsidered (Berry and Linoff 2011). However, upon comparison of the variables identified through both stepwise elimination and entropy tree (below), there is very little difference between the two models, as noted by Herzog (2006).



Variable Importance

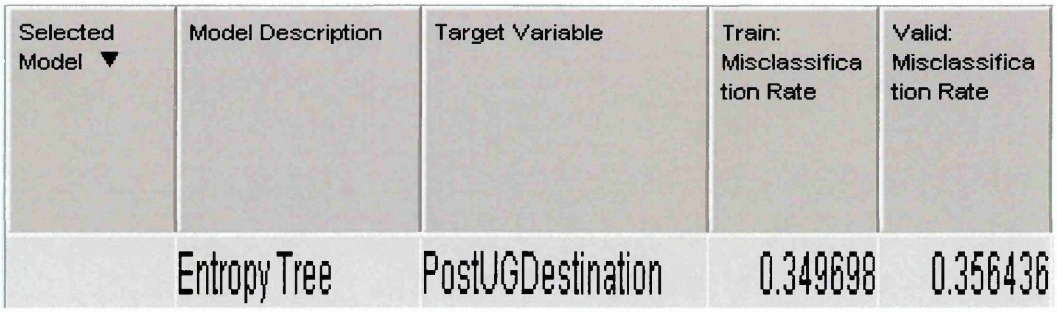
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
CourseGP	Node	4	1	1	1
AwardMark		4	0.311309	0.343064	1.102005
EthnicityGP	Node	1	0.188761	0.117789	0.624009
EntryAge		2	0.103172	0.021738	0.210697
EntryPoints		0	0	0	.
QAHE		0	0	0	.
Gender		0	0	0	.
QYPR		0	0	0	.



Effect	DF	Wald Chi-Square	Pr > ChiSq
AwardMark	6	76.1287	<.0001
CourseGP	42	199.9061	<.0001
EntryAge	6	26.2150	0.0002
EthnicityGP	6	32.5766	<.0001

Figure 7.25 – Variable Importance Tables.

Therefore, due to the similarities between the two models, it was decided to adopt the entropy tree model as the results are much easier to understand and interpret.



Selected Model ▼	Model Description	Target Variable	Train: Misclassification Rate	Valid: Misclassification Rate
	Entropy Tree	PostUGDestination	0.349698	0.356436

Figure 7.26 – Entropy Tree Misclassification Rates.

Assessing the misclassification rates, of the entropy tree (above), suggests that this model would be a good predictor of student employment. Indeed, the misclassification rates are relatively small (34% and 35%) and the difference between the two misclassification rates is also small.

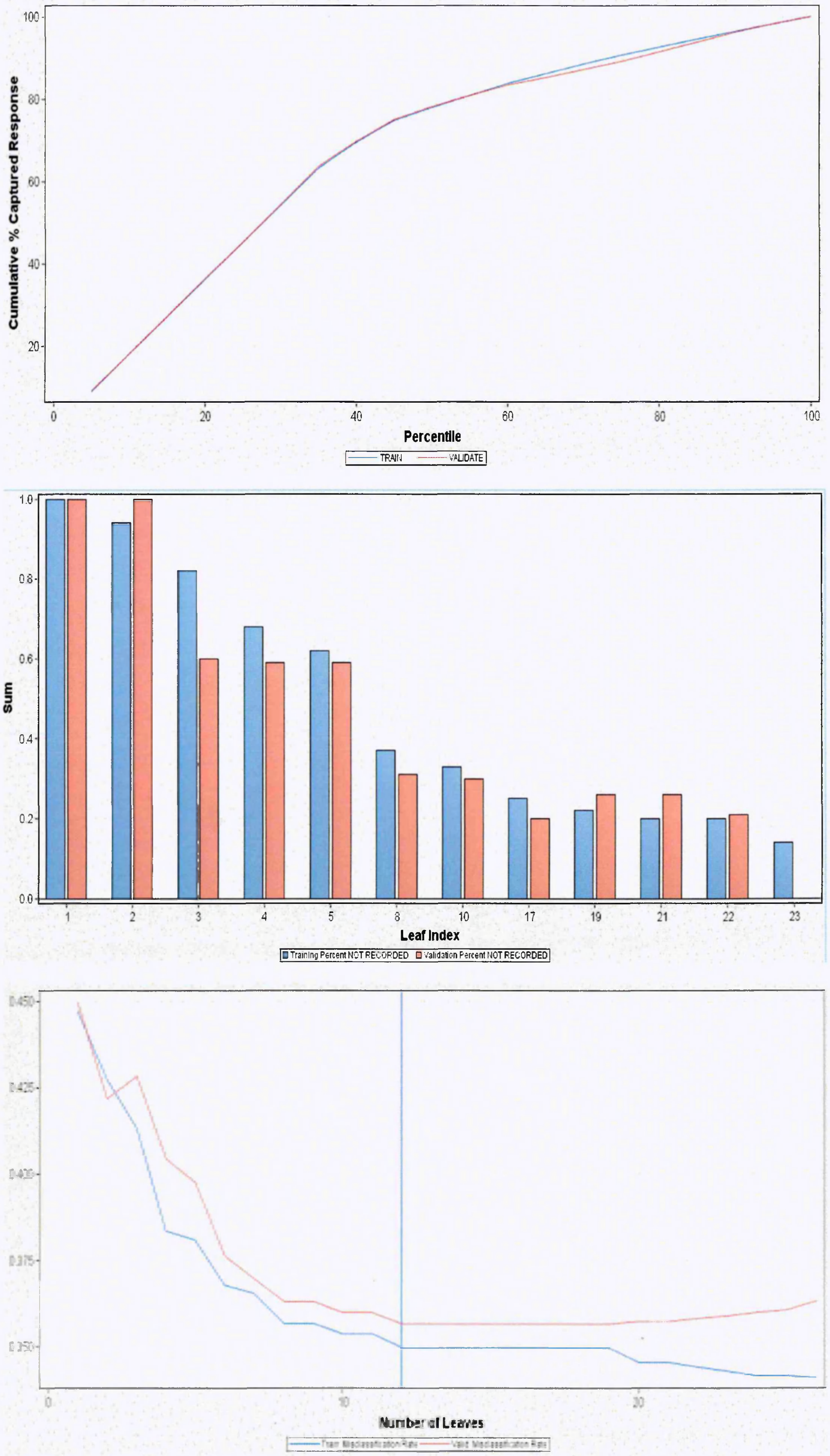


Figure 7.27 – Model Assessment.

The %captured response chart (top), shows that the top 20% of responses captures over 35% of the student employment type data for the validation and train data sets. These translate into reasonably high lift (see glossary page ix) rates (approximately 1.75) and there is very little difference between the train and validate lift rates. With the exception of one leaf, the leaf index bar chart (middle) shows a reasonably balanced amount of training and validation data in each leaf of the final model. The close proximity of the misclassification rates can be seen in the misclassification chart (bottom), this also confirms that a tree with 11 leafs is the optimal solution. Furthermore, the variables selected by the tree discussed early are also sensible when considered in relation to the student retention and EDM literature reviewed in Chapter 3.

7.1.3.6 The Final EM

Figure 7.28, below, provides an overview of the final Entropy tree model.

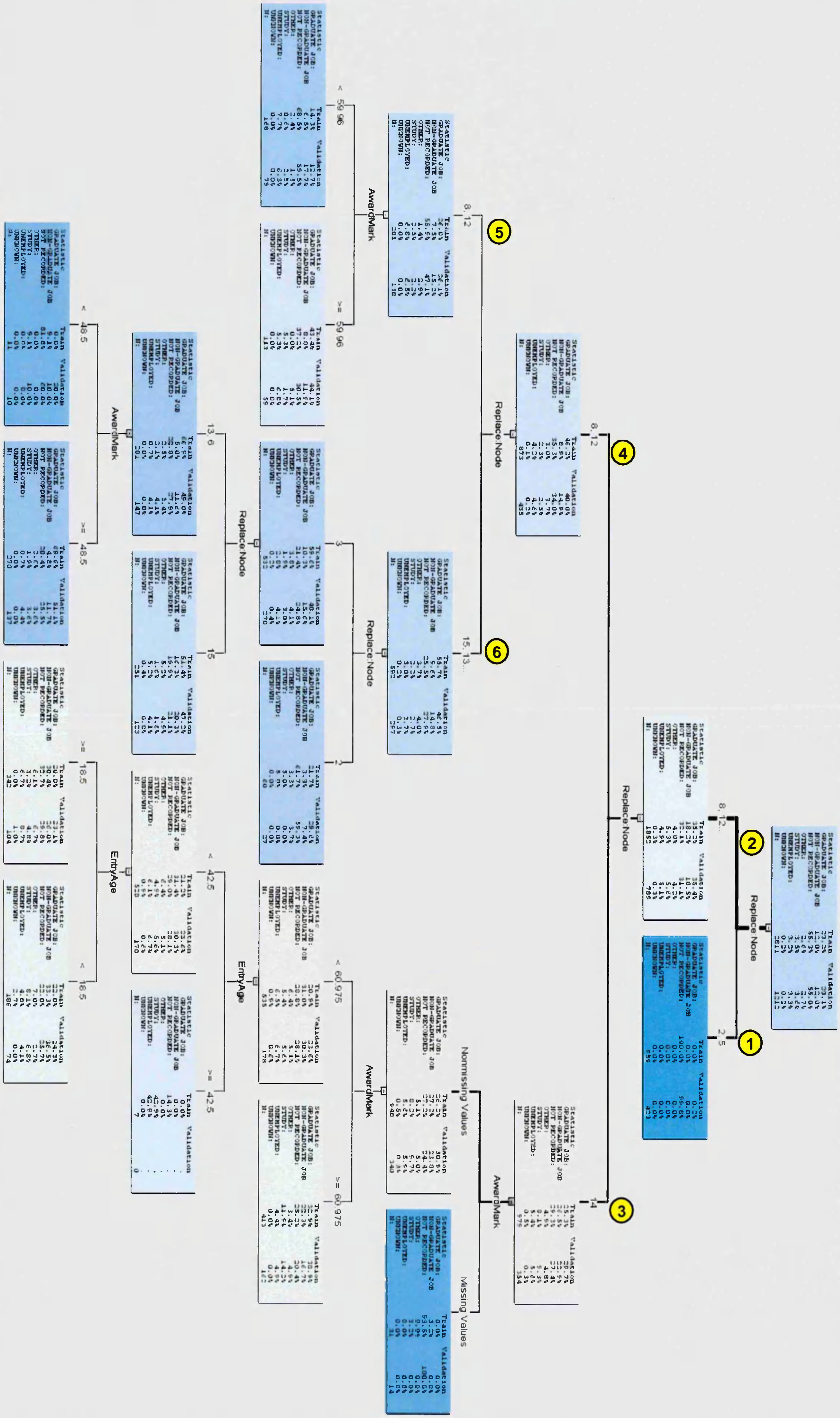


Figure 7.28 - Final Entropy Tree.

Firstly the tree splits CourseGP then, following the left side of the tree, splits CourseGP again into two groups. Moving down the right hand side (3) the tree then splits on AwardMark followed by EntryAge and finally further refines EntryAge. Moving down the left hand side (4) the tree splits out CourseGP again, the left hand side (5) is then split by AwardMark the right hand side (6) splits EthnicityGP followed by CourseGP and then AwardMark. This creates the following rules, below.

```
IF  Node IS ONE OF: 2 5
THEN                          ①
  NODE     :         3
  N        :       959
  NOT RECO:  100.0%
  UNEMPLOY:     0.0%
  NON-GRAD:     0.0%
  OTHER    :    0.0%
  GRADUATE:     0.0%
  STUDY    :    0.0%
  UNKNOWN :     0.0%
```

```
IF  AwardMark  IS MISSING
AND Node EQUALS 14
THEN
  NODE     :         9    ②
  N        :        31
  NOT RECO:    93.5%
  UNEMPLOY:     0.0%
  NON-GRAD:     3.2%
  OTHER    :    0.0%
  GRADUATE:     0.0%
  STUDY    :    3.2%
  UNKNOWN :     0.0%
```

```
IF  59.96 <= AwardMark
AND Node IS ONE OF: 8 12
THEN                              ③
  NODE     :        11
  N        :       113
  NOT RECO:    37.2%
  UNEMPLOY:     5.3%
  NON-GRAD:     8.8%
  OTHER    :    0.0%
  GRADUATE:    43.4%
  STUDY    :    5.3%
  UNKNOWN :     0.0%
```

