Forecasting and trading Bitcoin with machine learning techniques and a hybrid volatility/sentiment leverage

WEI, Mingzhe <http://orcid.org/0000-0002-8817-7788>, SERMPINIS, Georgios and STASINAKIS, Charalampos

Published version

WEI, Mingzhe, SERMPINIS, Georgios and STASINAKIS, Charalampos (2022). Forecasting and trading Bitcoin with machine learning techniques and a hybrid volatility/sentiment leverage. Journal of Forecasting.

Online Appendix

OA.1 Latent Dirichlet Allocation (LDA)

Developed by Blei et al. (2003), LDA is one of the most prevalent algorithms in topic modelling area. As a generative probabilistic model, LDA is used to identify latent topics in a large corpus of texture where each topic is characterized by a distribution over words. Given a corpus organized by D documents with T topic where each document d has N_i $(i \in 1, \dots, N)$words, we apply LDA algorithm to C in the following generative process:

1. For each topic t $(t \in 1, \dots, T)$ we draw a Dirichlet distribution over words $\beta_t^{iid} \sim Dirchilet\ (\eta)$, which is the probability of a word in topic t and η denotes the hyperparameter for prior distribution of β_i.
2. For each document d $(d \in 1, \dots, D)$, we draw a Dirichlet distribution over topics $\theta_d^{iid} \sim Dirchilet\ (\alpha)$,, indicating the distribution on the topics for document d. α denotes the hyperparameter for prior distribution of θ_d.
3. For each word w_{di} in document d, we have $d \in 1, \dots, D$,
   i. Choose a topic from $z_{di} \sim Multinomial(\theta_d)$, where z_{di} denote the topic from which w_{di} is drawn.
   ii. Choose an observed word from $w_{di} \sim Multinomial(\beta_{z_{di}})$,

In the above process, only words are observed variables and the rest parameters are latent variables. The probability of observed data is obtained as the product of marginal probability as follows:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \quad \text{(OA.1)}$$

OA.2 Implementation of LDA and relevant results

Consistent with previous studies in NLP, we use nltk Python library (Joakim, 2012) to preprocess the data, including converting words to lower cases, removing special characters (e.g., horizontal Tab, space, and comma) and stopping words (e.g., me, you, and I), stemming (tracking back the root of words, e.g., stopping back to stop) and transforming the cleaned corpus into term-document matrix. As pointed by Chen and Doss (2019), it is difficult to specify an optimal topic number using LDA in advance, we thus use 10 as the number of topics. Similar to Azqueta-Gavaldón (2020), we want to give an interpretable picture so that fewer topics can give a more concise result. Another reason is because we do not seek for specific topic tasks but to retrieve the sentiment scores, therefore 10 topics are adequate.

Table OA.1 displays ten topics produced by LDA. As suggested by Larsen and Thorsrud (2019), LDA is an unsupervised learning algorithm, and it does not generate labels through the computation procedure. In order to explicitly demonstrate results, we subjectively label each topic based on our understanding. Unsurprisingly, the largest topic corresponding to our corpus is about Finance and Economy, taking up 35.9% of all cryptocurrency narratives. Technology, the second largest topic contains words associated

with technical development, such as *digit, blockchain, future, system, data*, etc. Media and Politics are the next two topics, having 12.1% and 9.7% proportions, respectively. For Media, we have words like *press, report, journal, magazine* and Politics has words associated with politicians and governments, such as *trump, nation, regulator.* In the end, we have three topics, including Crim (7.97%), Accountancy (6,58%) and Corporation (6.44%). Additional topics are not worth mentioning, since related narratives are not the main targets in this study.

[Insert Table OA.1]

OA.3 RFE-RF Process

Granitto et al. (2006) and Darst, Malecki & Engelman (2018) suggest the combination of RFE and RF to perform feature selection through the iteration of model training, feature ranking and eliminating the ranked features below threshold. Similar evidence is also found in the occasion of correlated features (Gregorutti, Michel and Saint-Pierre, 2017). Generally, the process of RFE-RF in feature screening can be described as follows:

For target series y and a set of features x_m (m denotes the number of features and m = 1,2, …, m), RF algorithm is fit and exclude the features under the importance threshold (e.g., the least 25% important features are dropped). After the first round of model fit with RF, we obtain the reduced set of features x_k^1 (k < m) and fit with RF algorithm for the second time. After t-th iteration, we have the further reduced subset of x_j^t (j < k) features. In the manner of feature elimination, a user-specified stopping criterion is set so that necessary rounds of selection or a certain number of rules can be defined at start.

