ORIE 5740 - STATISTICAL DATA MINING

Predicting Cryptocurrency Prices using Machine Learning

May 6, 2021

Assigned by:

Prof. Damek Davis

Submitted by:

Niharika Dalsania (nd322)

Joel Dsouza (jnd74)

Sukriti Kumar (sk3335)

Abstract

Cryptocurrency has been gaining a lot of popularity recently because of its uncontrollable, untraceable nature as well as support from a lot of very influential people. This has led to an increase in popularity amongst academic researchers as well. The high volatility and the non linear nature of these prices makes it quite difficult to predict them. Various machine learning algorithms have been used to predict prices of cryptocurrencies but we look to including macro economic variables for this analysis. Various linear and non linear methods were used to predict the bitcoin prices. We have also tried to classify up and down movement in bitcoin prices.

1 Introduction

Cryptocurrency has gained immense attention recently because of its decentralized nature. Because of the high volatility and non-linear nature, cryptocurrencies have a very low signal-to-noise ratio which makes the price predictions quite cumbersome, hence we look at predicting daily percent returns. This makes more sense because prices are not stationary but returns are. In this analysis, we try to study the impact on cryptocurrency price movements of various macroeconomic factors, cryptocurrency specific features including market, activity, block, mining, economics as well as market sentiment factors and use them for price prediction.

The aim of our project is to forecast returns of bitcoin prices over changing crypto market regimes having bull and bear market periods. In the subsequent sections, we explain our choice of predictor variables (which have daily frequency). In order to understand the data first, we do exploratory data analysis on each of the feature variables to study if there exists a pattern between predictors and bitcoin prices, analyze the different distributions of the data, deal with outliers and with non-stationarity in time series. Next, we look at linear and non-linear models for the prediction of future returns and analyze feature importance. We also look at classification models to predict the up and down movement in price. Finally, we compare and summarize the results as well as extend on future work of this research.

2 Dataset

The dataset used ranges from January 2015 to March 2020. For a time-series dataset, we need to carefully divide our dataset to avoid data leaking into training data by incorporating information from the future. Therefore, we cannot simply pick a random subset of data for training and testing. We took the first 80% of the rows for training and the last 20% of the rows (recent years) to constitute our test dataset. Because our dataset is not very large (approximately 3.8 years of training data, 1 year of validation data and 1.3 years of test data) we have not considered a sliding window for the validation as is done for time series analysis. The training set is used to form