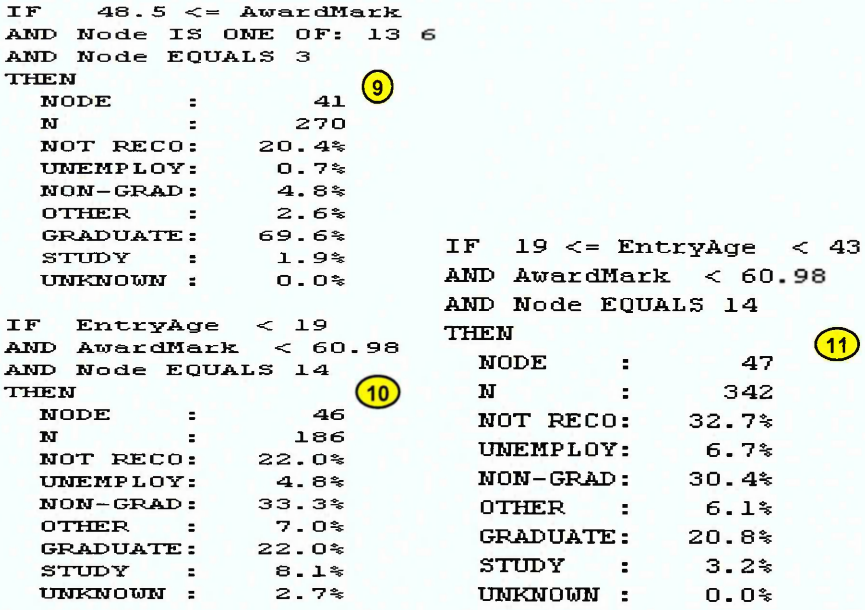
```
IF  Node EQUALS 2
AND Node IS ONE OF: 15 13 6
THEN
  NODE     :        13      ④
  N        :        60
  NOT RECO:    61.7%
  UNEMPLOY:     5.0%
  NON-GRAD:     3.3%
  OTHER    :    3.3%
  GRADUATE:    21.7%
  STUDY    :    5.0%
  UNKNOWN :     0.0%
```

```
IF  60.98 <= AwardMark
AND Node EQUALS 14
THEN                          ⑤
  NODE     :        15
  N        :       413
  NOT RECO:    25.2%
  UNEMPLOY:     4.4%
  NON-GRAD:    22.3%
  OTHER    :    3.4%
  GRADUATE:    32.9%
  STUDY    :   11.9%
  UNKNOWN :     0.0%
```

```
IF  Node EQUALS 15
AND Node EQUALS 3
THEN                          ⑥
  NODE     :        22
  N        :       251
  NOT RECO:    19.9%
  UNEMPLOY:     5.2%
  NON-GRAD:    16.3%
  OTHER    :    5.2%
  GRADUATE:    51.4%
  STUDY    :    1.6%
  UNKNOWN :     0.4%
```

```
IF  43 <= EntryAge
AND AwardMark  < 60.98
AND Node EQUALS 14           ⑦
THEN
  NODE     :        27
  N        :         7
  NOT RECO:    14.3%
  UNEMPLOY:    42.9%
  NON-GRAD:     0.0%
  OTHER    :    0.0%
  GRADUATE:     0.0%
  STUDY    :   42.9%
  UNKNOWN :     0.0%
```

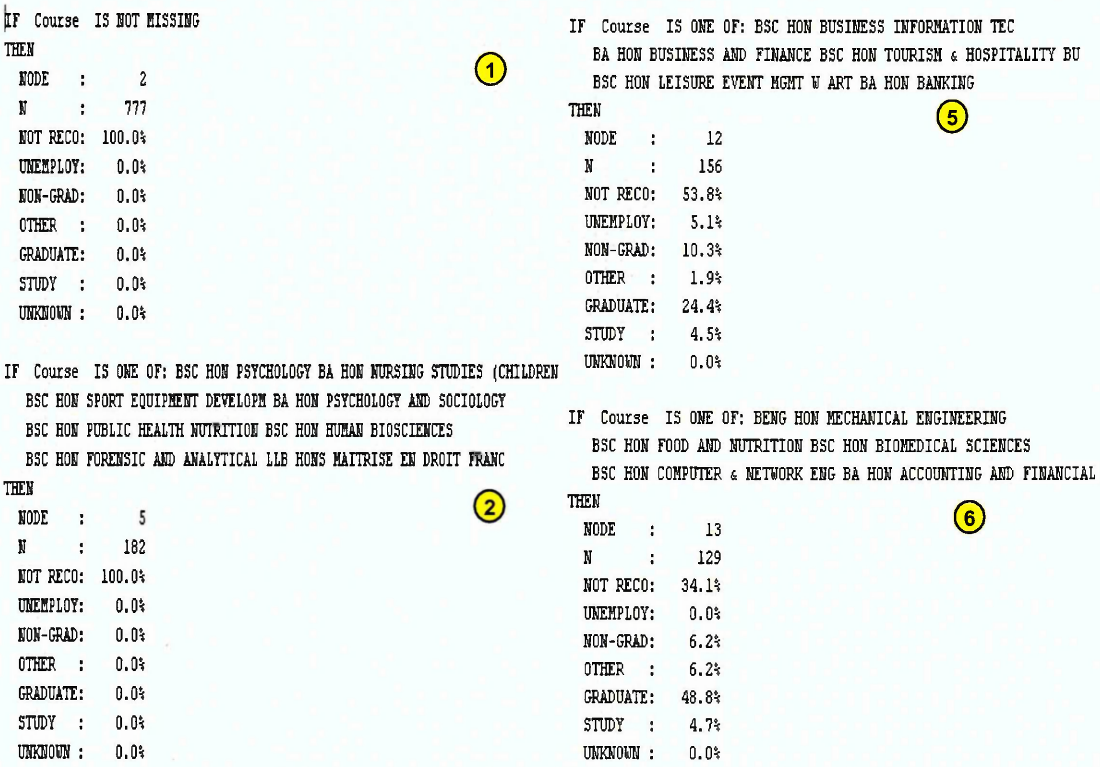
```
IF  AwardMark  <  48.5
AND Node IS ONE OF: 13 6
AND Node EQUALS 3             ⑧
THEN
  NODE     :        40
  N        :        11
  NOT RECO:    81.8%
  UNEMPLOY:     0.0%
  NON-GRAD:     9.1%
  OTHER    :    0.0%
  GRADUATE:     0.0%
  STUDY    :    9.1%
  UNKNOWN :     0.0%
```



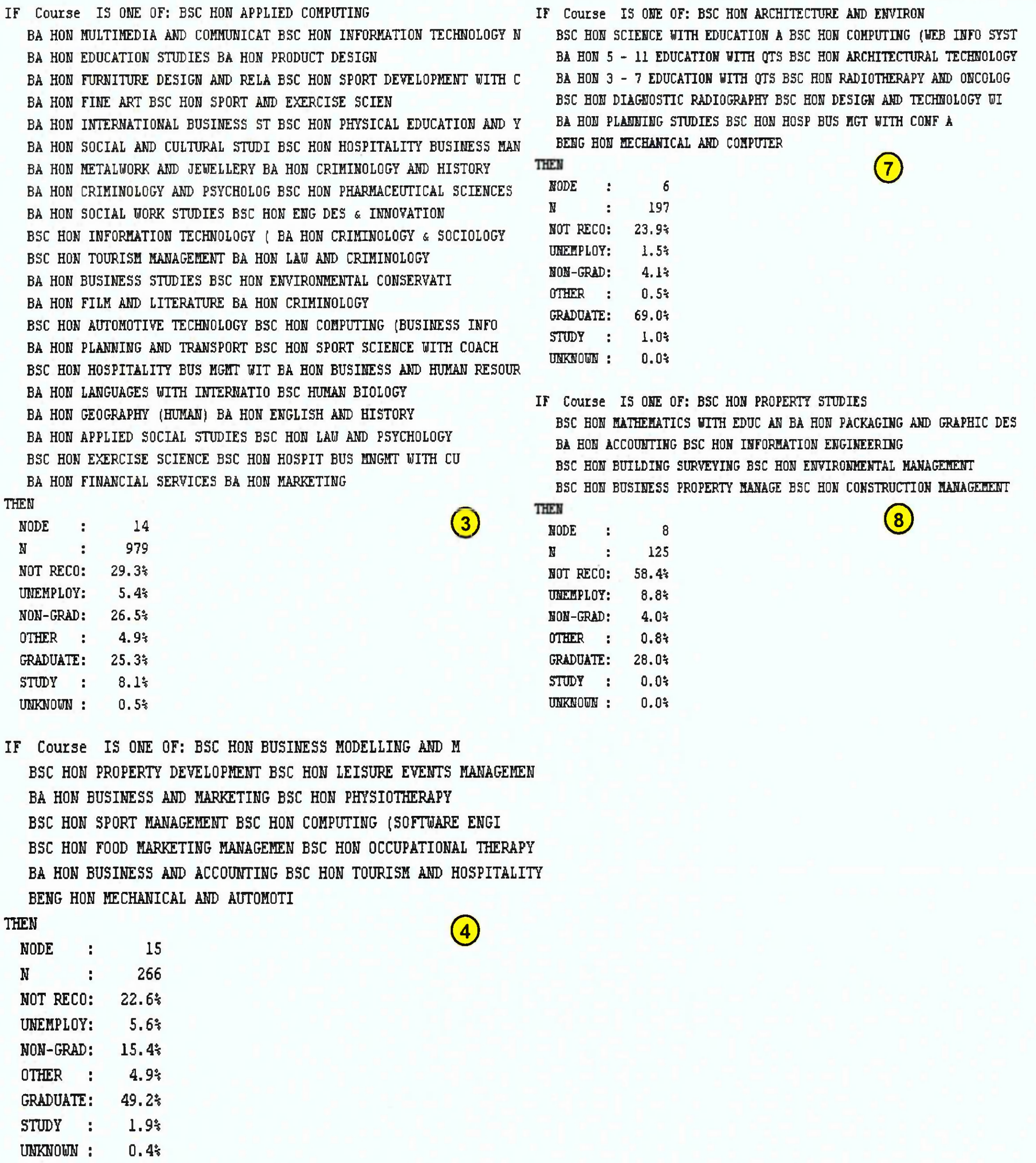
```
IF    48.5 <= AwardMark
AND Node IS ONE OF: 13 6
AND Node EQUALS 3
THEN
   NODE      :        41
   N         :       270
   NOT RECO:     20.4%
   UNEMPLOY:      0.7%
   NON-GRAD:      4.8%
   OTHER    :     2.6%
   GRADUATE:     69.6%
   STUDY    :     1.9%
   UNKNOWN  :     0.0%

IF   EntryAge   < 19
AND AwardMark   < 60.98
AND Node EQUALS 14
THEN
   NODE      :        46
   N         :       186
   NOT RECO:     22.0%
   UNEMPLOY:      4.8%
   NON-GRAD:     33.3%
   OTHER    :     7.0%
   GRADUATE:     22.0%
   STUDY    :     8.1%
   UNKNOWN  :     2.7%

IF   19 <= EntryAge   < 43
AND AwardMark   < 60.98
AND Node EQUALS 14
THEN
   NODE      :        47
   N         :       342
   NOT RECO:     32.7%
   UNEMPLOY:      6.7%
   NON-GRAD:     30.4%
   OTHER    :     6.1%
   GRADUATE:     20.8%
   STUDY    :     3.2%
   UNKNOWN  :     0.0%
```

Figure 7.29 - Final Rules.

As before in order to help understand these rules, the course groupings have been repeated below from the modify section.



```
IF  Course  IS NOT MISSING
THEN
  NODE     :       2
  N        :     777
  NOT RECO:  100.0%
  UNEMPLOY:    0.0%
  NON-GRAD:    0.0%
  OTHER    :   0.0%
  GRADUATE:    0.0%
  STUDY    :   0.0%
  UNKNOWN :    0.0%

IF  Course  IS ONE OF: BSC HON PSYCHOLOGY BA HON NURSING STUDIES (CHILDREN
   BSC HON SPORT EQUIPMENT DEVELOPM BA HON PSYCHOLOGY AND SOCIOLOGY
   BSC HON PUBLIC HEALTH NUTRITION BSC HON HUMAN BIOSCIENCES
   BSC HON FORENSIC AND ANALYTICAL LLB HONS MAITRISE EN DROIT FRANC
THEN
  NODE     :       5
  N        :     182
  NOT RECO:  100.0%
  UNEMPLOY:    0.0%
  NON-GRAD:    0.0%
  OTHER    :   0.0%
  GRADUATE:    0.0%
  STUDY    :   0.0%
  UNKNOWN :    0.0%

IF  Course  IS ONE OF: BSC HON BUSINESS INFORMATION TEC
   BA HON BUSINESS AND FINANCE BSC HON TOURISM & HOSPITALITY BU
   BSC HON LEISURE EVENT MGMT W ART BA HON BANKING
THEN
  NODE     :      12
  N        :     156
  NOT RECO:   53.8%
  UNEMPLOY:    5.1%
  NON-GRAD:   10.3%
  OTHER    :   1.9%
  GRADUATE:   24.4%
  STUDY    :   4.5%
  UNKNOWN :    0.0%

IF  Course  IS ONE OF: BENG HON MECHANICAL ENGINEERING
   BSC HON FOOD AND NUTRITION BSC HON BIOMEDICAL SCIENCES
   BSC HON COMPUTER & NETWORK ENG BA HON ACCOUNTING AND FINANCIAL
THEN
  NODE     :      13
  N        :     129
  NOT RECO:   34.1%
  UNEMPLOY:    0.0%
  NON-GRAD:    6.2%
  OTHER    :   6.2%
  GRADUATE:   48.8%
  STUDY    :   4.7%
  UNKNOWN :    0.0%
```



```
IF  Course  IS ONE OF: BSC HON APPLIED COMPUTING
   BA HON MULTIMEDIA AND COMMUNICAT BSC HON INFORMATION TECHNOLOGY N
   BA HON EDUCATION STUDIES BA HON PRODUCT DESIGN
   BA HON FURNITURE DESIGN AND RELA BSC HON SPORT DEVELOPMENT WITH C
   BA HON FINE ART BSC HON SPORT AND EXERCISE SCIEN
   BA HON INTERNATIONAL BUSINESS ST BSC HON PHYSICAL EDUCATION AND Y
   BA HON SOCIAL AND CULTURAL STUDI BSC HON HOSPITALITY BUSINESS MAN
   BA HON METALWORK AND JEWELLERY BA HON CRIMINOLOGY AND HISTORY
   BA HON CRIMINOLOGY AND PSYCHOLOG BSC HON PHARMACEUTICAL SCIENCES
   BA HON SOCIAL WORK STUDIES BSC HON ENG DES & INNOVATION
   BSC HON INFORMATION TECHNOLOGY ( BA HON CRIMINOLOGY & SOCIOLOGY
   BSC HON TOURISM MANAGEMENT BA HON LAW AND CRIMINOLOGY
   BA HON BUSINESS STUDIES BSC HON ENVIRONMENTAL CONSERVATI
   BA HON FILM AND LITERATURE BA HON CRIMINOLOGY
   BSC HON AUTOMOTIVE TECHNOLOGY BSC HON COMPUTING (BUSINESS INFO
   BA HON PLANNING AND TRANSPORT BSC HON SPORT SCIENCE WITH COACH
   BSC HON HOSPITALITY BUS MGMT WIT BA HON BUSINESS AND HUMAN RESOUR
   BA HON LANGUAGES WITH INTERNATIO BSC HUMAN BIOLOGY
   BA HON GEOGRAPHY (HUMAN) BA HON ENGLISH AND HISTORY
   BA HON APPLIED SOCIAL STUDIES BSC HON LAW AND PSYCHOLOGY
   BSC HON EXERCISE SCIENCE BSC HON HOSPIT BUS MNGMT WITH CU
   BA HON FINANCIAL SERVICES BA HON MARKETING
THEN
  NODE     :     14
  N        :    979
  NOT RECO:   29.3%
  UNEMPLOY:    5.4%
  NON-GRAD:   26.5%
  OTHER    :   4.9%
  GRADUATE:   25.3%
  STUDY    :   8.1%
  UNKNOWN :    0.5%
```

```
IF  Course  IS ONE OF: BSC HON ARCHITECTURE AND ENVIRON
   BSC HON SCIENCE WITH EDUCATION A BSC HON COMPUTING (WEB INFO SYST
   BA HON 5 - 11 EDUCATION WITH QTS BSC HON ARCHITECTURAL TECHNOLOGY
   BA HON 3 - 7 EDUCATION WITH QTS BSC HON RADIOTHERAPY AND ONCOLOG
   BSC HON DIAGNOSTIC RADIOGRAPHY BSC HON DESIGN AND TECHNOLOGY WI
   BA HON PLANNING STUDIES BSC HON HOSP BUS MGT WITH CONF A
   BENG HON MECHANICAL AND COMPUTER
THEN
  NODE     :      6
  N        :    197
  NOT RECO:   23.9%
  UNEMPLOY:    1.5%
  NON-GRAD:    4.1%
  OTHER    :   0.5%
  GRADUATE:   69.0%
  STUDY    :   1.0%
  UNKNOWN :    0.0%
```

```
IF  Course  IS ONE OF: BSC HON PROPERTY STUDIES
   BSC HON MATHEMATICS WITH EDUC AN BA HON PACKAGING AND GRAPHIC DES
   BA HON ACCOUNTING BSC HON INFORMATION ENGINEERING
   BSC HON BUILDING SURVEYING BSC HON ENVIRONMENTAL MANAGEMENT
   BSC HON BUSINESS PROPERTY MANAGE BSC HON CONSTRUCTION MANAGEMENT
THEN
  NODE     :      8
  N        :    125
  NOT RECO:   58.4%
  UNEMPLOY:    8.8%
  NON-GRAD:    4.0%
  OTHER    :   0.8%
  GRADUATE:   28.0%
  STUDY    :   0.0%
  UNKNOWN :    0.0%
```