OA.4 Hyperparameter Optimization and statistical/profitability metrics

Due to the complexity of most ML algorithms, tuning hyperparameters is a crucial step to convincible results. In this study, we apply grid search which is one of the most popular approach in terms of hyperparameters search methods. Grid search method is designed to find the optimal value by exhaustively searching in a specified hyperparameters space. In Table OA.2, we give brief introduction and the optimal value of our critical hyperparameters used in GBDT family mode.

[Insert Table OA.2]

Finally, the following table provides the statistical and profitability metrics used in this study.

[Insert Table OA.3]

OA.5 Factor Importance Ranking and hybrid PCA leverages

From both forecasting ability and trading performance, we find GBDT family has better performance than other models. This encourages us to further explore the significance of input factors in our best model. Figures OA.1 and OA.2 provide the top 10 features' (RFE-RF factors) contribution to the building of XGB and LBM, respectively (from the most important to the least important) in F1 forecasting exercise period.

[Insert Figures OA.1-OA.2]

[Insert Figures OA.3-OA.4]

OA.3 and OA.4 illustrate the top 10 features' contribution of PCA factors. Detailed explanation of importance scores can be referred to (Elith, Leathwick & Hastie, 2008; Hastie, Tibshirani & Friedman, 2009). Unlike other ML algorithms, GBDT family is good at interpretation of feature selection by retrieving important score during tree construction. The highest score indicates that specific attribute is used most frequently in tree split. From Figures OA.1 and OA.2, PMA ratios outperform other selected factors and occupy a large proportion in the construction of GBDT family. That is, PMA ratios make the greatest contribution to the forecasting accuracy of our model.

Finally, for consistency with the main text we present the assigned leverages for the hybrid strategy based on the PCA approach.

[Insert Figures OA.5-OA.7]

Online Appendix – Tables

Table OA.1 Summary of overall LDA results

Topic	Label	Score	Percentage (%)	Words
1	Finance and Economy (I)	0.038	21.4	bitcoin, financial, company, bond, business, bank, currency, money, market, capital, investors, payments, transaction, industry, dollar, security, exchange, trading, investment, deal
2	Finance and Economy (II)	0.040	14.5	mt, gox, people, cash, wsj, stock, earnings, fund, firm, customer, oil, on line, asset, buy, growth, price, gold, federal, economy, sale
3	Technology	0.109	15.3	digit, ethereum, blockchain, system, document, coin, technology, future, times, cryptocurrency, data, government, platform, global, virtual, tech, online, reserve, update
4	Media	0.027	12.1	press, state, including, public, percent, chief, media, president, label, low, forward, report, day, power, right, journal, read, real, fell, magazine
5	Politics	0.029	9.70	trump, government, America, China, UK, tax, power, rise, European, nation, English, north, inflation, Korea, top, move, base, regulator, service
6	Crime	0.009	7.97	crucial, position, stan, standout, poker, giant, cop, approximate, telegraph, illegal, credit, operative, knight, group, garage, hostage, terminate, wall, court, drug
7	Accountancy	0.010	6.58	ledger, information, atmosphere, statement, tax, load, decentralize, distance, stall, carrie, week, year, puzzle, ring, sign, inception, ltd, number, story, version, men
8	Corporation	0.022	6.44	adopt, book, start, conference, web, piece, ltd, centre, mail, wide, kind, electronic, body, road, tip, sense, entrepreneur, city, origination, team
9	Tech-related	0.033	3.27	ebay, yahoo, web, publication, artist, crowdfund, minute, let, accountant, ross
10	Unidentified	0.031	2.75	poker, bowl, forest, high, gmt, copyright, cryptographic, anyone, ny, vs

Note: This table reports the key words extracted for each topic based on LDA algorithm. For the corpus, Score denotes the sentiment score (polarity) of each topic and percentage (%) denotes the proportion of each topic. All the country names should in lower case (we use capital letters for a better view).