Predictor	**Rationale**
Lagged Bitcoin Price	Price of bitcoin the previous day. Autocorrelation reveals that only the first lag is significant.
Realised Volatility	Rolling 5 day standard deviation of bitcoin prices. Returns on bitcoin price will be impacted more in high volatility and less in low volatility.
Twitter Sentiment	Sum of sentiment scores (between -1 to 1, -1 being the most negative and 1 being the most positive) of all the tweets having keyword "bitcoin" on a specific day, filtered by choosing only tweets from people with more than 10k followers to get authentic/influential tweets. Scores are added, not averaged so that the information of number of tweets is also captured (prices are likely to be affected in light of an event that lead to many people tweeting about it)
Google Trends	Number of times the keyword "bitcoin" was searched worldwide on a specific day. This will capture the market sentiment in volatility of bitcoin price
DIA.US.Index	Dow Jones Industrial Index. It represents all companies without being heavily skewed by tech companies. In case of a market rally, the crypto currencies should also go up.
CNYUSD.Curncy	The currency exchange rate for Chinese Renminbi and US Dollar. A lot of crypto mining is done in China and hence including the exchange rate will implicitly capture this as well.
EURUSD.Curncy	The currency exchange rate for the European Union and US Dollar. The European Union is one of the biggest regions for trading of cryptocurrencies. Hence, demand of cryptocurrencies is implicitly captured in this variable and hence the prices of crypto currencies might get impacted by the currency exchange rate.
GDX.US.Equity	A lot of recent studies suggest that gold prices and bitcoin prices are treated as safe haven and thus the price of gold may have an impact on cryptocurrencies.
VIX.Index	Implied Volatility of S&P 500 index. It captures the market's sentiments of movement in the S&P 500 index. In case of volatility change, the crypto currency returns will also get impacted.
VVIX.Index	Implied volatility of implied volatility (second derivative) of S&P 500 index. It captures the market's sentiments of movement in the volatility of S&P 500 index. In case of volatility change, the crypto currency returns will also get impacted.
SHCOMP.Index	SSE Composite Index stock index. It represents all companies in the Chinese market. A lot of crypto mining is done in China and hence including the Chinese market will implicitly capture this as well.
TWSE.Index	Taiwan Capitalisation weighted stock index. It represents all companies in the Taiwan market. There are no regulations on the Taiwan market for cryptocurrencies and these are freely traded on Taiwan exchanges and hence taiwan market by impacting the supply and demand of cryptocurrencies a lot and hence the returns of crypto currencies.
SPX.Index	S&P 500 index. It represents all companies in the US market. In case of a market rally, the crypto currencies should also rally.
ADS_Index	The index is designed to track real business conditions in the US which has real implications on the stock market performance and volatility. As the real business conditions worsens, the stock market will crash and so should the bitcoin prices.
Oil prices	Crypto currencies are for investing. Oil is a consumption commodity so the intuitive guess is that when consumption commodity prices increases, the investing commodity prices should increase.
Lumber prices	Crypto currencies are for investing. Lumber is a consumption commodity so the intuitive guess is that when consumption commodity prices increases, the investing commodity prices should increase.
Bitcoin Specific Predictors	Bitcoin/Blockchain related features such as Market Capitalisation, Price Volatility, Two week hash growth, Miner Revenue Value, Fee Rate etc. are expected to have a strong predictive power for Bitcoin Prices and are used as predictor variables.

Table 1: Rationale for chosen predictors (Refer here for more.)

the model, the validation set is used to tune the hyperparameters and the test data is used to check the accuracy of the trained model.

Table 1 shows the predictor variables chosen for the problem and explains the rationale behind choosing them. A major predictor is the lagged cryptocurrency price itself.

Looking at the ACF plot (here) shows that only the first lag is significant. The data used has several bitcoin-specific predictors and a few macro variables as well. A minor discrepancy in data arises because the cryptocurrency market functions 24 x 7 year-round, whereas macro variables are only available on weekdays. To resolve this issue, we front fill on weekends for macro variables data. This makes sense because the information available for the weekend forecast is the data at the end of the previous week. But note that this front fill might not be very meaningful visually as it will have constant values on weekends, so we perform exploratory data analysis and collinearity analysis separately for macro variables and bitcoin-specific variables.

3 Exploratory Data Analysis

We initially thought of analysing behaviour of Bitcoin, Litecoin, Ethereum, Dogecoin and Ripple for this analysis. But due to lack of data for independent variables for all other crytpocurrencies, we cannot perform analysis for Litecoin, Ethereum, Dogecoin and Ripple.

As part of exploratory data analysis, we thus analysed the the behaviour of all the independent variables with Bitcoin prices only. Figure 1 shows selected predictors and their changing patterns with bitcoin prices. This is mostly in accordance with the reasons established in Table 1. The levels of all these independent variables have been normalised so that they can be accurately represented on the plot together. As seen from the graphs, some variables (like Avg UTXO Value, Transaction per block, SP 500 Index, TWSE Index, Dow Jones Index, Lumber prices, Oil prices and Google trends) are good predictors of bitcoin prices and they follow the trend that the bitcoin prices follow.

It is however observed that these prices are not stationary, hence to remove trends we difference all independent variables and the dependent variable. Only sentiment score has not been differenced because there was no pattern observed in the series. Some of the differenced series have been shown in Figure 2. These series now show no trend and are good to be used as predictors.

