Annals of Mathematics and Computer Science
ISSN: 2789-7206
Vol 11 (2023) 19-28

HIGH FREQUENCY BITCOIN PRICE PREDICTION: STATISTICAL LEARNING APPROACH

SAIDJON KAMOLOV[1]*

ABSTRACT. While the next day price financial forecasting has been the most popular scenario in the literature, high frequency forecasting has received considerably less attention. In this paper, we consider the problem of predicting the Bitcoin price in the context of high frequency trading. To this end, we compare the performance of the elastic net (Enet) regression against the standard ordinary least squares (OLS) linear regression. The results are surprising in that OLS is more accurate than the more sophisticated Enet. We also identify the driving covariates under each model.

1. INTRODUCTION

High frequency trading plays a major role in financial markets. In particular, the high frequency trading volume of crypto-currencies (cryptos) has exploded in recent years due to their superior market performance. Given the potential financial gains, a significant amount of research has been dedicated to forecasting the price of cryptos. However, modeling the price of a crypto is an increasingly challenging task. Indeed, it is practically impossible to forecast the price of a crypto if one believes in the efficient market hypothesis. Nevertheless, in this paper, we attempt to consider the feasibility of modeling the price of cryptos using ordinary least squares regression (OLS) and elastic net (Enet). We find interesting and unexpected conclusions regarding the relative performances of the two models.

To analyze high frequency crypto trading we focus on forecasting the next minute closing price of Bitcoin. We consider minute-based data from Oct 2, 2020 to Mar 31, 2021 which totals almost 600K minute observations. We compare the performances of the OLS and Enet under different scenarios. As part of our analysis, we determine the key factors that affect the price of Bitcoin on the minute to minute basis.

2. LITERATURE

The problem of predicting the price of a financial asset has attracted an enormous amount of research. Financial forecasting includes stock prices, market indices, exchange rates, and cryptos [6, 14, 15, 16, 17, 20]. In particular, forecasting Bitcoin price has recently garnered a tremendous amount of attention due to its rise in value

Date: Received: Oct 17, 2022; Accepted: Dec 20, 2022.
* Corresponding author.
2010 *Mathematics Subject Classification.* Primary 46L55; Secondary 44B20.
Key words and phrases. Bitcoin, forecasting, high frequency, elastic net, least squares regression.

[7]. As with many other financial assets, forecasting Bitcoin consists of three main approaches: statistical models, autoregressive models, and machine learning models. Statistical models employ least squares regression to predict the future price of Bitcoin. Autoregressive models employ ARIMA-GARCH based models to use the previous time-series values to predict the next-step series value. Machine learning models employ neural network and other AI algorithms to allow data-driven rules to predict the price of Bitcoin.

Statistical models have been shown to perform well. In [27], the authors used regression models to predict 1-minute Bitcoin data. In [4] the authors find that statistical model such as logistic regression (LR) and linear discriminant analysis outperform machine learning algorithms such as support vector machines and long short-term memory (LSTM) in predicting 5-minute interval price prediction for Bitcoin. On the other hand, it is argued in [22], based on a comprehensive survey, that traditional statistical methods, albeit simple to implement and interpret, need a lot of assumptions that could be unrealistic, which leaves machine learning as the best technology in the field of crypto, to predict price based on experience. The authors in [24] used differential evolution, a metaheuristic optimization technique, to fit polynomial regression on Bitcoin data. A number of statistical learning algorithm were compared in [30] and found that extra-trees regression provides optimal results. Quantile regression in [32] showed that investor attention has a positive relationship with crash risk when crash risk is low and negative when crash risk is high. Simple linear regression using univariate series forecast using with closing prices was employed in [36]. Linear discriminant analysis using sentiment analysis was employed in [8] to predict Bitcoin price movement. Similarly, support vector machines with sentiment analysis were used in [9].

A significant amount research to predict Bitcoin using machine learning techniques has been carried out. There exist many different proposed models that provide mixed results. In [26], the authors combine gated recurrent unit (GRU) and LSTM into a hybrid system to predict crypto values. Another hybrid approach that combines LR and LSTM was proposed in [2]. A comparison of recurrent neural network (RNN), GRU, and LSTM showed that GRU with recurrent dropout provides the best performance [5]. An empirical comparison done by [25] showed that LSTM outperforms other machine learning algorithms. The authors in [29] argue that a simple feedforward network is sufficient to outperform ARIMA and SVM models. Another hybrid approach based GRU and LSTM was proposed in [23]. An alternative approach using reinforcement learning was proposed in [31] that achieved highly accurate crypto price forecasts. One of the main issues in forecasting Bitcoin price could be imbalanced data and high dimensional data. There exist several approaches to deal with imbalanced [11, 12, 21, 34] as well as feature selection [10, 13, 19, 33, 35].