Table OA.2 Main hyperparameters and optimal value of GBDT family models

XGB model							
Hyperparameters	**Symbol**	**RFE(F1)**	**RFE(F2)**	**RFE(F3)**	**PCA(F1)**	**PCA(F2)**	**PCA(F3)**
Number of boosted trees to fit	N_estimators	500	500	500	500	500	500
Evaluation metrics	Eval_metrics	rmse	rmse	rmse	rmse	rmse	rmse
Subsample ratio of training data	Subsample	1	1	1	1	1	1
Minimum sum of data weight needed in a child	Min_child_weight	1	1	1	1	1	1
Maximum depth of a tree	Max_depth_	4	6	6	3	6	4
Minimum loss reduction required to make a further partition of a leaf node of the tree	Gamma	0	0	0	0	0	0
Subsample ratio of columns for construction of each tree	Colsample_bytree	1	1	1	1	1	1
L1 regularization term on weights	Alpha	0	0	0.01	0	0.01	0
L2 regularization term on weights	Lambda	0	0	0	0	0	0
Learning rate	Eta	0.041	0.045	0.024	0.046	0.055	0.050
LightGBM model							
Hyperparameter	**Symbol**	**RFE(F1)**	**RFE(F2)**	**RFE(F3)**	**PCA(F1)**	**PCA(F2)**	**PCA(F3)**
Number of boosting iterations	N_estimators	500	500	500	500	500	500
Maximum depth of a tree	Max_depth	6	3	6	7	4	8
Maximum number of leaves in one tree	Num_leaves	32	16	32	32	16	64
Subsample ratio of instances for construction of each tree	Colsample_bytree	1	1	1	1	1	1
L1 regularization term on weights	Reg_alpha	0.01	0	0	0.01	0.01	0
L2 regularization term on weights	Reg_lambda	0	0	0	0	0	0
Learning rate	Learning_rate	0.045	0.039	0.041	0.055	0.051	0.047

Note: *This table reports hyperparameters of XGB model and LBM model used in this study.*

Table OA.3 Statistical and Profitability Measurement and Formula

Statistical Measurements	Formula
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{\tau=t+1}^{t+n}\left\|\hat{Y}_\tau - Y_\tau\right\|$
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{\tau=t+1}^{t+n}(\hat{Y}_\tau - Y_\tau)^2}$
Mean Squared Error (MSE)	$\frac{1}{n}\sum_{\tau=t+1}^{t+n}(\hat{Y}_\tau - Y_\tau)^2$
Profitability Measurements	**Formula**
Annualized Return (R^A)	$R^A = 365 * \frac{1}{N} * (\sum_{t=1}^{N} R_t)$
Sharpe ratio (SR)	$SR = \frac{R^A - r_f^A}{\sigma^A}$
Maximum drawdown (MDD)	$Max(MD) where MD = \underset{i=1,\ldots,t;t=1,\ldots,N}{Min}(\sum_{j=i}^{t} R_j)$
Information ratio (IR)	$IR = \frac{R^A}{\sigma^A}$
Sortino ratio (SOR)	$SOR = \frac{R^A - r_f^A}{\sigma^D}$

Notes: *This table reports formulas of measurements used in this study.* $\hat{Y}_\tau$ *denotes the forecast values,* Y_τ *denotes the actual values,* σ^A *denotes the standard deviation and* σ^D *denotes the downside risk. In this paper, we apply 0.327% as risk-free rate, corresponding the 10-year treasury bond yield of US.*