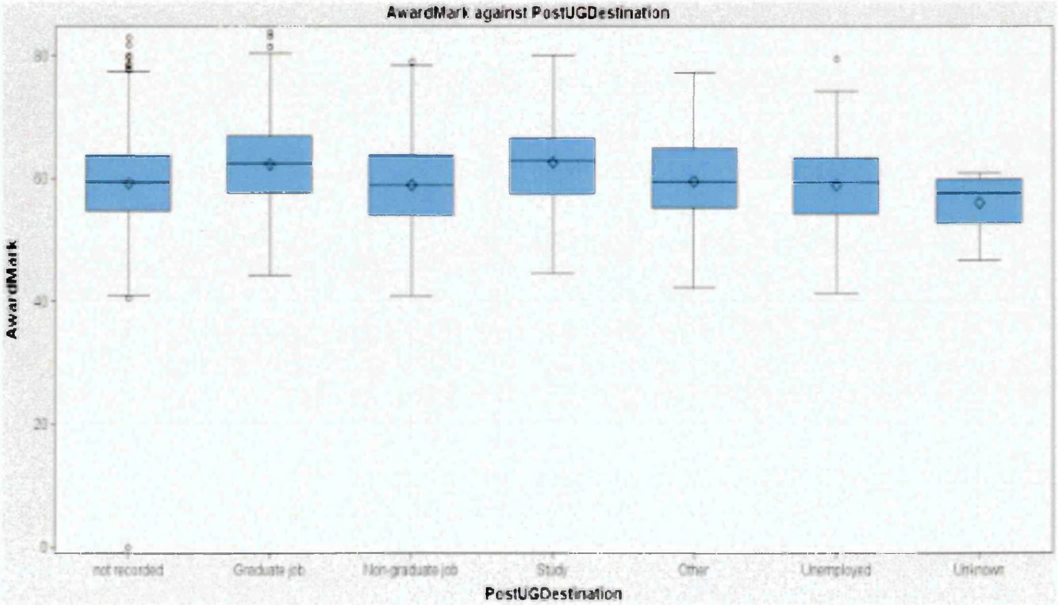
```
IF  Course  IS ONE OF: BSC HON BUSINESS MODELLING AND M
   BSC HON PROPERTY DEVELOPMENT BSC HON LEISURE EVENTS MANAGEMEN
   BA HON BUSINESS AND MARKETING BSC HON PHYSIOTHERAPY
   BSC HON SPORT MANAGEMENT BSC HON COMPUTING (SOFTWARE ENGI
   BSC HON FOOD MARKETING MANAGEMEN BSC HON OCCUPATIONAL THERAPY
   BA HON BUSINESS AND ACCOUNTING BSC HON TOURISM AND HOSPITALITY
   BENG HON MECHANICAL AND AUTOMOTI
THEN
  NODE     :     15
  N        :    266
  NOT RECO:   22.6%
  UNEMPLOY:    5.6%
  NON-GRAD:   15.4%
  OTHER    :   4.9%
  GRADUATE:   49.2%
  STUDY    :   1.9%
  UNKNOWN :    0.4%
```

These rules can be interpreted as follows:

Rule (1)	If CourseGP is CourseGP1 or CourseGP2 then student employment type is not recorded.
Rule (2)	If AwardMark is missing and if CourseGP is CourseGP3 then student employment type is likely to be not recorded.
Rule (3)	If AwardMark is greater than or equal to 59.96 and CourseGP is CourseGP8 or CourseGP5 then the student is 43.3% likely to obtain a graduate job.
Rule (4)	If EthnicityGP is Mixed Group and CourseGP is one of CourseGP4, CourseGP6 or CourseGP7 then the likely student employment type is not recorded.
Rule (5)	If AwardMark is greater than or equal to 60.68 and CourseGP is CourseGP3 the student is more likely to obtain a graduate job.
Rule (6)	If CourseGP is CourseGP4 and EthnicityGP is Mainly White then student employment type is more likely to be a graduate job.
Rule (7)	If EntryAge is greater than or equal to 43 and AwardMark is less than 60.98 and CourseGP is CourseGP3 then students are more likely to end up unemployed (42.9%) or studying (42.9%).

Rule (8)	If AwardMark is less than 48.5 and CourseGP is one of CourseGP6 or CourseGP7 and EthnicityGP is Mainly White. Then the employment type is likely to be not recorded.
Rule (9)	If AwardMark is greater than or equal to 48.5 and CourseGP is one of CourseGP6 or CourseGP7 EthnicityGP is Mainly White. Then it is highly likely (69.6%) that the student will end up in a graduate role.
Rule (10)	If EntryAge is less than 19 and AwardMark is less than 60.98 and CourseGP is CourseGP3 then it is likely that the student will end up in a non-graduate job.
Rule (11)	If EntryAge is greater than or equal to 19 and AwardMark is less than 60.98 and CourseGP is CourseGP3 then it is likely that the student employment type will be not recorded (32.7%) or non-graduate job (30.4%).

The rules above suggest that the non-white ethnic groups are less likely to obtain a graduate job. This further highlights the possibility of a dependency between ethnicity and award mark, which ultimately impacts upon employment type. Arguably, of all of the three models developed the student employment type model has provided the best train and misclassification rates, 34% and 35% respectively. The variables and values selected by the final model are sensible given the exploration of the data, carried out previously. The author is confident that this model will prove to be a good predictor of student employment type, which is confirmed through an assessment of the exploration graphs in relation to rule 9.



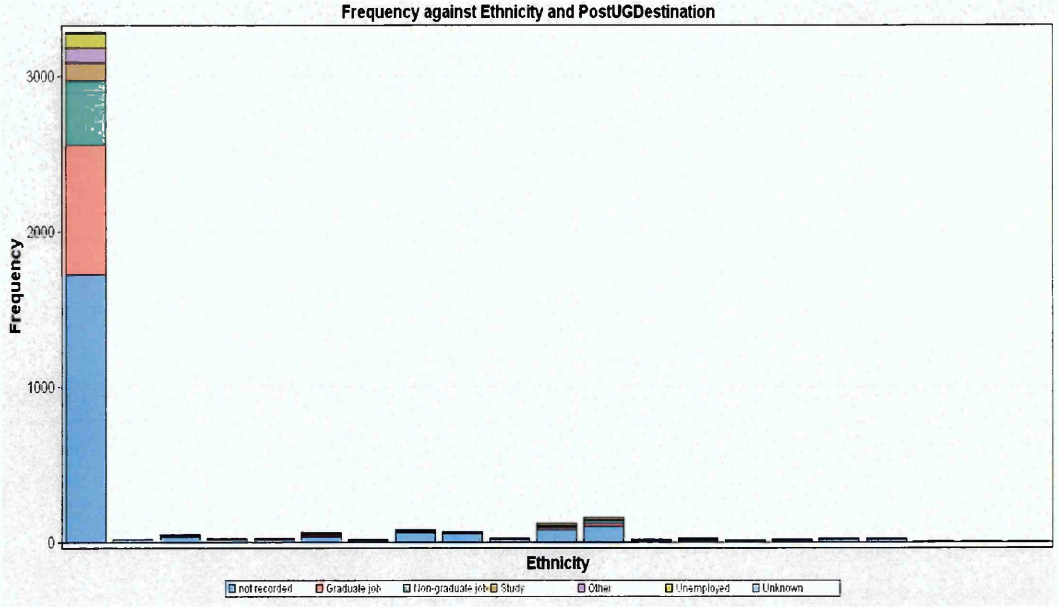


Figure 7.30 – Assessing Rule 9 Plots.

The box plot above confirms that the majority of students who obtain graduate jobs on average are awarded a mark that is greater than 60%. The bar chart, below, also suggests that more White British students obtain graduate level jobs than any other ethnic group.

7.2 Summary

In this chapter, the process of transforming structured data into knowledge and ultimately intelligence, as outlined in the Knowledge Pyramid in figure 4.1, is illustrated using SAS® SEMMA. The resulting models can be used to predict student award classification, progression onto postgraduate studies at SHU, and employment type. Whilst the predictive accuracy of some of the models developed is questionable, it is thought that the predictive accuracy would be much improved with a larger data set. Overall, the results of the modelling process are appropriate in the context of each individual target variable. Further to this, the misclassification rates are also representative of similar studies that have tried to predict student behaviour (Herzog 2006, Superby *et al.* 2006).

8 Findings

Having now built the three models (award classification, progression onto postgraduate studies and student employment type) this chapter will discuss the main finding from each and where applicable relate it to the exploration of the data and the literature.

8.1 Award Classification

Through the modelling of the data, outlined in Chapter 7, thirteen rules were identified (see section 7.1.1.6) these rules show that there are six key variables for predicting student award classification. These variables, in order of importance are:

- Course
- Entry Points
- Ethnicity
- Undergraduate Entry Age
- QYPR
- QAHE.

These variables are sensible predictors of award classification when they're considered in relation to the research that has been conducted previously and the exploration of the data carried out in section 6.1.3.2.4.

Indeed, it is noted that the students are more likely to progress when they make informed decisions about their course (Yorke and Longden 2008, Moxley *et al.* 2001). Superby *et al.* (2006) and Dekker *et al.* (2009) also identified course has been a key variable in predicting student behaviour. Unfortunately, the importance of the course variable, in predicting award classification, was not recognised at the data exploration stage due to the large number of levels associated with this variable.

The student's previous academic ability, which in this case is reflected in their entry points, is also identified as being fundamental to student progression (Yorke and Longden 2008, Moxley *et al.* 2001). The EDM literature also suggests that the previous academic experience is vital in predicting student behaviour. The relationship between entry points and award classification was identified at the data exploration stage and it suggested that students entering with higher entry points achieve a better award classification. Therefore, the inclusion of this variable in some of the final rules was anticipated.

The identification of ethnicity as a predictor of award classification is relatively unique in terms of the EDM literature. However, the inclusion of this variable in some of the rules is not surprising as the 1997 Labour Government identified that there was a need to widen and increase participation of students from certain ethnic groups. The data exploration and DM process confirmed that certain ethnic groups are indeed underrepresented and those groups tended to achieve a lower award classification.

The relationship between the students maturity and how it can effect progression are well documented (Yorke and Longden 2008). Herzog (2006) also identifies that students age is an important variable in predicting student behaviour. The rules containing undergraduate entry age reflect the patterns identified at the data exploration stage. In that students who entered their degree at a younger age tended to achieve more 1st and 2:1 classifications whereas students obtaining 2:2 and 3rd class honours tended to be older. Interestingly, this contradicts some of the progression literature where the students age is considered to have an effect on progression. Nonetheless this still indicates that age is important in trying to predict student behaviour.

Arguably, the incentives around the widening and increasing participation agenda suggests that there are some links to social deprivation and the students award classification. Therefore, it is sensible to include Young Peoples Participation in HE by Postcode (QYPR) and Adults with Higher Education Qualifications by Postcode (QAHE) in the final model. However, some of the initial exploration of these variables identified a number of anomalies such as students from where QYPR was high tended to achieve a 2:1 whereas students from areas where QYPR was low-medium tend to obtain more 1st classifications, a similar pattern was also identified for QAHE. This trend is believed to be valid due to the fact that, as stated in 3.2.1, non-traditional students are being catered for by the post 1992 universities (Archer 2002). In addition to this, the National Audit Office (2007) report, shows that SHU managed to increase participation whilst also improving student progression from 91.2% in 2001-02 to 92.3% in 2004-05.