The ARIMA model has been shown to provide underwhelming results. In [3], the authors showed that the ARIMA model is unable to capture the sharp fluctuations in the price. Similarly, the ARIMA model was implemented in [28]. Nontraditional sources of covaraites such as tweet volumes and sentiment analysis have also been employed in attempt to predict Bitcoin price [1].

3. METHODOLOGY

3.1. **Data.** The dataset used in our study consists of minute-based information regarding the values of various factors that can potentially affect the price of Bitcoin. The list of the variables included in our experiments is given in Table 1. As shown in Table 1, we consider factors such as the prices of other cryptos, foreign exchange rates, gold, oil, market indices and others. There are a total of 108 variables used as inputs in the regression model. The values are given on the per minute basis. The data ranges from Oct 2, 2020 8:01 to Mar 31, 2021 23:59. There are a total of 598,098 observations. The data is aggregated from various sources that are publicly available online. The data is normalized to have mean 0 and standard deviation 1. Normalization allows us to compare the values of the coefficients directly. It also makes it easier to set the L_1 ratio parameter.

TABLE 1. The list of covariates used in the prediction of Bitcoin price.

ADA-open	ADA-high	ADA-low	ADA-close	ADA-volume
ADA-tradecount	BNB-open	BNB-high	BNB-low	BNB-close
BNB-volume	BNB-tradecount	BTC-open	BTC-high	BTC-low
BTC-close	BTC-volume	BTC-tradecount	DOGE-open	DOGE-high
DOGE-low	DOGE-close	DOGE-volume	DOGE-tradecount	ETH-open
ETH-high	ETH-low	ETH-close	ETH-volume	ETH-tradecount
LINK-open	LINK-high	LINK-low	LINK-close	LINK-volume
LINK-tradecount	LTC-open	LTC-high	LTC-low	LTC-close
LTC-volume	LTC-tradecount	XRP-open	XRP-high	XRP-low
XRP-close	XRP-volume	XRP-tradecount	AUDUSD-Open	AUDUSD-High
AUDUSD-Low	AUDUSD-Close	EURUSD-Open	EURUSD-High	EURUSD-Low
EURUSD-Close	GBPUSD-Open	GBPUSD-High	GBPUSD-Low	GBPUSD-Close
NZDUSD-Open	NZDUSD-High	NZDUSD-Low	NZDUSD-Close	CADUSD-Open
CADUSD-High	CADUSD-Low	CADUSD-Close	USDCHF-Open	USDCHF-High
USDCHF-Low	USDCHF-Close	JPYUSD-Open	JPYUSD-High	JPYUSD-Low
JPYUSD-Close	ASX200-Open	ASX200-High	ASX200-Low	ASX200-Close
DAX30-Open	DAX30-High	DAX30-Low	DAX30-Close	FTSE100-Open
FTSE100-High	FTSE100-Low	FTSE100-Close	NASDAQ100-Open	NASDAQ100-High
NASDAQ100-Low	NASDAQ100-Close	NIKKEI225-Open	NIKKEI225-High	NIKKEI225-Low
NIKKEI225-Close	S&P500-Open	S&P500-High	S&P500-Low	S&P500-Close
Gold-Open	Gold-High	Gold-Low	Gold-Close	Crudeoil-Open
Crudeoil-High	Crudeoil-Low	Crudeoil-Close		

We split the data into two parts. The first part comprises the initial 80% of the observations which is used to train the regression models. The second part consists of the remaining 20% of the data. Note that since we are dealing with time series data, the split must be performed using the temporal structure. In particular, the traditional random split is not acceptable to avoid leaking the test data into the training process.

After the regression coefficients are fitted on the train set, the model accuracy is evaluated on the unseen test set. This process allows us to measure the generalizability of the model.