Online Appendix - Figures

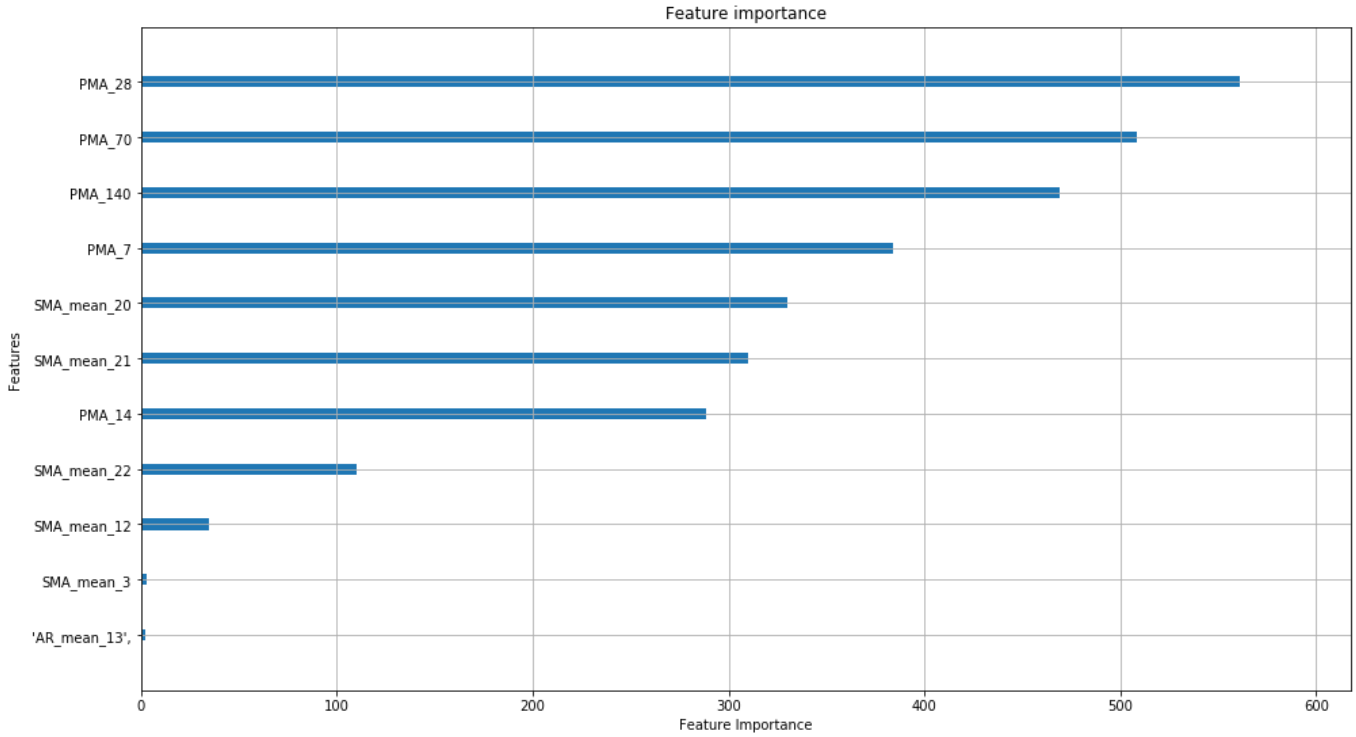


Figure OA.1 Top 10 features' contribution of RFE-RF factors in the construction of XGB. **Note:** Factors are named by their models with specific parameter. For example, PMA_28 denotes the 28 lag of PMA ratio and SMA_mean_3 denotes the simple moving average model with parameter 3.