Furthermore, through the process of mining the data, a number of key rules came to light in relation to student award classification. It is perhaps important to point out that the grouping, such as the students ethnicity, were automatically grouped by the SAS® software. The final rules were:

- Students are more likely to obtain a 2:1 on a business, management or marketing related course;
- Young (age on entry <23) white students who enter with less than 205 entry points onto certain health, property, design, biology, sports or engineering related courses are more likely to achieve a lower award classification;
- White students who enrol, with less than 205 entry points, onto food, health, business, sports, property, social care, education or film related courses tend to achieve a better award classification than non-white students;
- Students who enrol, with 300 or more entry points, onto food, health, business, sports, property, social care, education or film related courses tend to achieve a higher award classification;
- Students are less likely to achieve a high classification when studying certain health, sports, social work, English, film, computing or tourism related courses when entering with less than 300 entry points;
- Older (age on entry >=23) white students entering certain food, health, business, sports, property, social care, education or film related courses with high entry points are likely to achieve a high award classification;
- Older (age on entry >=23) white students with lower entry points, who study food, health, business, sports, property, social care, education or film related courses, tend to achieve better award classifications from areas where adult participation in HE is higher;
- Black African students with less than 300 entry points are more likely to achieve lower classifications on certain sports, education, business, management, health or property related courses than other students; and
- Students on food and nutrition, property development, science with education, business or media studies related courses are more likely to achieve a higher award classification.

8.2 Postgraduate Studies

The modelling of progression onto postgraduate studies identified 5 rules (see section 7.1.2.6), these rules indicate that there are four variables that are fundamental to predicting progression onto postgraduate studies. These are:

- Undergraduate Course
- Undergraduate Award Mark
- QYPR
- Undergraduate Entry Points.

Given the findings of the literature review and the exploration of the data these variables are sensible predictors of the target.

Indeed, as previously discussed, the importance of the course variable in predicting student behaviour is well documented in the EDM literature and the student progression literature also places a large emphasis on the course (Yorke and Longden, 2008, Moxley *et al.* 2001, Superby *et al.* 2006, Dekker *et al.* 2009). The inclusion of course isn't surprising as the analysis of the JACS subjects (course groups) identified that some of the JACS subjects had more students who progressed onto postgraduate students than others (see section 6.1.3.2.4).

The inclusion of the undergraduate award mark variable is also relevant in that the data exploration highlighted that the undergraduate award mark of those students who go on to study a postgraduate degree at SHU is higher than those that don't, this is particularly true of those students who obtain a 1st or a 2:1 classification. However, this trend may be due to entry requirements for postgraduate courses. Additionally, in unrelated studies, Dekker *et al.* (2009) and Superby *et al.* (2006) have identified past achievements (in terms of grades) has been a key predictor of student behaviour.

The presence of Young People in HE by Postcode (QYPR) variable in the final model is reasonable. As previously pointed out this variable poses some interesting questions regarding the type of students SHU caters for. Indeed, contrary to the authors own perceptions, it was identified at the data exploration stage that more students from areas where QYPR is low tended to go on to take postgraduate studies at SHU, this could be due to the fact that support, in terms of funding, is more widely available to students from those areas and/or ‘added value’ (see glossary page vii). Furthermore, this trend is reaffirmed by the fact that students from areas where QYPR is low-medium tended to achieve a higher award classification.

Again, the inclusion of undergraduate entry points in this model is sensible, as previous academic ability/experience have been identified as being essential to understanding student behaviour. From the author's perspective, undergraduate entry points and undergraduate entry qualifications are the only measure that is available in the SI data to determine the student's previous academic experience. A relationship between those students who go on to study a postgraduate degree and undergraduate entry points was identified at the data exploration stage. This suggested that students with higher entry points tended to take further postgraduate studies at SHU.

Additionally, through the process of mining the data, a number of key rules came to light in relation to student progression onto postgraduate studies at SHU. These rules were:

- Certain undergraduate courses have higher employability rates (such as law, education and computing (specialist)) and are therefore less likely to take postgraduate studies, at SHU, soon after completing a degree;
- Students are more likely to take postgraduate studies, at SHU, after studying undergraduate courses such as those associated with business, planning or psychology;
- Students who take business, finance, education QTS, design and technology or planning related undergraduate courses and achieve a low award mark are less likely to take postgraduate studies at SHU; and
- Students who achieve an award mark greater than 55 but entered HE with less than 300 entry points are more likely to take postgraduate studies (at SHU) when they have studied business, finance, education QTS, design and technology or planning related undergraduate courses.

8.3 Student Employment Type

The DM process determined eleven rules for predicting student employment type, see section 7.1.3.6. These rules are composed of four variables that are key to predicting the target. The variables include:

- Course
- Award Mark
- Ethnicity
- Undergraduate Entry Age.

Arguably, in the light of the literature review and data exploration these variables are practical.

Again, as with the previous two models the course variable is the most pertinent variable in predicting student employment type, the importance of this variable in both the student progression and EDM literature has been noted previously. Due to the large number of course values within this variable no trends were identified. However, the exploration of JACS subject did identify certain course groups that have large number of students who are unemployed, don't respond to the DOL survey, obtain a graduate job and the like. This raised some interesting questions at the data exploration stage as to why some SHU courses had a 100% nonresponse rate to the DOL survey.

The inclusion of award mark in the student employment type model is sensible as the final award classification, determined by the award mark, is a key requisite to obtaining a graduate job. Indeed, the data exploration identified that students who achieved a 1st classification are more likely to obtain a graduate job or take further studies. This supports the inclusion of award mark in the model built to predict progression onto postgraduate studies at SHU. Further general patterns found in the data, reflected in the final model include students obtaining a 2:2 or a 3rd tended to end up in non-graduate jobs or unemployed and student response rates to the DOL survey tends to decrease as the award mark reduces. In addition to this, previous EDM studies have indicated the importance of award mark in predicting student behaviour (Dekker *et al*. 2009, Superby *et al* 2006).

The inclusion of the Ethnicity variable in the final model is interesting, as previously pointed out there has been some indication, through the widening and increasing participation agenda, of the importance of this variable. The data exploration identified that the majority of students who obtained a graduate job were white and students obtaining graduate jobs from other ethnic groups were in the minority. It is possible that further investigation into this trend could identify some interesting barriers to obtaining a graduate job, such as English verbal and written skills.

In addition to this, the inclusion of undergraduate entry age in the final model is valid as maturity related problems are documented in the student progression literature and it is also noted as being a key variable in predicting student behaviour (Yorke and Longden 2008, Herzog 2006). The relationship between the students maturity and how it can effect progression are well documented (Yorke and Longden 2008). Herzog (2006) also identifies student's age as being an important variable in predicting student behaviour. The exploration of undergraduate entry age against student employment type indicated that older students were likely to find it harder to obtain a job after completing their undergraduate degrees. Indeed, this is reflected in rule 7 in section 7.1.3.6.

Further to this, through the process of mining the data, a number of key rules came to light in relation to student employment type. Again it is important to state that the groupings, such as the students ethnicity, were automatically created by the SAS® software. These rules were:

- There is either a problem with recording DOL data or students who study certain undergraduate courses and achieve low award marks tend not to respond to the DOL survey;
- Students are more likely to obtain a graduate job when they obtain a higher award mark in certain business, management, property, design, finance, sports, humanities, computing, art, social, criminology or education related courses;
- White students are more likely to obtain a graduate job when studying management, finance or business related courses;
- White students on engineering, food, computing (specialist), science, education, health, planning or design and technology related courses are more likely to end up in graduate roles even when achieving a lower award mark;
- Older (age on undergraduate entry >=43) students who achieve a lower award mark (<2:1), on computing, education, design, sports, art, business, criminology, social, management, planning, English or humanities related courses are less likely to end up in employment; and
- Younger (age on undergraduate entry <19) students who achieve a lower award mark on certain computing (general), education, design, sports, art, business, criminology, social, management, planning, English or humanities related courses are less likely to get a graduate job.

8.4 SUMMARY

This chapter sets out the main findings from the data modelling stage for each of the predictive models developed. It presents some discussion around the key variables for all of the models. The modelling process identified that course, award mark, undergraduate entry age, entry points, ethnicity, QAHE and QYPR were key in predicting student behaviour. These variables thus offer sensible, predictors of the three target variables, as the literature on student progression and EDM also highlights the importance of such variables in predicting student behaviour (Yorke and Longden 2008, Herzog 2006, Dekker *et al.* 2009, Superby *et al.* 2006). Overall, the final models developed through the DM stage of this research are valid within the context of SHU. Indeed, the difficulties associated with classifying students in an objective way are well documented, as student behaviour is dependent on many factors which are likely to change from institution to institution (Dekker *et al.* 2009).

9 Recommendations

In realising objective 6, this chapter will present the main findings from the research as a number of recommendations. These recommendations will be useful for the HE stakeholders (SHU, students and the state) introduced in Chapter 3, vendors of DM software and future researches interested in the subject area. The recommendations will be complied around data capture, use of DM, required skill levels to carry out such projects, further work involving DM, mining other available data and potential new studies involving DM.

9.1 Data Capture

Through the DW process it transpired that there were a number of issues with the SI data. These issues related to incorrect and incomplete data being included in the subset of data obtained from SHU. There appears to be a lack of validation in the front end software that has affected the quality of the data. Therefore, it is recommended that SHU make some investment in training staff about the importance of complete and correct data. Furthermore, difficulties were also identified with the Entry Qualification type variable in that it wasn't possible to separate students entering their degrees with A-Level/GNVQ3 qualifications. Therefore, the Award Mark was favoured over the Entry Qualification type variable. Although, there seems to be a lack of transparency regarding how Entry Qualifications are translated into Award Marks. Consequently, it is recommended that SHU split the A-Level/GNVQ3 qualification into two distinct values, as it will improve the ability of any future DM models. However, this recommendation has been superseded as GNVQ no longer exists.

It is also important to note that further models could be built if the data was available, but SHU doesn't record anything about progression. Finally, SHU should try to improve responses to the DOL survey, as this would improve the predictive power of any future models developed whilst also providing a greater understanding of student opinions. Indeed, it was identified that some SHU courses had a 100% non-response rate to the DOL survey and as student award classifications decreased so did the likelihood of the student responding.

In summary the main recommendations from this sub-section are:

1. Improve the quality and completeness of the SI data;
2. Improve staff awareness regarding the importance of good quality data;
3. Separate dual values within variables and stop using the duplicate;
4. Improve transparency regarding how student entry point are calculated; and
5. Investigate poor student response rates to DOL survey.

9.2 Use of Data Mining

The review of the EDM literature showed that there were numerous potential uses for DM within HE, which was highlighted by the previous studies reviewed as part of the EDM review in section 3.3. The results of the EDM review emphasised problems, regarding high levels of misclassification, with such studies that use DM in HE to predict student behaviour. However, DM will prove useful in the future as institutions become better at collecting and understanding their data. Indeed, SHU are in the process of developing their own data warehouse for reporting. This potentially will provide researchers and SHU the opportunity to access a huge repository of clean, verified and merged SI data. This level of data would help to validate some of the patterns identified as part of this study. The DM techniques used here were selected based on the results of previous studies and whilst, in some cases, the decision tree didn't produce the best model. The future use of decision trees, in this context, will provide models that are easier to interpret and thus apply within institutions.

9.3 Required Skills Level

Arguably, the ability of the researcher to carry out the investigation will affect the overall outcome of any research project. Indeed, the complexity of DM and statistical techniques require a certain level of understanding which, if lacking, will inevitably increase the time to complete such studies in the future and potentially affect the quality of any outcomes. In conjunction with this, understanding national and local HE policies, student progression issues and how institutions record and store their data will also prove vital to the future success of any studies in this area.

9.4 Further Work Involving Data Mining at SHU

Further work involving DM in HE to predict student behaviour at SHU will be considered, from three perspectives, in this section. These can be grouped into the following categories: the further exploration of the data collected (both from SI and an online survey); capturing additional data from SHU; and considering current changes to the HE landscape.

9.4.1 Further Exploration of the Collected Data

The further exploration of data falls into two categories, the further exploration of the SI data and the exploration of the data collected through an online survey, each will be discussed separately.

A number of intriguing relationships were identified at the data exploration stage and through the mining of the data. It is these links that inform the following recommendations:

1. examine the link between ethnicity (mainly black) and the small number of these students who take postgraduate studies at SHU;
2. investigate the association between ethnicity (mainly black) and the small number of these students who obtain a graduate role;
3. carry out a similar study but include DLHE data so that this can be compared with other institutions;
4. investigate the significance of the benefits of added value (see glossary page vii) and widening participation;
5. examine the association between high demand skills that are in low supply and obtaining a graduate role;
6. investigate the association between low demand skills that are in high supply and students progressing onto postgraduate studies at SHU; and
7. obtain further insights into the association between older students and unemployment\low levels of employment.

The SI data could also be further explored, through the application of unsupervised DM techniques (such as clustering and market basket analysis), to determine patterns and trends in the data that haven't been explored as part of this directed DM study. Additionally, this research focuses on one method of data analysis, which was of special interest to the author as it is relatively new to HE. However, now that the data has been constructed in the DM mart it would be useful to create cubes (see glossary page vii) and report from them.

Further to this, the data collected as part of an online survey could be analysed, using both directed and undirected DM techniques, for patterns and trends in the data set. The results from this could then be compared to the results of Burleys (2007) study. This would then help determine if the problems identified by Burley are local to the Department of Computing at SHU or if they are transferable across all faculties at the university.

9.4.2 Capturing Additional Data from SHU

Arguably, the models developed as a result of this study were based on a snapshot of the SI data. Therefore it would be interesting to gather further level 6, postgraduate studies and DOL data from other years, about 30,000 records, to determine if the patterns and trends identified as part of this study are reflective across other final years and student groups. However, the author is aware of the problems with obtaining and cleaning 4023 student records. In the future, these problems should be greatly reduced with the advent of the new SHU data warehouse.