3.2. **Models.** We consider two different regression models in our study: OLS and Enet. The OLS model is a classical regression method that fits a hyperplane to multidimensional dataset. The objective of OLS is to determine the coefficients of the covariates so as to minimize the total mean squared error between the predicted and the actual values of the target variable. Since the data is normalized in the preprocessing stage we fit OLS without fitting the intercept. The second model used in our study is Enet which is a regularized regression model. Enet is based on the OLS model plus a penalty on the size of the regression parameters. The penalty is introduced to restrict the values of the coefficients and avoid overfitting, particularly in high dimensional settings. The Enet penalty consists of L_1 and L_2 components. The objective function of Enet is given by the following equation

$$\frac{1}{2n} \cdot \|y - Xw\|_2^2 + \alpha \cdot r \cdot \|w\|_1 + 0.5 \cdot \alpha \cdot (1 - r) \cdot \|w\|_2^2, \tag{1}$$

where X is the matrix of input variables, y is the target variable, w is the vector coefficients, α is the regularization parameter, and r is the trade-off parameter between L_1 and L_2 regularization. Thus, Enet is a combination of Lasso (L_1) and Ridge (L_2) regressions. The amount of penalty and the weight of the L_1 and L_2 penalties is controlled by the corresponding parameters α and r.

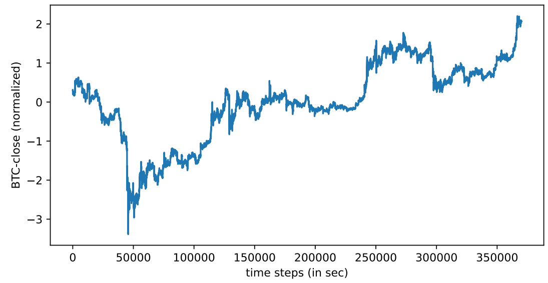
All the numerical experiments are conducted in Python using the scikit-learn library. We perform model parameter tuning using the train set. Unless otherwise mentioned, all the default settings of the models are applied.

4. Numerical experiments

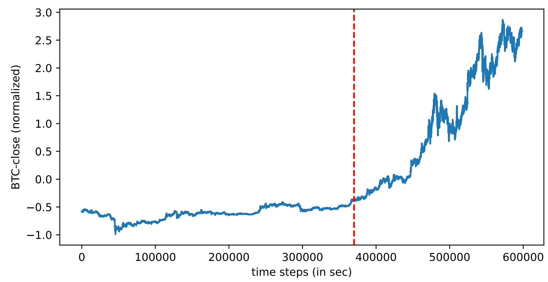
We carry out 4 major experiments. In the first 2 experiments, we forecast the closing price of Bitcoin in the next time step $(t + 1)$ using the covariate values in the current time step (t). While in the second 2 experiments, we explore the relationship between the price of Bitcoin at the current time step (t) and the non-Bitcoin covariates at the current time step (t).

We consider the full dataset as well as a partial dataset. Due to the nature of Bitcoin price movement, there is a structural break at the time step 370,000. After this time step the price of Bitcoin grows exponentially. Given the significant change in the behavior of Bitcoin, it is decided to separate the data into pre and post growth periods. The price of Bitcoin over the initial 370K time steps and the total 600K time steps is illustrated in Figure 1. Thus, we conduct separate experiments using the partial data consisting of the initial 370K observations. In addition, we conduct experiments with full dataset of 600K observations.

In the first set of experiments, we predict the price of Bitcoin at time $t + 1$ using the information from time t. The target variable is the closing price of Bitcoin at the end of a minute $(t + 1)$. The input variables are all the covariate values from the previous minute (t). The input variables include the price of other cryptos, gold, oil, and other



(A) The price of Bitcoin over the initial 370,000 minutes.



(B) The price of Bitcoin over the full 600,000 minutes.

FIGURE 1. The normalized closing price of Bitcoin over two time periods.

factors. The input features also include the Bitcoin price information from the previous minute (t). The model can be described as follows

$$y_{t+1} = c_1 x_1 + c_2 x_2 + ... + c_{108} x_{108}. \tag{2}$$

The results of fitting Enet to the initial 370K time steps are shown in Figure 2. The minimum MSE achieved by the model is 2.5512e-05. The model performs optimally the lowest values of the parameter α. It is a surprising outcome given the likelihood of overfitting. On the other hand, we observe from the right-hand graph that the values of most of the coefficients converge to zero at values of α.