Next, we look at the collinearity with features in order to remove redundant features. The heatmap in Figure 3 reveals correlations among macro variables. We can see a strong positive correlation between SP500 and Dow Jones as well as VIX and VVIX. Also, both VIX and VVIX are strongly negatively correlated with Dow Jones

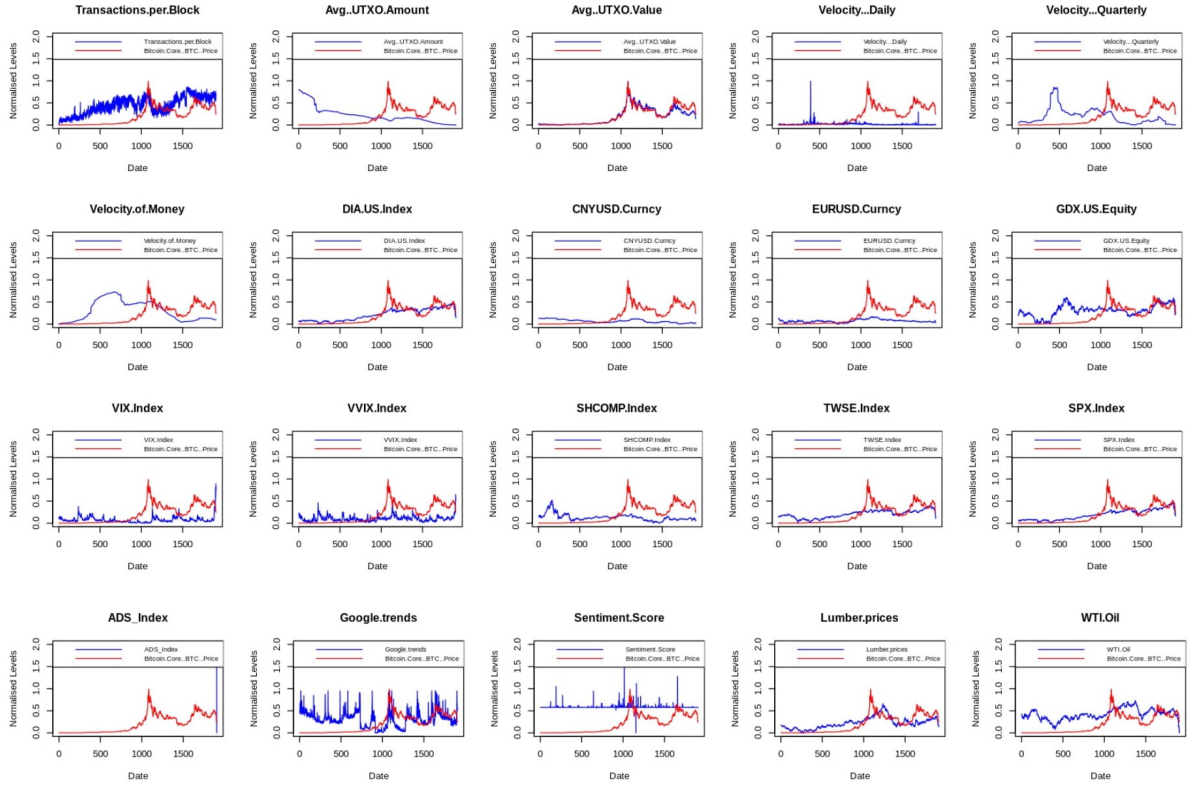

Figure 1: Time series of selected predictor variables. (Refer here for more.)