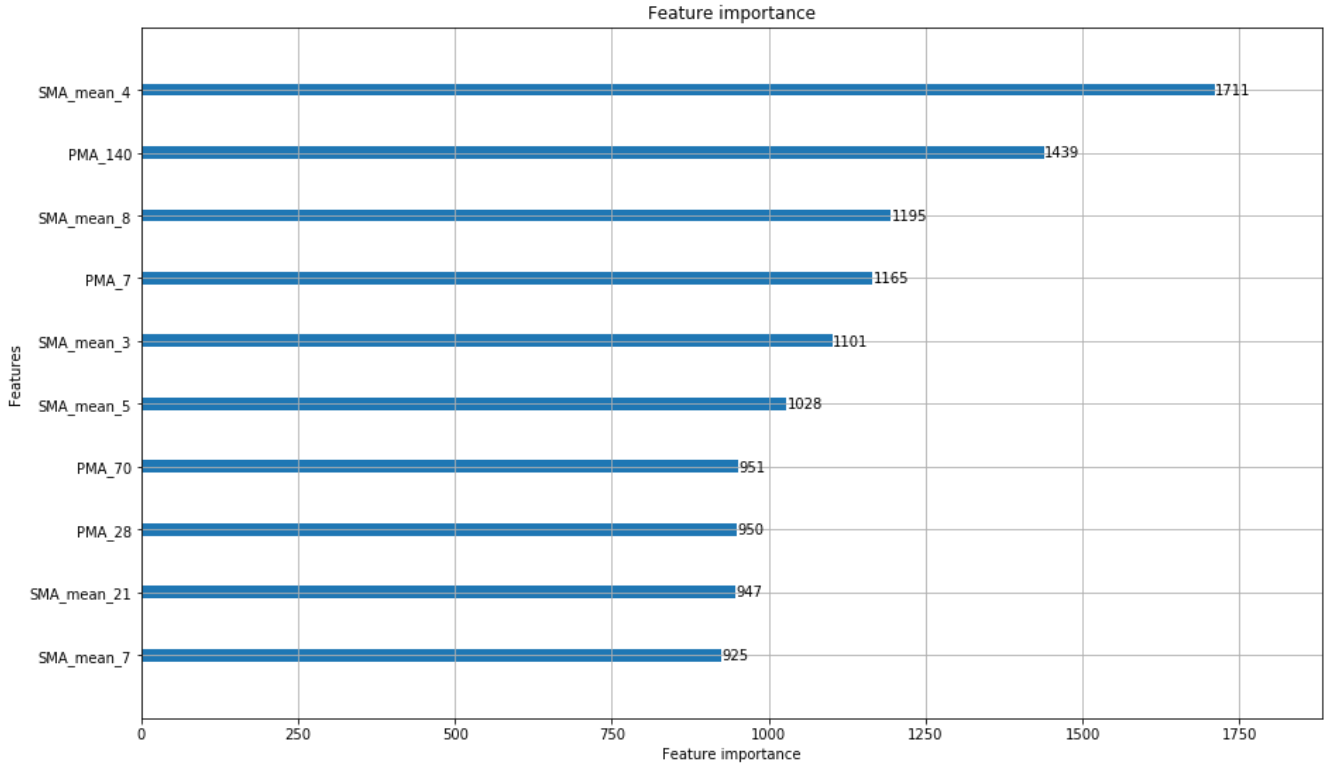


Figure OA.2 Top 10 features' contribution of RFE-RF factors in the construction of LBM. **Note:** Factors are named by their models with specific parameter. For example, PMA_28 denotes the 28 lagged days of PMA ratio and SMA_mean_3 denotes the simple moving average model with parameter 3.