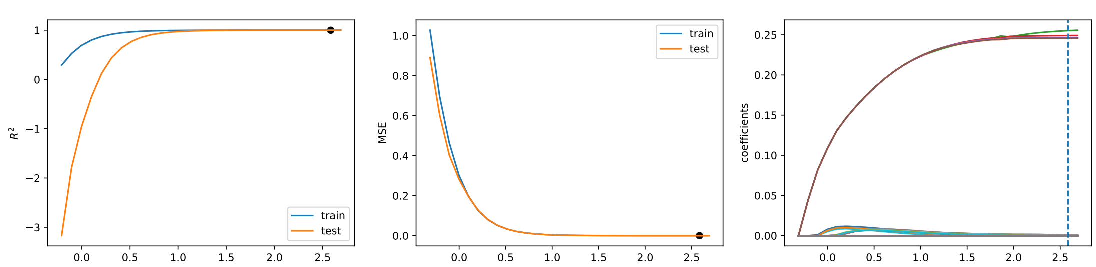


FIGURE 2. The performance of the model over a range of values of the α parameter using the initial 370K observations.

The corresponding nontrivial covariates of the optimal model are presented in Figure 3. It is not surprising to see that Bitcoin-based variables are the most relevant in determining the Bitcoin price in the next time step. The minimum MSE achieved by the OLS model is 1.8295e-05 which is lower than the Enet model. It is a surprising outcome given that OLS is a basic model. On the other hand, it can be seen that Enet performs optimally at the lowest value of α which is equivalent to the OLS setting.

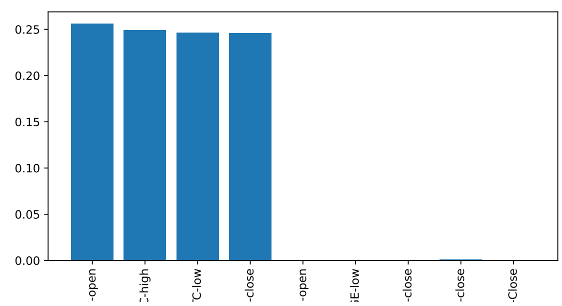


FIGURE 3. Top features in the optimal model with 370K observations.

The results of fitting Enet to the full set of data of 600K observations are presented in Figure 4. The results are in line with the outcome based 370K observations. In particular, the model achieves the optimal performance at the lowest value of the parameter α. The minimum test MSE achieved is 4.0028e-05.

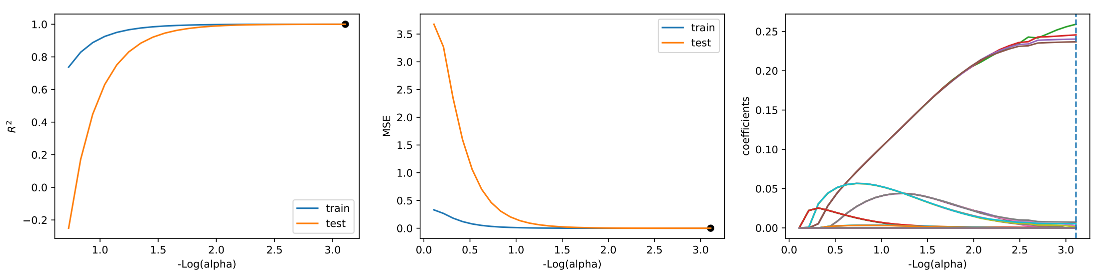


FIGURE 4. The performance of the model over a range of values of the α parameter using the full 600K observations.

The corresponding relevant variables of the optimal model are presented in Figure 5. It is not surprising to see that Bitcoin-based variables are the most important covariates that drive the price of Bitcoin in the next time step. The minimum test MSE achieved by the OLS model on the full dataset is 2.1643e-05 which is lower than the Enet model. It is a both surprising and expected outcome. On one hand, Enet is a more advanced model and should achieve better results than OLS. On the other hand, Figure 5 shows that Enet performs best at the lowest value of the parameter α which is equivalent to the OLS model.