9.4.3 Considering Current Changes to the HE Landscape

Since starting this study in 2009, there have been some significant changes in the HE landscape, the main ones being the reduction in the number of university places and the emphasis on student led HE degree funding. The reduction in student numbers has already had an impact on the calibre of students recruited by SHU. Indeed, the cap placed on entry requirement has led to SHU being able to recruit a much higher calibre of student in 2012. It is anticipated that this could produce some interesting DM models for comparison to the models built as part of this study. However, the effects of the cap on student numbers and the increase in tuition fees in predicting undergraduate student award classification, progression onto postgraduate studies at SHU and student employment type will not be modelled until 2015/2016 (depending on students mode of study). This problem is further compounded when modelling progression onto postgraduate studies at SHU. Indeed, this would have to be modelled even later, as the students would have had to complete their undergraduate degrees.

9.5 Data Mining other Available HE Data

There are also further applications of DM in HE outside of SHU. Indeed, this sub-section considers these from four perspectives, the extension of the study to other post and pre 1992 universities, the mining of the National Student Assessment data and exploration of the data submitted by post and pre 1992 universities to HESA.

9.5.1 Extending the Study to other Post 1992 Universities

It would be interesting to determine whether the patterns and trends found as part of this study were repeated at another post-1992 university, such as Nottingham Trent, Huddersfield or Leeds Metropolitan University. Arguably, given the high levels of competition between universities to attract students, it could be potentially very difficult to convince a competitor of SHU to relinquish its student data.

9.5.2 Extending the Study to Pre 1992 Universities

Again it would be interesting to find out if the patterns and trends found whilst undertaking this study would be repeatable at a pre-1992 university such as Sheffield University. However, as pointed above there could be difficulties in obtaining such data from a competitor of SHU.

9.5.3 National Student Assessment Data

Whilst there are benefits to creating a questionnaire that is tailored to the research being carried out. There are numerous difficulties associated with gathering student opinions from a large number of universities, such as location, poor response rates and data confidentiality. However, the National Student Assessment is conducted nationwide and could provide a rich source of data for mining. Whilst using such data would affect the direction of the research it could potentially provide a researcher with access to a large repository of student opinions.

9.5.4 HESA Data

There are national requirements on universities to submit data to HESA, such submissions include non-progressions and student award classifications. Therefore, as noted above, whilst the subject of any research would have to be tailored to fit the available data, this could provide a rich source of data for future studies.

9.6 Possible New Studies Involving Data Mining

There are also some other interesting studies, involving DM, that could be conducted outside of the HE domain. Indeed, such studies involve the mining of Further Education (FE) data and the data retained by UCAS. Each of these will be discussed separately below. Arguably, both 9.6.1 and 9.6.2 are a departure from the remit of this study. However, as the review of the literature and the exploration of the data highlighted, understanding such data could prove invaluable in determining the successful progression of students in HE.

9.6.1 Data Mining Further Education Data

The importance of students selecting the right university and course, at A-Level, has been highlighted in Chapter 3 as being vital to their progression at university (Yorke and Longden, 2008). However, the difficulties associated with making decisions about universities and courses, at such an early age and without the actual grades, have been identified in recent media reports (see Chapter 3). Therefore, it is believed that DM could help FE to predict job types, universities, university courses and A-Level results.

9.6.2 UCAS Data

In addition to mining data at the individual FE institution, there is also a vast wealth of data collected by UCAS, such data includes student demographics, students level 3 results, students choices of university and course, and there entry points. Again there are issues of data confidentiality as competition to attract students to universities will inevitably increase in the near future.

9.7 SUMMARY

In realising objective 6 – compile recommendations, this chapter sets out a number of recommendations based on the author's experience of carrying out this research project. These fall into the following five categories:

- Data Capture
- Use of DM
- Required Skill Level
- Further work involving DM at SHU
- Mining other available HE data.

From these thirteen recommendations were compiled, which will be recapped below. These can be grouped into SHU specific recommendations and recommendations for future studies. Hence the first set of recommendations, below, are for the institution to consider proceeding which are set of recommendations for future research in the area.

The SHU specific recommendations are to:

- improve the quality and completeness of the current data that they retain;
- increase staff awareness regarding the importance of good quality data;
- separate, where possible, dual values within variables and stop using the duplicate;
- improve transparency regarding the calculation of student entry points; and
- Investigate poor student response rates to DOL survey.

Future researchers wishing to further understand student behaviour in HE may wish to carry out the following.

Recommendations for future researches:

- further exploration of the SI data collected as part of this study;
- exploration of the data collected as part of an online survey;
- explore a larger data set of student records from SHU to validate the patterns identified as part of this research;
- mine data from another post-1992 university to validate the patterns identified in this research;
- mine data from a pre-1992 university to see if the patterns identified in this research are similar;
- mine data collected by HESA and from the National Student Assessment survey;
- explore how current changes to HE will affect future predictive models; and
- consider how data collected by FE institutions and UCSA could be used to predict student behaviour in HE.

Overall, it seems that SHU are already trying to improve the quality of their data with the introduction of a data warehouse and the review of the EDM literature has highlighted the potential benefits that DM could bring in helping to understand student behaviour in HE.

10 Reflective Summary

This chapter will attempt to evaluate the successes and areas of difficulty encountered whilst carrying out the research. This chapter will discuss where things could have been done differently, highlight potential missed opportunities and examine some areas of the research that never came to fruition. The above criteria will be used to examine the focus of the research along with the review of the literature, the uniqueness of the research, the research approach adopted, the building and understanding of the data set, mining the data, and the findings.

Throughout the process, every possible effort has been made to ensure that the direction of the research remained on course. However, whilst this may not have been possible at all times, the research did stay focused on its primary objectives set out in Chapter 2. The research question, aim and objectives were as follows:

Research Question

How can Business Intelligence be used to predict student behaviour as an aid to improving student progression?

Research Aim

Explore, through the application of BI tools, the issues that affect the progression of all undergraduate students at SHU. It is intended that a number of predictive models will also be constructed to predict student behaviour.

Research Objectives

No.	Objective	Measure
1.	Review, compare and contrast existing knowledge to develop a theoretical framework on which to base the rest of the study.	Completed literature review.
2.	Develop knowledge of the relevant SHU information systems and DM software to form an understanding of the underlying data structures and mining software.	Understanding of the student data and SAS® software through speaking to experts.
3.	Explore existing data sets, inductively, to build inferences and determine patterns in the data.	Reduced variables in the data set and the introduction of new variables through the iterative use of DM.
4.	Apply suitable DM techniques to build a number of predictive models.	Final models built and assessed.
5.	Validate the findings of study by comparing the results of the quantitative analysis to the current body of knowledge.	Completed findings.
6.	Compile a list of recommendations for the future uses of DM in this area based on the findings of the study.	Completed list of recommendations.

10.1 Focus of the Research

Firstly, there have been some significant changes in the HE landscape, brought about by a change of government, since starting this project in 2009. These relate to a reduction in student numbers and an increased emphasis placed on student led funding. As a result the models built as part of this study are likely to need revising in the future. However, building models that take into consideration the current situation wouldn't be possible for at least three to four years, post HE changes, as this will not be reflected in the universities data until the new students have graduated. In addition to this, since AAB and ABB students are outside the universities number cap, in 2012/13, it is likely that this has the potential for the post 1992 universities to attempt to recruit higher calibre students, which could have a positive outcome on final classifications and ultimately university rankings. Indeed, the initial exploration of the data (section 6.1.3.2.4) indicated that students with higher entry points tended to achieve better award classifications. This would have a positive effect on progression and university finances, as the loss of fees due to non-progression could reduce as more student's progress.

Furthermore, the change in the HE landscape will have a negative impact on the number of students, from non-traditional backgrounds, as there is likely to be less emphasis on the increasing and widening participation agenda going forwards. This is likely to change the nature of courses run by FE colleges who are now allowed to award degrees. Plus the large increases in student fees will increase the amount of student debt. However, student loans have been increased to meet the new tuition fees, which will put new students in much greater debt when they eventually leave university. This problem is further compounded by the current economic climate and a lack of jobs for young people (Coughlan 2012). The reduction in the number of university places may also lead to some students selecting there university in a hurry, which as discussed in section 3.2.3 could have a negative impact on student progression.

Although interviews and an online survey were conducted during the course of the research, it was decided that these were no longer relevant in the context of this study. As upon reflection there was no obvious connection to the DM exercise and these were also two large studies in their own right. Therefore, efforts were refocused so that more time could be spent on building the predictive models. It is anticipated that the results of the interviews and online survey will be used to form part of a follow up paper after the PhD has been completed.

10.2 Previous Research

The literature review helped to give a wider appreciation of student progression and EDM. The process was found to be fairly monotonous due to the large amount of material and the changing HE landscape, due mainly to the changes in student numbers and funding. This resulted in the review being one of the most time consuming to complete. A contributing factor to this was the vast amount of literature in the area. For this reason, it was decided to focus the student progression part of the review on literature taken from the mid-1990s onwards, this was mainly due to the changes taking place around student numbers, funding and student debt. It is possible that a conceptual model, such as a research territory map (see glossary page xi), that identifies relationships between topics may have helped to improve the management of the review (Dawson 2000). Even though the author has had past experience of writing such reviews, nothing could have prepared him for the amount of rework required in keeping the review up-to-date prior to handing in this thesis.

Arguably, there are areas of the literature review that could be improved. These relate to the inclusion of the section about the increasing and widening of student participation. Whilst this policy has been relaxed, it was decided to include this in section 3.2.1 as this was the policy at the time (2006) from when the data was drawn for the DM analysis. Furthermore, it could be argued that the NAO data used is quite old, taken from 2007 survey. However, this was the most update NAO survey at the time the thesis was completed. Indeed, a new NAO survey was completed close to the finalisation of this document. However, this was published too late to make the final version of this thesis.

With regards to the review on EDM and student progression in HE, the review found that the literature in this area was sparse. It could also be argued that the review of student progression and EDM should have also considered literature from FE as well as HE. Whilst some of the FE literature would have been relevant there are a number of differences between the two. In the main, these relate to the age of the students and the fact that they are, in the main, still at home living with family/carers and student fees are significantly less.

A further criticism is the length of the review which will undoubtedly affect its readability. Furthermore, it is fair to state that the review and the whole project places a large emphasis on material cited from Yorke and Longden. Whilst the author fully acknowledges this he would argue that it is only logical for the review and the project to have a considerable emphasis on these authors, as they are probably the biggest subject matter experts in the UK. Finally, the accuracy of some of the material used may be questionable as a large percentage of the material was gathered from trusted Internet and on-line journals, due in part to the changing nature of HE and the rapid growth of EDM. However, the comments of Dawson (2000) were taken into consideration and the review does reference some material that was sourced was from recognised authors, such as Yorke and Longden, Luan and the like.

10.3 Uniqueness of the Research

Having worked in a data analysis background for a number of years the author was keen to apply his knowledge and skills to the HE domain. He was also keen to develop new skills and apply previous knowledge gained whilst undertaking an MSc in Business Intelligence and SAS® training.

Whilst the area of EDM has enjoyed rapid growth since 2008/09, it is fair to state that there is almost a complete lack of literature that considers EDM and student progression in HE. This made it very difficult to put the research into context. However, previous research by Burley (2007), Luan (2001), Luan (2002), Luan (2004) and Luan (2006) have highlighted the applicability of DM in the HE context. Overall the DM process (including the data understanding phase) has helped to gain new insight into SHUs student data and ultimately the use of DM as a tool to predict student behaviour in the future, which further demonstrates the value of DM within HE.

10.4 Research Approach

This chapter had to be significantly reworked when it was decided to drop the acquisition of new data through interviews and questionnaires, see section 10.1 paragraph 3. However, the research approach was one of the most fulfilling and yet challenging chapters of the project. It involved a large amount of new learning and the theory discussed was found to be quite complex. On the whole the research approach adopted (Chapter 5) was found to be satisfactory in the gathering the data, building the DM marts and constructing the final predictive models and findings.

Indeed, the strong quantitative aspects of this research dictated the research strategy in terms of: the type of research, epistemology and ontology. Due to the inductive nature of the DM techniques applied and the relating of the results to previous literature the boundaries between theory and research are somewhat blurred. However, as pointed out by Bryman (2012, p614) "research methods are much more free floating than is sometimes supposed" and the distinctions made between the two approaches in section 5.1 are not as deterministic. Given the quantitative nature of the research, the cross sectional research design was found to be most appropriate. Obtaining a sample size of greater than 4,000 records helped to improve the reliability of the research as there was no reason to suppose that this cohort of students would be different from any other similar time period.

Furthermore, whilst the rules developed might not be transferrable to other time periods or institutions, the research is repeatable as the process followed is well documented. Indeed, all things being equal, anyone wishing to repeat this process would be able to use the descriptive framework and research sequence, outlined in Chapter 5, in conjunction with the further detail in Chapters 6 and 7. However, it is important to state that the results may be different as this is dependent on the techniques applied and skills of the individual carrying out the DM analysis. Moreover, the validity of results the results is substantiated by the fact that the variables and rules determined by the final models are visible in both the traditional student progression and the EDM literature. This research has also further highlighted the applicability of DM in the HE domain. Finally, the chosen DW approach (BDLD) and DM methodology, SAS® SEMMA, was successfully implemented in the context of this research.

10.5 Building and Understanding the Data Set

This stage of the research proved to be the most interesting and one of the most time consuming parts of the project. The creating of the three separate DM marts wasted a little time. Indeed it was identified, through the data understanding process, that this could have been done in a single table and that this would be more manageable. Additionally, Faculty and JACS subject had to be added to the data set after the ETL process as it was determined that Course variable had too many overall values. However, time was wasted through adding these extra variables as categorical variable consolidation (see glossary page vii), using decision trees, highlighted that the Course variable was indeed a better predictor of the target variable. Ultimately, this was due to the author's inexperience at carrying out DW and DM projects. As well as adding an extra dimension to the research, it is hoped that the addition of the DOL to the study will have added extra value for SHU.

Furthermore, it is important to remember that the data obtained from the SI database is a snapshot at a point in time. As a result there are still some blank values in the award classification (target) variable, this means that it wasn't possible to determine if the blanks (unknown) in the award classification variable are fails or if SHU hasn't recorded them yet. However, this was one of the reasons why the data was taken from a point in time where changes in HE were relatively stable and which afforded the best opportunity to obtain a more complete data set, the 2006 academic year. In addition to this, the postgraduate studies data set is limited in that it only considers students from 2007-2009 other students could have gone onto postgraduate studies at SHU after the snapshot was taken. However, the cut off had to happen somewhere and this was the latest data at the time of this study.