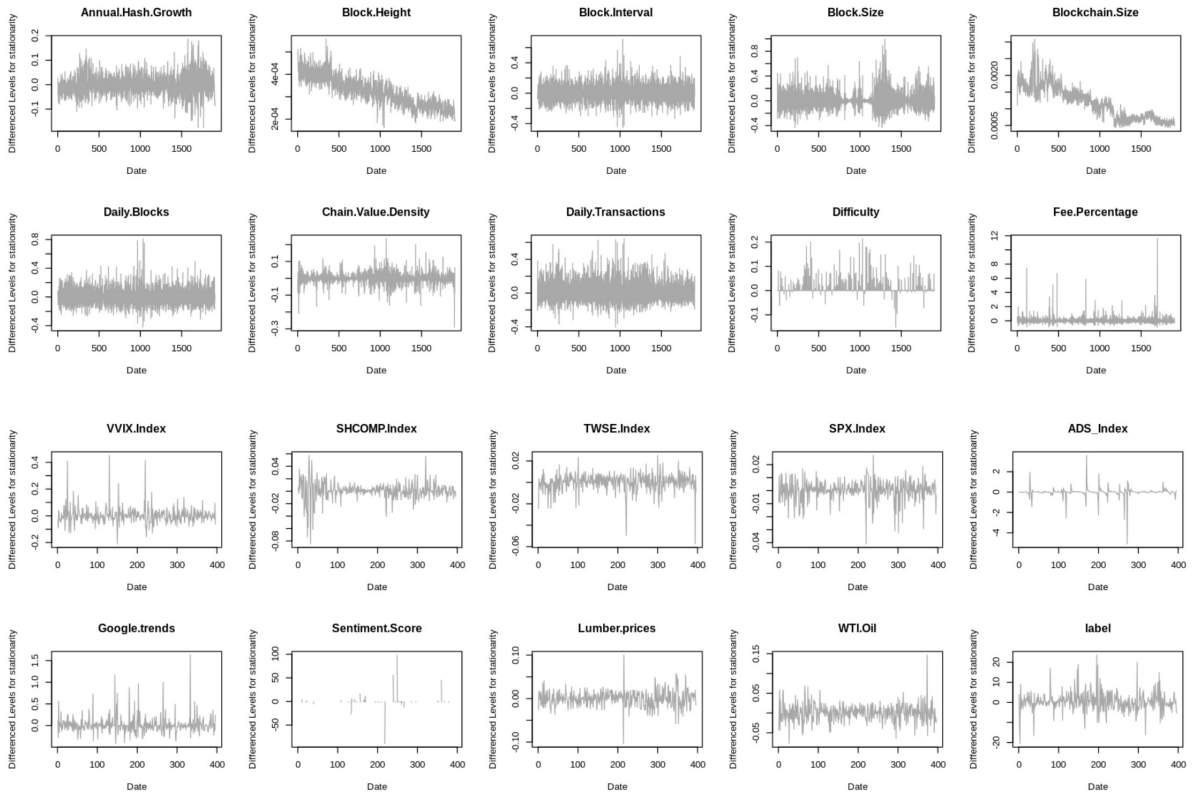

Figure 2: Stationery time series of selected predictor variables. (Refer here for more.)

index. We believe that a correlation of greater than 0.75 in magnitude between any predictors would lead to biased estimates in regression. So taking a threshold of 0.75, we remove all redundant predictor variables but one that have a positive or negative correlation beyond that threshold. Here we keep Dow Jones index and drop the other 3. Similar analysis and reduction is performed for bitcoin specific features as well. (Refer here for more.)