In the second part of our study, we consider the relationship between the Bitcoin price and its covariates. To this end, we consider the following model

$$y_t = c_1 x_1 + c_2 x_2 + \ldots + c_{104} x_{104}.$$

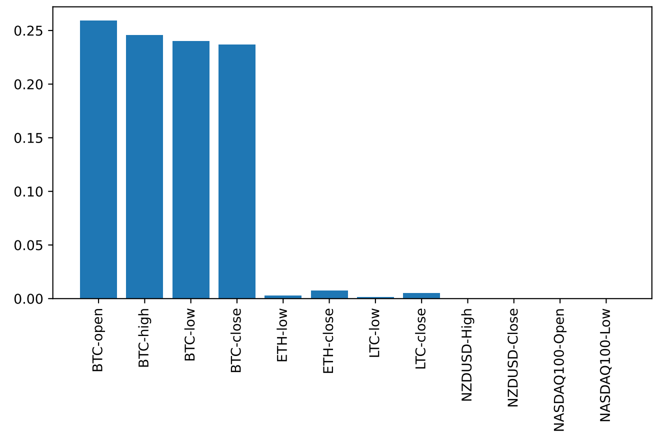


FIGURE 5. Top features in the optimal model with 600K observations.

In this model, the target variable is the Bitcoin price at time t and the input variables are the values of the covariates also at time t, excluding the four Bitcoin-related features. The results of fitting Enet on initial 370K time steps are presented in Figure 6. It can be seen from the figure that the minimum test MSE is occurs at around $-log(\alpha) = 1.6$. The test MSE exhibits the classical convex shape. It shows that too much de-regularization can lead to overfitting. The minimum test MSE is 0.1229.

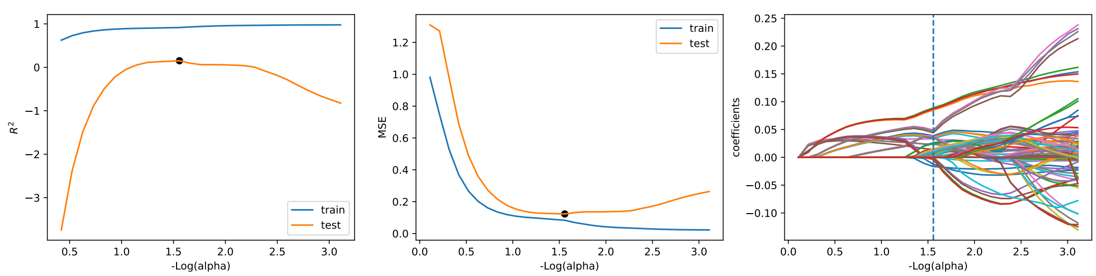


FIGURE 6. The performance of the model without the Bitcoin covariates using the initial 370K observations.

The corresponding nonzero coefficients of the optimal model are presented in Figure 7. It can be seen that several factors affect the Bitcoin price. Most notably ETH, AUDUSD, NASDAQ, and XPR are the main driving forces of the Bitcoin price according to the Enet model on the initial data. The corresponding test MSE achieved by the OLS model is 0.1766 which is higher than the Enet model. The OLS model is less accurate than the Enet model in this particular setting. This is not surprising given the the optimal α in the Enet MSE graph is in the middle of the range. It suggests that the best performance takes place away from the OLS setting.

Finally, we train and test the above model on the entire dataset. The results using the Enet regression are presented in Figure 8. The results of using the full dataset are significantly different than using the initial 370K time steps. In particular, the test MSE continues to decrease as the regularization is relaxed. Thus the lowest test MSE

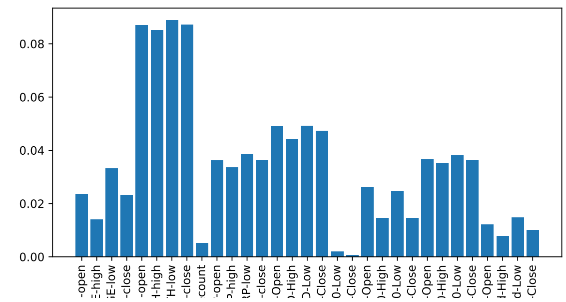


FIGURE 7. Top features in the optimal model without the Bitcoin covariates using the initial 370K observations.

is 0.3059 which is achieved at the lowest value of the parameter α. It means that to achieve the best results there should be no restriction on the values of the parameters.