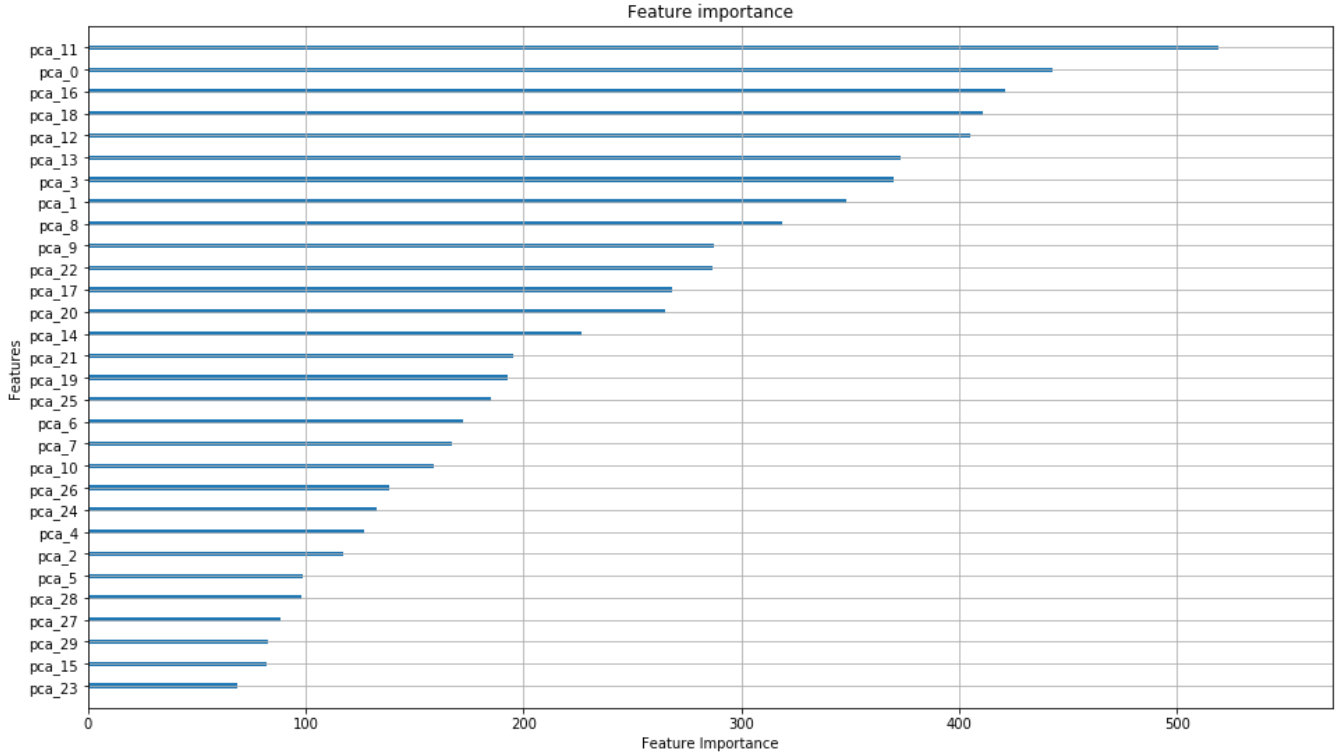


Figure OA.3 Top 10 features' contribution of PCA factors in the construction of XGB. Note: Factors reported are all principal components.