The inclusion of the location based POLAR2 (see glossary page x) data was useful as it confirmed that SHU does indeed cater more for the non-traditional type of students. In addition to this, POLAR2 data was also used as a predictor variable in the final models for two out of the three target variables. Further to this, the Entry Qualifications variable may have been considered at the modelling stage. However, it wasn't possible to split the GNVQ3/A-level qualification value. Ultimately, Entry Points was a better variable to use as other Entry Qualifications can be translated into comparable Entry Points. However, it would be help if SHU could make the process of mapping Entry Qualifications to Entry Points more transparent.

10.6 Data Mining

Having created a single table that would be utilised by all models, SAS® and the authors knowledge of HE was used to group certain values and reduce the number of variables. This proved a useful aid in determining how best to go about reducing and grouping the data set. In order to use some of the variable selection techniques, with SAS®, a number of dummy variables had to be created, which wasn't fully anticipated at the DW stage. Again this proved to be a pointless exercise as, in the end, categorical variable consolidation (see glossary page vii) and the decision trees were able to handle these groupings without any manual intervention.

Furthermore, it wasn't possible to rename the 'Node' variable, which was automatically created by SAS® when carrying out categorical variable consolidation (see glossary page vii), to something more meaningful. Therefore, the final decision trees had to be explained in greater detail, which undoubtedly increased the final word count for this chapter. Arguably, given that all models will use the same single data set, the data understanding section is a bit repetitive and it is difficult to determine the groups within each bar. However, the graphs presented in conjunction with the text in this document provide enough information for the reader to get a feel for the data. Indeed, the graphs have identified some interesting and valid information that could be classified as new knowledge.

After completing the data understanding stage of the DM, questions were raised regarding the number of events. The progression onto postgraduate studies at SHU modelling is a good example of this where prior probabilities and oversampling had to be applied to increase the 'Yes' event. This undoubtedly added an extra layer of complexity to the DM process. Ultimately, the final models produced from this are not very good predictors due to the sample size of 321 records (193 "Yes" and 128 "No" events).

10.7 Findings

The findings of the research are somewhat narrowed down by the fact that this investigation is using directed DM. Indeed, much more could have been found in the data if undirected DM had been applied to the data set. However, this wasn't possible as the research focused on predicting student behaviour in three defined areas:

1. award classification;
2. progression onto postgraduate studies at SHU; and
3. employment type.

Overall, the findings that have come out of this research are more than satisfactory. In that there is conformity with other research studies, the models developed are sensible and the issues with poor misclassification rates appear to be common in this type of research. Furthermore, the rules developed as part of this research are useful as they can be easily applied manually or coded to provide some type of instant benefit to SHU when trying to understand student behaviour. However, it is important to note that these models will have to be revalidated as the HE landscape changes.

It is however important to note that the findings are limited by the quality and amount of data received from SHU. Whilst a great deal of work was made to improving the quality of this data, it was impossible to invent data where data didn't exist, was missing or incomplete. In addition to this, due to SAS® grouping the course variable, through categorical variable consolidation (see glossary page vii), it was difficult to interpret the original rules. However, the original groupings were reiterated when discussing the final models in sub sections 7.1.1.6, 7.1.2.6 and 7.1.3.6.

Even though Course groupings were difficult to interpret the DM highlighted that some Courses were more difficult to achieve a high award classification than others, all other things being equal such as Entry Points, Entry Age *et cetera*. Additionally, many of the findings highlighted things that one might expect intuitively, for instance students on Courses with high employment prospects are less likely to go onto postgraduate studies at SHU. However, this further validates the applicability of DM in the HE context.

10.8 SUMMARY

This chapter set out to evaluate the success and difficulties faced whilst carrying out this research. It initiates by reiterating the research objectives set out in Chapter 2, these are then used to direct the assessment of this work in six key areas: previous research; uniqueness of the research; the research approach; building and understanding the data set; data mining; and the findings. The predictive models developed as part of this research study are more than satisfactory, but it is disappointing to note that these will have to be revalidated as the HE landscape changes. Indeed, the current changes brought about by the reforming of HE will have undoubtedly limited the longevity of any models developed as part of this research. Overall, the findings, from the DM and ultimately the research, further validate the applicability of DM in the HE context.

11 Conclusions

This research has used BI tools and techniques to predict student behaviour in three areas student award classification, progression onto postgraduate studies at SHU and employment type. An extensive review of the literature highlighted that the traditional literature around student progression has been well documented and that literature in the area of student progression and EDM was sparse. Data was obtained from SHU's SI database and a single DM mart was developed using DW processes. SAS® Enterprise Miner was later used to build three predictive models. This resulted in a number of interesting rules that could be linked to both the traditional and EDM student progression literature.

Having provided a brief overview of this research, above, this final Chapter will summarise the main findings and recommendations that were identified through carrying out this research.

11.1 Summary of Student Award Classification Results

The mining of the SI data identified a number of key variables that could be used to predict student award classification at SHU. These variables included: Course; Entry Points; Ethnicity; Undergraduate Entry Age; QYPR; and QAHE. Upon comparison to the literature discussed in Chapter 3, which are sensible predictors of the respective target variables. Indeed, these variables can be identified in traditional and EDM student progression research carried out by Moxley *et al.* (2001), Archer (2002), Superby *et al.* (2006), Herzog (2006) Yorke and Longden (2008), Dekker (2009) and in some of the policies implemented by the 1997 Labour Government, such as widening and increasing participation.

Constructing the models to predict award classification identified a number of rules which will be fundamental in helping both students and SHU to improve award classifications.

These included:

- Students are more likely to obtain a 2:1 on a business, management or marketing related course;
- Young (age on entry <23) white students who enter with less than 205 entry points onto certain health, property, design, biology, sports or engineering related courses are more likely to achieve a lower award classification;
- White students who enrol, with less than 205 entry points, onto food, health, business, sports, property, social care, education or film related courses tend to achieve a better award classification than non-white students;
- Students who enrol, with 300 or more entry points, onto food, health, business, sports, property, social care, education or film related courses tend to achieve a higher award classification;
- Students are less likely to achieve a high classification when studying certain health, sports, social work, English, film, computing or tourism related courses when entering with less than 300 entry points;
- Older (age on entry >=23) white students entering certain food, health, business, sports, property, social care, education or film related courses with high entry points are likely to achieve a high award classification;
- Older (age on entry >=23) white students with lower entry points, who study food, health, business, sports, property, social care, education or film related courses, tend to achieve better award classifications from areas where adult participation in HE is higher;
- Black African students with less than 300 entry points are more likely to achieve lower classifications on certain sports, education, business, management, health or property related courses than other students; and
- Students on food and nutrition, property development, science with education, business or media studies related courses are more likely to achieve a higher award classification.

11.2 Summary of Progression onto Postgraduate Studies at SHU Results

The results from mining this data identified that there were four key variables to predicting student progression onto postgraduate studies. These variables included: Undergraduate Course; Undergraduate Award Mark; QYPR; and Undergraduate Entry Points. Again as before these variables are sensible predictors when considering them in relation to the HE polices at the time, particularly widening and increasing participation, and the literature discussed in Chapter 3.

Through the process of building this model a number of interesting rules were identified. These could be used by the university to help to foster a better understanding of students who take Postgraduate studies at SHU and to focus marketing resources.

These rules were:

- Certain undergraduate courses have higher employability rates (such as law, education and computing (specialist)) and are therefore less likely to take postgraduate studies, at SHU, soon after completing a degree;
- Students are more likely to take postgraduate studies, at SHU, after studying undergraduate courses such as those associated with business, planning or psychology;
- Students who take business, finance, education QTS, design and technology or planning related undergraduate courses and achieve a low award mark are less likely to take postgraduate studies at SHU; and
- Students who achieve an award mark greater than 55 but entered HE with less than 300 entry points are more likely to take postgraduate studies (at SHU) when they have studied business, finance, education QTS, design and technology or planning related undergraduate courses.

11.3 Summary of Student Employment Type Results

The student employment type model built through carrying out the DM in Chapter 7 determined that four variables were key to predicting undergraduate student employment type. These variables were Course; Undergraduate Award Mark; Ethnicity; and Undergraduate Entry Age. In terms of the literature discussed in Chapter 3, the inclusion of these variables in the final model are sensible as these were identified in the literature and HE polices, such authors include Herzog (2006) and Yorke and Longden (2008).

Building the student employment type model highlighted a number of rules which will be useful to both students and SHU.

These are outlined below:

- There is either a problem with recording DOL data or students who study certain undergraduate courses and achieve low award marks tend not to respond to the DOL survey;
- Students are more likely to obtain a graduate job when they obtain a higher award mark in certain business, management, property, design, finance, sports, humanities, computing, art, social, criminology or education related courses;
- White students are more likely to obtain a graduate job when studying management, finance or business related courses;
- White students on engineering, food, computing (specialist), science, education, health, planning or design and technology related courses are more likely to end up in graduate roles even when achieving a lower award mark;
- Older (age on undergraduate entry >=43) students who achieve a lower award mark (<2:1), on computing, education, design, sports, art,

business, criminology, social, management, planning, English or humanities related courses are less likely to end up in employment; and

- Younger (age on undergraduate entry <19) students who achieve a lower award mark on certain computing (general), education, design, sports, art, business, criminology, social, management, planning, English or humanities related courses are less likely to get a graduate job.

11.4 Summary of Recommendations

Through carrying out this research a number of recommendations were made in Chapter 9. These recommendations are summarised below.

11.4.1 Data Capture

It transpired through the building of the DM mart that there were a number of issues with the SI data. These issues related to incorrect and incomplete data being retained by SHU. Therefore, it was recommended that SHU should improve the quality and completeness of the SI data by improving staff awareness regarding the importance of good quality data. It was also identified that dual value variables should be separated and that transparency regarding how student entry points are calculated should be improved. Furthermore, SHU should also investigate the large number of none responses to the DOL survey.

11.4.2 Further Exploration of the Collected Data

The recommendations fall into two categories, the further exploration of the SI data and the exploration of the data collected through an online survey.

1. Apply unsupervised DM techniques (such as clustering and market basket analysis), to determine patterns and trends in the DM mart developed in this study.
2. From the rules, outlined above, recommendations were made to:
    a. examine the link between ethnicity (mainly black) and the small number of these students who take postgraduate studies at SHU;
    b. investigate the association between ethnicity (mainly black) and the small number of these students who obtain a graduate role;
    c. carry out a similar study but include DLHE data so that this can be compared with other institutions;
    d. investigate the significance of the benefits of added value (see glossary page vii) and widening participation;
    e. examine the association between high demand skills that are in low supply and obtaining a graduate role;
    f. investigate the association between low demand skills that are in high supply and students progressing onto postgraduate studies at SHU; and

   g. obtain further insights into the association between older students and unemployment\low levels of employment.

3. Investigate the data collected from SI using unsupervised DM techniques.
4. Use the DM mart to create data cubes and report from them.
5. Investigate the data collected as part of an online survey using both directed and undirected DM techniques.

11.4.3 Capturing Additional Data from SHU

Repeat the study with a much larger amount of SHU level 6, postgraduate and DOL data (30,000 records) taken from other years, to determine if the patterns and trends are repeatable across other final years and student groups.

11.4.4 Considering Current Changes to the HE Landscape

Given the changes to the current HE landscape that has come about, due to austerity measures, in 2012. It would be interesting to repeat this study to determine what effect the cap on student numbers and increases in tuition fees has had on the models developed in this research.

11.4.5 Extending the Study to other Post 1992 Universities

It would be interesting to determine whether the patterns and trends found as part of this study were repeated at another post 1992 university, such as Nottingham Trent, Huddersfield or Leeds Metropolitan University.

11.4.6 Extending the Study to Pre 1992 Universities

Again it would be interesting to determine if there are any similarities, in the patterns and trends found as part of this study, between the post and pre 1992 universities, such as Sheffield University.

11.4.7 Mining Nationwide HE Data

Potentially there are two rich external sources of HE data that could be mined, using both undirected and directed DM techniques, to determine student behaviour. Such sources include:

1. The National Student Survey; and
2. Data collected by HESA.

11.4.8 Possible New Studies Involving Data Mining

There are also some other interesting studies, involving DM, that could be conducted outside of the HE domain. These include:

1. The exploration of FE data to help student select the right university and course whilst studying A-Levels.
2. The exploration of UCAS data to help determine university and course selection and potentially the type of student who are likely to enter clearing.

11.5 Post Script

Considering the recommendations formed from this study, there is still much research to be done in this area. The evolution of HE, in terms of its strong political association's and attendant policy revisions, will undoubtedly result in changes to future DM models. However, this study has laid the foundations for future research in this area.

The investigation has found associations between the variables identified through DM and research conducted by contemporary authors, such as Yorke and Longden (2008). It has also developed a number of rules that could be used to help institutions, students and future researchers to understand student behaviour.

Indeed, this research has further proved the applicability of DM in HE. The lessons learnt through the data cleansing and modelling process, outlined in this thesis, will be useful to other HE institutions who are interested in using BI to predict student behaviour. Furthermore, the DM results have raised some interesting questions that other institutions may wish to investigate. These include the significance of 'added value' and 'widening participation', and the small number of students from ethnic minorities who undertake postgraduate studies and obtain a graduate role.

Overall, a number of new insights have been identified that will prove valuable to both students and SHU, and will also add to the knowledge domain in this area. However, it must be emphasised that this study focuses on a specific subset of data from one university and it would be unwise to assume that the findings are fully transferable to other academic years and institutions.