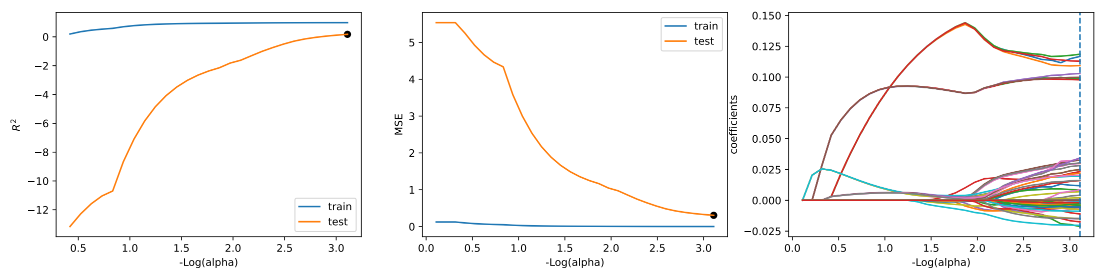


FIGURE 8. The performance of the model without the Bitcoin covariates using the full 600K observations.

The corresponding nontrivial coefficients of the model are presented in Figure 9. We observe from the figure that ETH, LTC, DODGE, NZDUSD, and NASDAQ are the most significant drivers of the Bitcoin price in this model. The corresponding test MSE using the OLS model is 0.2001 which lower than Enet model. It is not surprising given the MSE curves in Figure 8 where we see that the optimal result is achieved with very little regularization i.e. the setting of the OLS model.

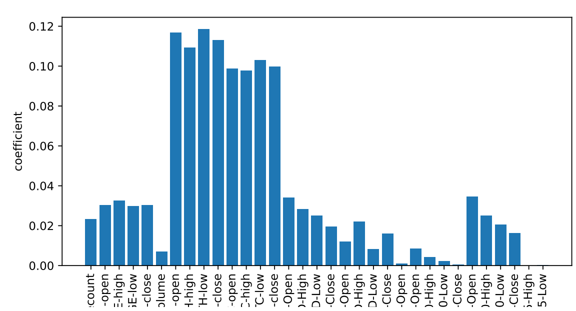


FIGURE 9. Top features in the optimal model without the Bitcoin covariates using the full 600K observations.

5. CONCLUSION

In this paper, we considered the task predicting the Bitcoin price in the context of high frequency trading. Concretely, we analyzed minute-based trading data that included 108 covariates and 600K observations. We compared the performance of the Enet model to OLS and discovered surprising results. In particular, we found that OLS outperformed Enet in several scenarios. We also observed that Bitcoin-related variables such as previous closing, open, and high prices are the most important factors in determining the price in the next time step. The results of the present study indicate that Bitcoin-based information together with a highly flexible model can be the optimal choice in predicting the next time step price of Bitcoin.

REFERENCES

[1] Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3), 1.

[2] Andi, H. K. (2021). An accurate bitcoin price prediction using logistic regression with LSTM machine learning model. Journal of Soft Computing Paradigm, 3(3), 205-217. Chicago

[3] Azari, A. (2019). Bitcoin price prediction: An ARIMA approach. arXiv preprint arXiv:1904.05315.

[4] Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. Journal of Computational and Applied Mathematics, 365, 112395.

[5] Dutta, A., Kumar, S., & Basu, M. (2020). A gated recurrent unit approach to bitcoin price prediction. Journal of risk and financial management, 13(2), 23.

[6] Gurrib, I., & Kamalov, F. (2019). The implementation of an adjusted relative strength index model in foreign currency and energy markets of emerging and developed economies. Macroeconomics and Finance in Emerging Market Economies, 12(2), 105-123.

[7] Gurrib, I., Elshareif, E. E., & Kamalov, F. (2020). The effect of energy cryptos on efficient portfolios of key energy listed companies in the S&P composite 1500 energy index. International Journal of Energy Economics and Policy, 10(2), 1-15.

[8] Gurrib, I., & Kamalov, F. (2021). Predicting bitcoin price movements using sentiment analysis: a machine learning approach. Studies in Economics and Finance.

[9] Gurrib, I., Kamalov, F., & Smail, L. (2021, August). Bitcoin Price Forecasting: Linear Discriminant Analysis with Sentiment Evaluation. In The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research (pp. 1-5).

[10] Kamalov, F. (2018, December). Sensitivity analysis for feature selection. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1466-1470). IEEE.

[11] Kamalov, F. (2020). Kernel density estimation based sampling for imbalanced class distribution. Information Sciences, 512, 1192-1201. Chicago

[12] Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. Knowledge-Based Systems, 207, 106368.