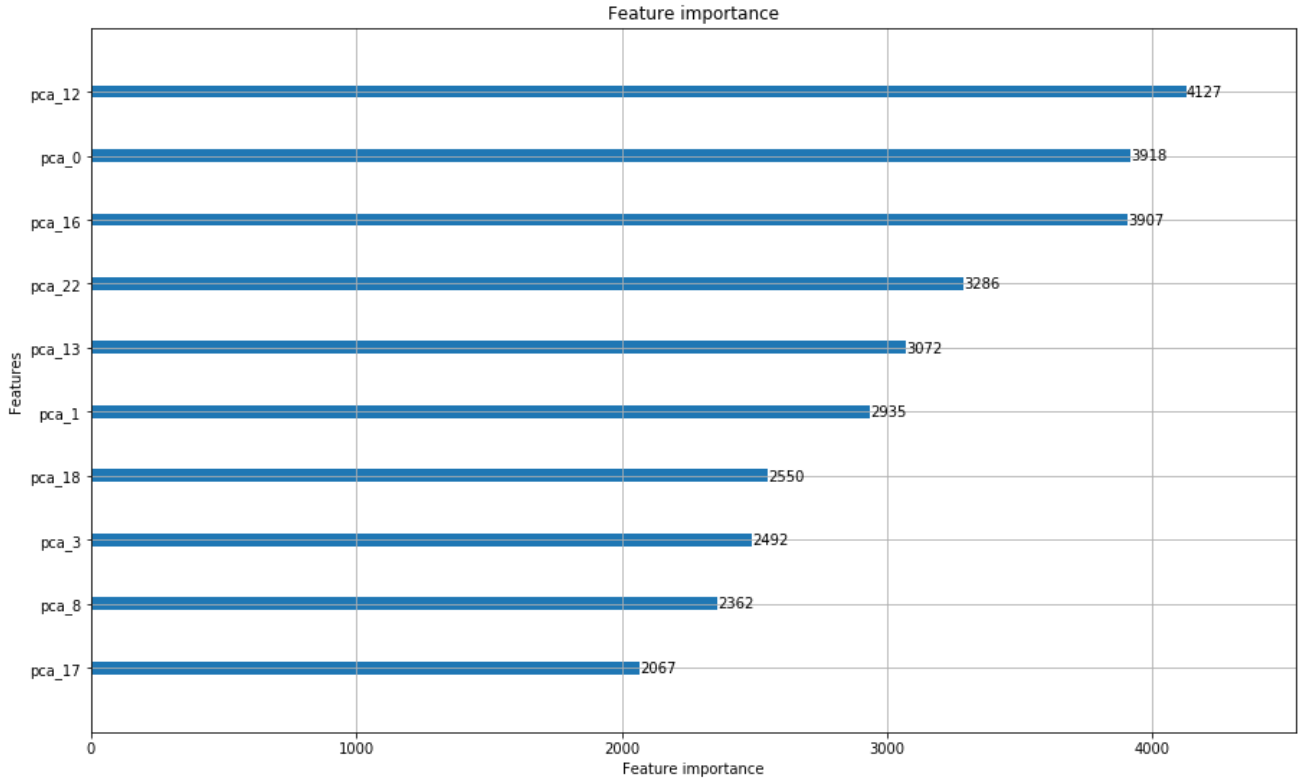


Figure OA.4 Top 10 features' contribution of PCA factors in the construction of LBM. Note: Factors reported are all principal components.

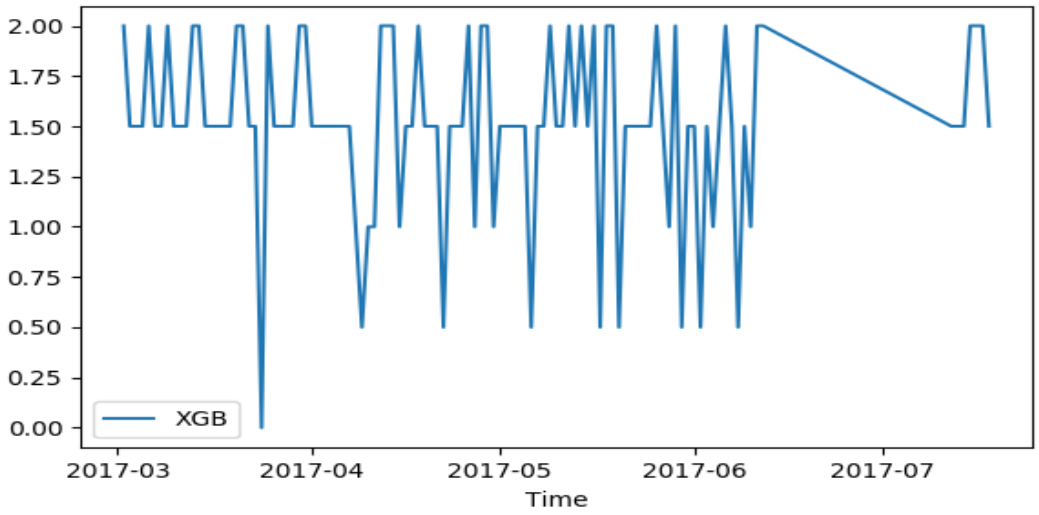
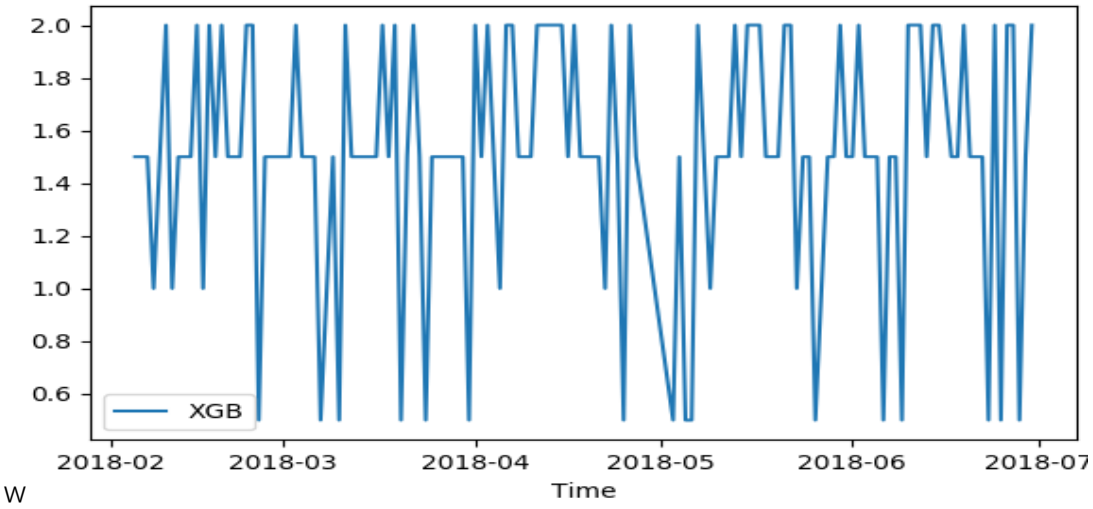