12 REFERENCES

ARCHER, Louise (2002). *A Question of Motives*, Times Higher Education. [Online]. Last accessed 7 February 2013 at http://www.timeshighereducation.co.uk/story.asp?storyCode=169435&sectioncode=26

BAKER, Ryan SJD and YACEF, Kalina (no date). "*The State of Educational Data Mining in 2009: A Review and Future Visions"* [Online]. Last accessed 7 February 2013 at *http://www.educationaldatamining.org/JEDM/images/articles/vol1/issue1/JEDMVol1Issue1_BakerYacef.pdf*

BURLEY, Keith M and WILSON, Richard S (2012), *Understanding Student Progression for Data Mining Analysis*, HEIR, Presented at the Fifth Annual Conference of the Higher Education Institutional Research Network for the United Kingdom and Ireland.

BURLEY, Keith (2007). Data Mining Techniques in Higher Education Research: The Example of Student Retention, 29th Annual EAIR Forum, 26th to 29th August 2007, Innsbruck.

BURLEY, Keith (2006). Data Mining Techniques in Higher Education Research: The Example of Student Retention, Sheffield Hallam University, December 2006.

BURLEY, Keith (2003). *Data Mining Structures*. [lecture] Held 2003, Sheffield Hallam Universtiy.

BERRY, Micheal JA and LINOFF, Gordon S (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management,* 3rd Edition. USA, John Wiley & Sons, Inc.

BERRY, Micheal JA and LINOFF, Gordon S (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management,* 2nd Edition. USA, John Wiley & Sons, Inc.

BERSON, Alex, SMITH, Stephen J and THEARLING, Kurt (1999). *Building Data Mining Applications for CRM.* USA, McGraw-Hill.

BRYMAN, Alan (2012). *Social Research Methods*. 4th Edition. USA, Oxford University Press Inc.

CALLENDER, Claire (2001). Changing student finances in higher education: Policy contradictions under New Labour, *Journal of Widening Participation & Lifelong Leaning*, 3 (2), 5-15.

CHAPMAN, Pete, CLINTON, Julian, KERBER, Randy, KHABAZA, Thomas, REINARTZ, Thomas, SHEARER, Colin and WIRTH, Rudiger (2000). CRISP-DM 1.0 step-by-step data mining guide*, SPSS Inc 2000*, 1-78.

COHEN, Louis, MANION, Lawrene, and MORRISON, Keith (2011). *Research Methods in Education,* 7th Edition. Great Britain, RoutledgeFalmer.

COHEN, Louis, MANION, Lawrene, and MORRISON, Keith (2000). *Research Methods in Education,* 5th Edition. Great Britain, RoutledgeFalmer.

COUGHLAN, Sean (2012). *Graduates 'facing tougher times'*, BBC News. [Online]. Last accessed 7 February 2013 at http://www.bbc.co.uk/news/education-20237664

COUGHLAN, Sean (2009). *'No fee degree' university plan*, BBC News. [Online]. Last accessed 7 February 2013 at http://news.bbc.co.uk/1/hi/education/8139803.stm

CROWN (2011). *Students at the Heart of the System*, Department for Business Innovation and Skills, Crown Copyright. [Online]. Last accessed 4 December 2011 at http://c561635.r35.cf2.rackcdn.com/11-944-WP-students-at-heart.pdf

CRESWELL, John W (2009). *Research Design Qualitative, Quantitative, and Mixed Methods Approaches*. 3rd Edition. London, Sage.

DAVIES, Rhys and ELIAS, Peter (2002). *Dropping Out: A Study of Early Leavers from Higher Education,* Institute for Employment Research. Research Report Number 386, December 2002, 1-81.

DAWSON, Christian (2000). *The Essence of Computing Projects a Students Guide,* Essex, Pearson Education Limited.

DEKKER, Gerben, PECHENIZKIY, Mykola and VLEESHOUWERS, Jan (2009). *Predicting Students Drop Out: A Case Study*. [Online]. Last accessed 7 February 2013 at http://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf

EASTERBY-SMITH, Mark, THORPE, Richard and JACKSON, Paul (2012). *Management Research*. 4th Edition. London, Sage.

ECKERSON, Wayne (2004). *Four Ways To Build A Data Warehouse*, BI Best Practices. [Online]. Last accessed 13 November 2009 at http://www.bi-bestpractices.com/view-articles/4770

ENGLISH, Larry P (1999). *Improving Data Warehouse and Business Information*. USA, John Wiley & Sons Inc.

ERTL, Hurbert and WRIGHT, Susannah (2008). *Reviewing the literature on the student learning experience in higher education*, London Review of Education. 6 (3), 195-210.

GARNER, Richard (2008). *Average student debt now £4,500 a year*, independent.co.uk. [Online]. Last accessed 7 February 2013 at http://www.independent.co.uk/news/education/education-news/average-student-debt-now-1634500-a-year-892852.html

GEOGHEGAN, Ben. (2009). *Fears over student place shortage,* BBC News. [Online]. Last accessed 7 February 2013 at http://news.bbc.co.uk/1/hi/education/8133859.stm

GEORGES, Jim, THOMPSON, Jeff and WELLS, Chip (2010). *Applied Analytic Using SAS® Enterprise Miner$_{TM}$ 6.1 Course Notes".* USA, SAS Institute Inc.

GILL, John and JOHNSON, Phil (2010). *Research Methods for Managers,* 4th Edition. London, Sage.

GREENFIELD, Larry (2004). *An (Informal) Taxonomy Of Data Warehouse Data Errors*. [Online]. Last accessed 7 February 2013 at http://www.dwinfocenter.org/errors.html

HAND David, MANNILA, Heikki and SMYTH, Padhraic (2001). *Principles of Data Mining.* USA, Massachusetts Institute of Technology.

HEFCE (2012b). *Key Information Sets*. HEFCE. [Online]. Last accessed 7 February 2013 at http://www.hefce.ac.uk/whatwedo/lt/publicinfo/kis/

HEFCE (2012a). *About HEFCE*. HEFCE. [Online]. Last accessed 7 February 2013 at http://www.hefce.ac.uk/about/

HEFCE (2012). *Higher Education Funding Council For England*. HEFCE. [Online]. Last accessed 7 February 2013 at http://www.hefce.ac.uk/whatwedo/wp/ourresearch/polar/polar2/

HEFCE (2009). *HEFCE's funding policy on student completion.* HEFCE. [Online]. Last accessed 7 February 2013 at http://www.hefce.ac.uk/whatwedo/wp/currentworktowidenparticipation/studentretentionandsuccess/faqonstudentcompletion/

HERZOG, Serge (2006). Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-a-Vis Regression, *New Directions for Institutional Research.* Fall 2006. No 131, 17-33.

HESA (no date a). *Overview*, HESA. [Online]. Last accessed 7 February 2013 at *http://www.hesa.ac.uk/content/view/4/54/*

HESA. (no date). *Joint Academic Coding System (JACS) Version 3.0*, HESA. [Online]. Last accessed 7 February 2013 at http://www.hesa.ac.uk/content/view/1776/649/

INMON, Willaim H (2005). *Building the Data Warehouse*, 4th Edition. Indianapolis, Wiley Publishing Inc.

INMON, William H. (2002). *Building The Data Warehouse,* 3rd Edition. USA, John Wiley & Sons, Inc.

INMON, William .H. (1999). *Data Mart Does Not Equal Data Warehouse*, Information Management and Source Media. [Online]. Last accessed 7 February 2013 at http://www.information-management.com/infodirect/19991120/1675-1.html.

JANALTA INTERACTIVE Inc. (2013). *Structured Query Language (SQL)*. [Online]. Last accessed 7 February 2013 at http://www.techopedia.com/definition/1245/structured-query-language-sql

JOHN, George H and LANGLEY, Pat (1996). Static Versus Dynamic Sampling for Data Mining, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Menlo Park, California: AAAI Press, 367-370.

KEMBER, David (1995). *Open Learning Courses for Adults: A Model of Student Progress,* Englewood Cliffs, New Jersey, Education Technology Publications.

KENNEDY, H. (1997). *Learning Works: Widening Participation in Further Education*. Coventry, FEFC.

KIMBALL, Ralph, REEVES, Laura, ROSS, Margy and THORNTHWAITE, Warren (1998). *The Data Warehouse Lifecycle Toolkit: Expert Methods For Designing, Developing, And Deploying Data Warehouses*. USA, John Wiley & Sons, Inc.

KIMBALL, Ralph, ROSS, Margy, THORNTHWAITE, Warren, MUNDY, Joy and BECKER, Bob (2008). *The Data Warehouse Lifecycle Toolkit,* 2nd Edition. Indianapolis, Wiley Publishing Inc.

KIMBALL, Ralph and Ross, Margy (2002). *The Data Warehouse Toolkit,* 2nd Edition. USA, John Wiley & Sons Inc.

KIMBALL, Ralph. (1997). *A Dimensional Modelling Manifesto*. Miller Freeman Inc. [Online]. Last accessed 7 February 2013 at http://www.dbmsmag.com/9708d15.html

LUAN, Jing (2006). Impact of Tutoring on Student Success at Cabrillo College (Draft), *Planning, Research & Knowledge Systems*, Canada.

LUAN, Jing (2004). Data Mining Applications in Higher Education, *SPSS Incorporated Executive Report*. Chicago, SPSS.

LUAN, Jing. (2002). Data Mining and Knowledge Management in Higher Education, *Presentation at AIR Forum*. Canada.

LUAN, Jing. (2001). Data Mining as Driven by Knowledge Management in Higher Education – Persistence Clustering and Prediction. *Keynote for SPSS Public Conference*. University of California, San Francisco.

MARCO, David (2003). *A Meta-Data Repository Is The Key To Knowledge Management,* Enterprise Warehousing Solutions. [Online] Last accessed 7 February 2013 at http://www.tdan.com/i024fe02.htm

MARTINEZ, Paul (2001). *Improving student retention and achievement. What do we know and what do we need to find out?*. Learning and Skills Development Agency. [Online]. Last accessed 15 October 2011 at http://www.ulster.ac.uk/star/resources/lsda_report.pdf

MARTINEZ, Paul (1996). *Student Retention: case studies of strategies that work.* 1(6). Bristol,FEDA.

MARTINEZ, Paul (1995). *Student Retention in further and adult education*, Bristol, FEDA.

MCGIVNEY, Veronica. (2003). *Staying or Leaving the Course*, Leicester, NIACE.

MCGIVNEY, Veronica. (1996). *Staying or Leaving the Course*, Leicester, NIACE.

MILES, Matthew B and HUBERMAN, Michael A (1994). *Qualitative Data Analysis, An Expanded Sourcebook,* 2nd Edition. USA, Sage.

MOORE, Rebbeca (1995). *Retention Rates: Research Project*. Sheffield, Sheffield Hallam University.

MOXLEY, David, NAJOR-DURACK, Anwar and DUMBRIDGE, Cecille (2001) *Keeping Students in Higher Education.* London, Kogan.

NATIONAL AUDIT OFFICE (2007), *Staying the course: The retention of student in higher education,* National Audit Office. [Online]. Last accessed 7 February 2013 at http://web.nao.org.uk/search/search.aspx?Schema=&terms=Staying+the+course:+the+retention+of+students+in+higher+education

OATES, Tim and JENSEN David (1997). The Effects of Training Set Size on Decision Tree Complexity. *Machine Learning: Proceedings of the Fourteenth International Conference*. San Francisco, Morgan Kaufmann. 254-262.

O'DONNELL, Peter, ARNOTT, David, and GIBSON, Marcus (2002). *Data warehousing development methodologies: A comparative analysis.* (Working Paper. No. 2002/02). Monash University. Australia, Decision Support Systems Laboratory.

OPPENHEIM, Bram (1992). *Questionnaire Design, Interviewing and Attitude Measurement.* New Edition. Great Britain, Printer Publishers Ltd.

ORACLE (no date). *The Oracle E-Business Intelligence Enterprise Data Warehouse (EDW).* Oracle. [Online]. Last accessed 11 June 2011 at http://iprod.auc.dk/misg/courses/Administrative_IT_Systemer/2003/ssel/The%20Oracle%20E%20business%20Intelligence%20%20%20EDW%20%20A%20white%20paper.pdf

PARMENTIER, Philippe (1994). *La reussite des etudes universitaires :facteurs structurels et processuels de la peformance academique en premiere annee en medicine.* Ph.D. Faculte de Psychologies et des Sciences de l'Education, Catholic University of Louvain.

PEELO, Moira. and WAREHAM, Terry (2002). *Failing Students in Higher Education*. The Society for Research into Higher Education. Buckingham, Open University Press.

PITKETHLY, Anne and PROSSER, Micheal (2001). The First Year Experience Project: A Model for University Wide Change. *Higher Education Research and Development*. 20(2). 185-198.

RAMSDEN, Paul. (no date). *The Future of Higher Education Teaching and Student Experience*. [Online]. Last accessed 7 October 2010 at http://www.bis.gov.uk/assets/BISCore/corporate/docs/H/he-debate-ramsden.pdf

REED, Micheal (no date). *A Definition of Data Warehousing,* Technology Evalution.com. [Online]. Last accessed 7 October 2010 at http://www.intranetjournal.com/features/datawarehousing.html

RICHARDSON, Bill and RICHARDSON, Roy (1992). *Business Planning, An Approach To Strategic Management,* 2nd Edition. Great Britain, Pitman Publishing.

ROMERO, Cristobal, VENTURA, Sebastian, PECHENIZKIY, Mykola, and BAKER Rayn SJD (2011). *Handbook of Educational Data* Mining. USA,CRC Press.

ROMERO, Cristobal, VENTURA, Sebastian, ESPEJO, Pedro G and HERVAS, César (2008). *Data Mining Algorithms to Classify Students.* [Online] Last accessed 7 February 2013 at http://sci2s.ugr.es/keel/pdf/specific/congreso/Data%20Mining%20Algorithms%20to%20Classify%20Students.pdf

ROMERO, Cristobal and VENTURA, Sebastian (2006). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Application*. 33. 135-146.

ROUSE, M. (2005). *Entity-Relationship Model (ERM or ER model)*, TechTarget. [Online] Last accessed 7 February 2013 at http://searchsqlserver.techtarget.com/definition/entity-relationship-model

SAMLI, AC POHLEN, TL and BOZOVIC, N (2002). A Review of Data Mining Techniques as they Apply to Marketing: Generating Strategic Information to Develop Market Segments. *The Marketing Review 2002*. 3. 211-227.