[13] Kamalov, F., & Leung, H. H. (2020). Outlier detection in high dimensional data. Journal of Information & Knowledge Management, 19(01), 2040013.

[14] Kamalov, F. (2020). Forecasting significant stock price changes using neural networks. Neural Computing and Applications, 32(23), 17655-17667.

[15] Kamalov, F., Smail, L., & Gurrib, I. (2020, December). Forecasting with Deep Learning: S&P 500 index. In 2020 13th International Symposium on Computational Intelligence and Design (ISCID) (pp. 422-425). IEEE.

[16] Kamalov, F., Smail, L., & Gurrib, I. (2020, November). Stock price forecast with deep learning. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 1098-1102). IEEE.

[17] Kamalov, F., Gurrib, I., & Rajab, K. (2021). Financial forecasting with machine learning: price vs return. Journal of Computer Science, 17(3), 251-264.

[18] Kamalov, F., Sulieman, H., & Santandreu Calonge, D. (2021). Machine learning based approach to exam cheating detection. Plos one, 16(8), e0254340.
[19] Kamalov, F. (2021). Orthogonal variance decomposition based feature selection. Expert Systems with Applications, 182, 115191. Chicago
[20] Kamalov, F., & Gurrib, I. (2022). Machine learning-based forecasting of significant daily returns in foreign exchange markets. International Journal of Business Intelligence and Data Mining, 21(4), 465-483.
[21] Kamalov, F., Moussa, S., & Avante Reyes, J. (2022). KDE-Based Ensemble Learning for Imbalanced Data. Electronics, 11(17), 2703.
[22] Khedr, A. M., Arif, I., El-Bannany, M., Alhashmi, S. M., & Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. Intelligent Systems in Accounting, Finance and Management, 28(1), 3-34.
[23] Hamayel, M. J., & Owda, A. Y. (2021). A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms. AI, 2(4), 477-496.
[24] Jana, R. K., Ghosh, I., & Das, D. (2021). A differential evolution-based regression framework for forecasting Bitcoin price. Annals of Operations Research, 306(1), 295-320.
[25] Ji, S., Kim, J., & Im, H. (2019). A comparative study of bitcoin price prediction using deep learning. Mathematics, 7(10), 898.
[26] Patel, M. M., Tanwar, S., Gupta, R., & Kumar, N. (2020). A deep learning-based cryptocurrency price prediction scheme for financial institutions. Journal of information security and applications, 55, 102583.
[27] Phaladisailoed, T., & Numnonda, T. (2018, July). Machine learning models comparison for bitcoin price prediction. In 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 506-511). IEEE.
[28] Poongodi, M., Vijayakumar, V., & Chilamkurti, N. (2020). Bitcoin price prediction using ARIMA model. International Journal of Internet Technology and Secured Transactions, 10(4), 396-406.
[29] Rajabi, S., Roozkhosh, P., & Farimani, N. M. (2022). MLP-based Learnable Window Size for Bitcoin price prediction. Applied Soft Computing, 129, 109584.
[30] Saheed, Y. K., Ayobami, R. M., & Orje-Ishegh, T. (2022). A Comparative Study of Regression Analysis for Modelling and Prediction of Bitcoin Price. In Blockchain Applications in the Smart Era (pp. 187-209). Springer, Cham.
[31] Shahbazi, Z., & Byun, Y. C. (2021). Improving the cryptocurrency price prediction performance based on reinforcement learning. IEEE Access, 9, 162651-162659.
[32] Smales, L. A. (2022). Investor attention and cryptocurrency price crash risk: a quantile regression approach. Studies in Economics and Finance.
[33] Thabtah, F., Kamalov, F., & Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. International journal of medical informatics, 117, 112-124.
[34] Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. Information Sciences, 513, 429-441.
[35] Thabtah, F., Kamalov, F., Hammoud, S., & Shahamiri, S. R. (2020). Least Loss: A simplified filter method for feature selection. Information Sciences, 534, 1-15.
[36] Uras, N., Marchesi, L., Marchesi, M., & Tonelli, R. (2020). Forecasting Bitcoin closing price series using linear regression and neural networks models. PeerJ Computer Science, 6, e279.

[1]FACULTY OF ENGINEERING, TAJIK TECHNICAL UNIVERSITY, DUSHANBE, TAJIKISTAN.
Email address: said.kamolov@ttu.tj