Figure OA.5 Hybrid strategy leverages PCA factors (F1)



W

Figure OA.6 Hybrid strategy leverages PCA factors (F2)

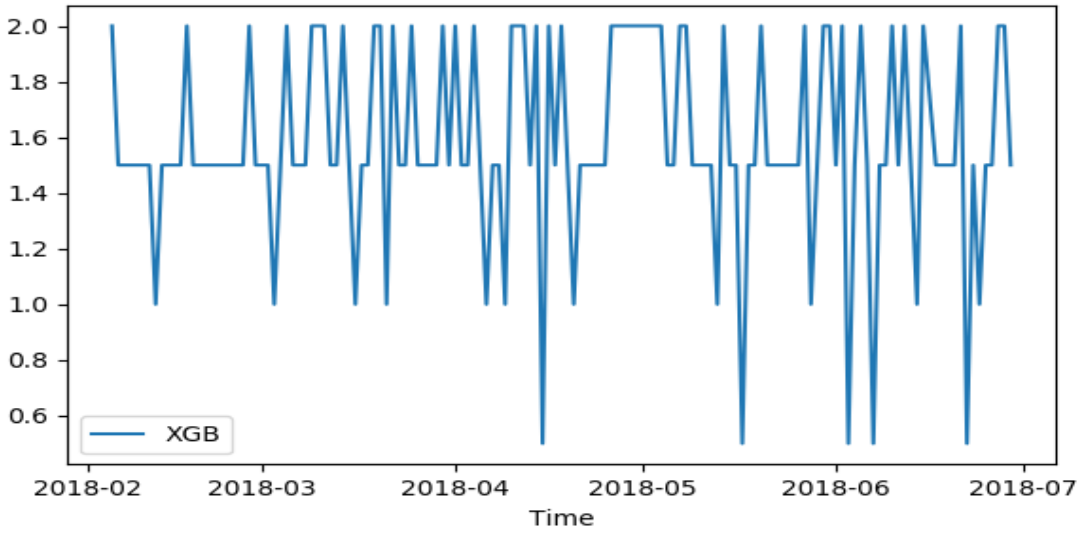


Figure OA 5 Hybrid strategy leverages PCA factors (F3)

Reference:

Azqueta-Gavaldón, A. (2020). Causal inference between cryptocurrency narratives and prices: Evidence from a complex dynamic ecosystem. *Physica A: Statistical Mechanics and its Applications*, 537, 122574.

Blei, D.M., Blei, Ng, A.Y., Jordan, M.I., Jordan and Lafferty, J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Chen, Z. and Doss, H., 2019. Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modeling. *Journal of Computational and Graphical Statistics*, *28*(3), pp.567-585.

Darst, B.F., Malecki, K.C. and Engelman, C.D., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics*, *19*(1), pp.1-6.

Elith, J., Leathwick, J.R. and Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), pp.802-813.

Gregorutti, B., Michel, B. and Saint-Pierre, P., 2017. Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), pp.659-678.

Granitto, P.M., Furlanello, C., Biasioli, F. and Gasperi, F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, *83*(2), pp.83-90.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Joakim, C., 2012. Explore Python, machine learning, and the NLTK library. *IBM Developer Works*.

Larsen, V.H. and Thorsrud, L.A., 2019. The value of news for economic developments. *Journal of Econometrics*, *210*(1), pp.203-218.