SAS (2013). *Usage Note 24205: Rare event oversampling for model fitting in SAS® Enterprise Miner(tm)*, SAS Institute Inc. [Online] Last accessed 7 February 2013 at http://support.sas.com/kb/24/205.html

SAS (no date). *Enterprise Miner SEMMA*, SAS Institute Inc. [Online] Last accessed 7 February 2013 at http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html

SAS INSTITUTE Inc. (2001), *SAS Rapid Data Warehousing Methodology*, SAS Institute Inc 2001. 1-14.

SCHWANDT, Thomas A (1997), *Qualitative Inquiry, a Dictionary of Terms*. USA, Sage Publications, Inc.

SCHWATZ, KD. (1996). *Data Warehousing: in Search of the Whole Story*. Enterprise Reengineering, August 1996. [Online] Last accessed 8 October 2010 at http://www.defenselink.mil/nii/bpr/bprcd/5776.htm

SHEFFIELD HALLAM UNIVERSITY. (no date), *International Standing*. [Online] Last accessed 7 February 2013 at http://www.shu.ac.uk/university/overview/international/

SINGH, Harry S (1998). *Data Warehousing: Concepts, Technologies, Implementations, and Management*. USA, Prentice Hall.

SLACK, Kim and CASEY, Lorraine (2002). *The Best Years of Your Life? Contrasting the Local and the Non-Local Student Experience of Higher Education*. Stoke on Trent, BERA Staffordshire University Press.

SPELLINGS, Margaret (2006). *A Test of Leadership: Charting the Future of U.S. Higher Education*. U.S. Department of Education Report. [Online] Last accessed 7 February 2013 at http://www.naspaa.org/accreditation/standard2009/docs/SpellingsCommissionReport09.2006.pdf

SUPERBY, J -F, VANDAMME, Jean-Philippe and MESKENS, Nadine (2006). *Determination of Factors Influencing the Achievement of the First-Year University Students using Data Mining Methods*. . [Online] Last accessed 7 February 2013 at http://www.educationaldatamining.org/ITS2006EDM/superby.pdf

THE QUALITY ASSURANCE AGENCY FOR HIGHER EDUCATION (2012). *About Us,* The Quality Assurance Agency for Higher Education. [Online] Last accessed 7 February 2013 at http://www.qaa.ac.uk/AboutUs/Pages/default.aspx

THOMANN, James and WELLS, David L (no date). Evaluating Data Warehousing Methodologies: Objectives and Criteria. *The Journal of Data Warehousing*. [Online] Last accessed 13 October 2010 at http://www.infocentric.org/articles/dw_methodology_1of3.PDF

THOMAS, Liz (2002). Student Retention in Higher Education: The Role of Institutional Habitus. *Journal of Educational Policy*. 17(4), 423-442.

TINTO, Vincent (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition. 2nd ed.*Chicago, University of Chicago Press.

TINTO, Vincent (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research.*Review of Education Research*, Winter 1975. 45(1). 89-125.

TWO CROWS (no date). *Data Mining Glossary*, Two Crows. [Online] Last accessed 7 January 2010 at http://www.twocrows.com/glossary.htm

TWO CROWS CORPORATION (1999). *Introduction to data mining and knowledge discovery,* 3rd Edition, USA, Two Crows Corporation.

UCAS (no date) *What is higher education?*. UCAS. [Online] Last accessed 7 February 2013 at http://www.ucas.ac.uk/students/wheretostart/heexplained/

UNIVERSITY BUSINESS. (2004). *University of Alabama uses SAS to Identify, Mentor at-Risk Students and Advance Student Retention*. [Online] Last accessed 31 May 2009 at http://findarticles.com/p/articles/mi_m0LSH/is_9_7/ai_n6187110

UPTON, Graham and COOK, Ian (2002). *A Dictionary Of Statistics*. Oxford University Press, 2002. [Online]. Last accessed 2 October 2009 at http://www.oxfordreference.com/views/BOOK_SEARCH.html?book=t106&subject=s23&authstatuscode=202.

WETHERILL, Barrie G (1986). *Regression Analysis with Applications*. Great Britain, Chapman and Hall Ltd.

YORKE, Mantz (1999). *Leaving Early*. London, Falmer.

YORKE, Mantz and LONGDEN, Bernard (2008). *Retention and Student Success in Higher Education*. Glasgow, Bell & Bain Ltd.

YORKE, Mantz and LONGDEN, Bernard (2008b). *The first-year experience of higher education in the UK*. The Higher Education Academy. January 2008.

YORKE, Mantz and LONGDEN, Bernard (2004). *Retention and Student Success in Higher Education*. The Society for Research into Higher Education. Maidenhead, Open University Press.

YORKE, Mantz and THOMAS, Liz (2003). Improving the Retention of Students from Lower Socio-economic Groups. *Journal of Higher Education and Policy and Management*. 25(1). 63-74.

13 BIBLIOGRAPHY

BERRY, Micheal JA and LINOFF, Gordon S (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*. USA, John Wiley & Sons, Inc.

BURLEY, Keith. (2008). *Tutor Intervention and Student Progression Myth or Reality*. 30th EAIR Forum, 24 August to 27 August 2008, Copenhagen.

CHUO-HAN, Lee. (2004), *MOLAP, ROLAP, and HOLAP*. [Online] Last accessed 7 February 2013 at http://www.1keydata.com/datawarehousing/molap-rolap.html

DAWSON, Christian (2009). *Projects in Computing and Information Systems A Student's Guide*. 2nd Edition, Essex, Pearson Education Limited.

GLASER, Barney G and STRAUSS, Anslem L(1967). *The Discovery Of Grounded Theory: Strategies for Qualitative Research*. USA, Aldine Publishing Company.

JARKE, Matthias, LENZERINI, Maurizio, VASSILIOU, Yannis. and VASSILIADIS, Panos (2000). *Fundamentals Of Data Warehouses*. 2nd Edition. New York, Springer.

REH, John F (no date). *Key Performance Indicators*. About Inc. [Online] Last accessed 7 February 2013 at http://management.about.com/cs/generalmanagement/a/keyperfindic.htm

JOHNES, Jill (1990). Determinants of Student Wastage in Higher Education. *Studies in Higher Education*, 15(1). 87-100.

KIMBALL, Ralph and BECKER, Bob (2007). *Think Critically When Applying Best Practices*. [Online] Last accessed 7 February 2013 at http://www.intelligententerprise.com/showArticle.jhtml?articleID=198700049

KIMBALL, Ralph and CASERTA, Joe (2004). *The Data Warehouse ETL Toolkit*. USA, Wiley Publishing Inc.

PECHENIZKIY, Mykola, CALDERS, Toon, VASILYEVA, Ekaterina, DE BRA, Paulv(2008). Mining the Student Assessment Data: Lessons Drawn from Small Scale Case Study. *Proceeding of the 1st International conference on Educational Data Mining (EDM'08)*. 187-191.

RICHARDSON, Hannah (2011). *Post-Result University Admission Urged,* BBC News. [Online] Last accessed 7 February 2013 at http://www.bbc.co.uk/news/education-15492470?print=true

RUGG, Gordon. and PETRE, Marian. (2004). *The Unwritten Rules of PhD Research*. Open University Press. McGraw-Hill Education, England.

SALANT, Priscilla and DILLMAN, Don A (1994) *How to Conduct Your Own Survey*. Canada, John Wiley & Sons,Inc.

SMITH, Mark J (2003). *Social Science in Question*, London, Sage.

THOMANN, James and WELLS, David L (no date). The Keys to the Data Warehouse. *The Journal of Data Warehousing*. [Online] Last accessed 13 October 2010 at http://www.infocentric.org/articles/keys_to_dw.PDF

WATTERSON, K. (1998). *Warehousing,* SunExpert Magazine. October 1998. 58-65.

YIN, Robert K (2009). *Case Study Research Design and Methods,* 4th Edition, USA: Sage Publications, Inc.

YIN, Robert K (2003). *Case Study Research Design and Methods,* 3rd Edition. USA, Sage Publications, Inc.

YIN, Robert K (1989). *Case Study Research Design and Methods*. Revised Edition. USA, Sage Publications, Inc.

Word Count: 48,164

14 APPENDICES

APPENDIX I – FACULTY DEPARTMENTS

ACES
Art & Design
Computing
Engineering & Mathematics
Media Arts and Communication

D&S
Humanities
Teacher Education
Education, Childhood and Inclusion
Architecture & Planning
Built Environment
Law & Criminology
Psychology, Sociology & Politics

HWB
Biosciences
Allied Health Professionals
Social Work
Sport
Nursing & Midwifery

SBS
Leisure & Food Management
Management
Accounting, Finance & Operation Systems

APPENDIX II – JACS GROUPED SUBJECT LIST

Missing
BUSINESS STUDIES
SOFTWARE ENGINEERING
SPORTS SCIENCE
LAW BY AREA
PSYCHOLOGY
COMPUTATIONAL SCIENCE FOUNDATION
SOCIAL POLICY
OTHERS IN EDUCATION
ENGLISH STUDIES
ACCOUNTING
PHYSIOTHERAPY
BIOLOGY
TOURISM TRANSPORT AND TRAVEL
ACADEMIC STUDIES IN NURSERY EDUCATION
CINEMATICS AND PHOTOGRAPHY NOT E
DESIGN STUDIES
HISTORY BY PERIOD
FINE ART
OTHERS IN MASS COMMUNICATIONS AN
INTERNATIONAL BUSINESS STUDIES
EVENT MANAGEMENT
ARCHITECTURE
MEDIA STUDIES
CINEMATICS AND PHOTOGRAPHY
FOOD AND BEVERAGE TECHNOLOGY
URBAN STUDIES
CHEMISTRY
GENERAL ENGINEERING
OTHERS IN TECHNOLOGY
HUMAN AND SOCIAL GEOGRAPHY
BUILDING SURVEYING
INDUSTRIAL/PRODUCT DESIGN
MATHEMATICS
APPLIED SOCIOLOGY
SOCIOLOGY
LAW BY TOPIC
SOCIAL WORK
TELECOMMUNICATIONS ENGINEERING
ACADEMIC STUDIES IN EDUCATION
ADULT NURSING
NUTRITION
OTHERS IN SOCIAL STUDIES

PRODUCTION AND MANUFACTURING ENG
RADIOGRAPHY DIAGNOSTIC
OCCUPATIONAL THERAPY
MECHANICAL ENGINEERING
MULTIMEDIA DESIGN
AUTOMOTIVE ENGINEERING
APPLIED STATISTICS, BANKING, COMMUNICATIONS ENGINEERING, COMMUNITY NURSING, COMPUTER-AIDED ENGINEERING, CONSTRUCTION MANAGEMENT, ELECTRONIC AND ELECTRICAL ENGINE, ELECTRONIC ENGINEERING,

APPENDIX III – MEETING NOTES

This Appendix provides further information about two meetings that took place between the author and the Information Department, at Sheffield Hallam University, to obtain data from the Student Information Services database. In order to protect the identities of the individuals involved in this meeting they are referred to as Person A and Person B.

Date	Attendees	Notes
15/12/2009	Person A Author	Purpose introductions and outline the research and find out what data is available and how it is currently used. • Many fields available within the SIS database; o The SIS is a type of SQL relational database; o Data is extracted by the University Systems Group; o Speak to service manager for MIPI (Ext 2641) about getting data; o Currently you can access data on recent cohorts, social class, ethnicity , disability, postcodes (term time and home address), previous education (A-levels, GCSEs etc.), previous school (FE college, 6th form etc.) and school name; o Also contains a SIS status flag (enrolled, withdrawn, transferred etc.) o Date of withdrawal in year o Potential withdrawal academic failure • It is possible to link the SIS database to other systems to produce a fuller data set, these include: o Data from UCAS – the data from the application process is stored and successful students are migrated over into the SIS database (this could be used to access data on social class, previous grades etc.); o Student Internal Survey (gathering opinions of the students – this is around the student experience) o Destination of leavers – careers database • Student code is the key in linking all the different databases • HEFCE POLAR – mapped the country down to electoral ward (quartiles) – SNAC files • Might be worth looking at full programmes of study as it's not clear what happens to some of the students • Data is very dirty • Person A mentioned that it might be worth noting that assessment methods are different across different subject areas, which may cause problems when predicting award classification. • Person A is currently writing a paper for the management of the university which details the fields from the SIS database that she believes to be an important indicator for retention – she has agreed to share this with me. • Person B works for Person A and knows a lot about the database • Arrange a further meeting with Person A and Person B after the New Year. • It also seems that there is currently no Data Warehouse in place and that no Data Mining is been carried out. The majority of the analysis is statistical in terms of averages etc. However, they are planning to carry out some text mining on some qualitative data that they have recently received from the student internal survey.

Date	Attendees	Notes
02/02/2010	Person A Person B Author	I will need to seek approval for the online questionnaire as there is a whole process to sending out a questionnaire at SHU. Need to obtain approval from ACES ethical committee for my research. It is possible to determine what the students went on to do from the Destination of Leavers data, there is about 5 years' worth of data here. There is approximately 4 years' worth of assessment data, which includes: • overall classification • measures of performance • module marks Person A suggested that I might wish to look at student employability as the university are interested in this There are four different ways to calculate a students awards. Data items discussed • Gender • Disability • Ethnicity • Age • Local Education Authority • Home postcode • Nationality • Fee status • Socio Economic Group • Tariff Points – entry points • Entry qualifications – highest education obtained before entering university A-levels etc. • Age on entry • Type of Study – UG, PG • Course • Faculty • University Entry Date • SCE – Start Date (academic enrolment date) • SCJ – Start Date (student registration on the course) • SOC – Parents standard occupation code Person A suggested: • HEFCE carried out some work looking at the levels of attainment on entry qualifications and age. • MS – postgraduate student perceptions of the university Person B will provide me with data for full time L6 undergraduates from 2006/07 including demographics, modules, entry qualifications, DOL and postgraduates ID's and course name from 2007 onwards.