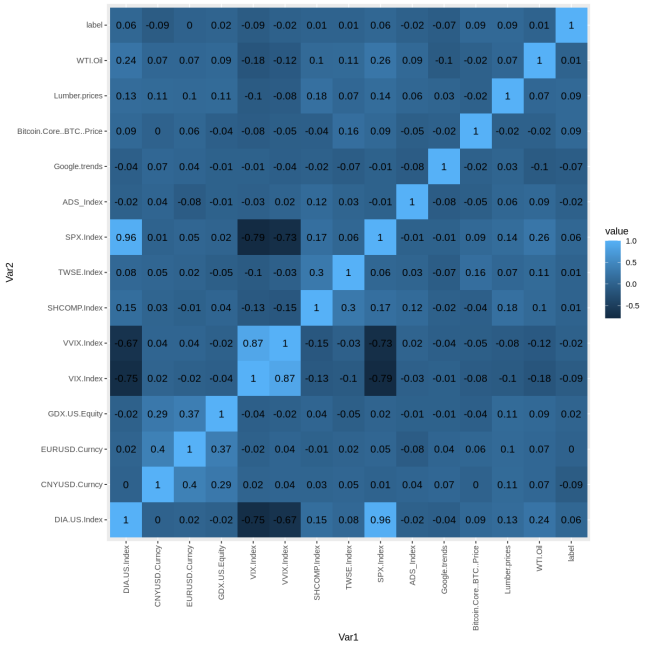

Figure 3: Multi-collinearity Analysis for macro variables

4 Models for Regression

In this section we first look at baseline ARIMA-GARCH model and then explore different models for regression to predict bitcoin returns and compare its performance with the baseline.

4.1 Baseline Model

In the finance industry, the norm is to use ARIMA and GARCH model for prediction of returns. We have thus compared the results of our models with the ARIMA and GARCH model, treating it as the baseline model for our purposes. Auto.arima() function is used to estimate the ARIMA part of the model. For the GARCH (p,q) model we looked at the autocorrelation and the partial autocorrelation of the bitcoin prices where p is the number of lagged return terms being considered (from the ACF graph) and q is the number of lagged error variance terms being considered (from the PACF graph). For our model, we have used the ARIMA (0,1) and GARCH (1,1) model. The parameters p and q for the GARCH model have been found using the Autocorrelation Function as well as the Partial Autocorrelation Function PACF (Refer here for more).

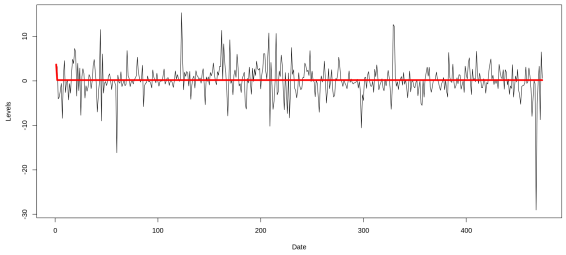

Figure 4: Performance of Baseline Model

4.2 Linear Models

For the initial model development, we first consider running an Ordinary Least Squares regression. The results are not great, as expected. To reduce more multicollinearity in the simple linear model, we also introduced three regularized models - Ridge, Lasso and ElasticNet. The penalty term in ElasticNet is simply a combination of the Ridge and Lasso penalties. The optimal regularisation parameters were obtained using the validation set. Unsurprisingly, Lasso and ElasticNet set most of the coefficients to zero barring the coefficients of only lagged price, meaning that the model is learning everything from the lagged price only and nothing else. Figure 5 shows the forecasted returns using Elastic Net, compared to true returns. RMSE performance results shown in Figure 8.

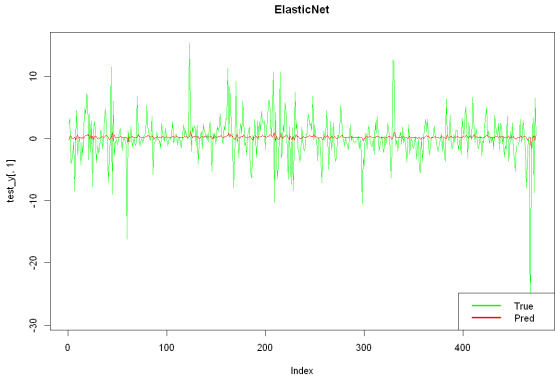

Figure 5: Performance of Elastic Net on Test Set

4.3 Non-Linear Models

We now try an ensemble method to predict the returns as it implicitly performs feature selection to generate uncorrelated random trees. This method is expected to perform better than Lasso and Ridge regressions because it captures non-linear trends as well. We used a validation set to optimize ntree hyperparameter.

As seen in Figure 7, on running the model on train data, we see that Google Trends has the maximum information along with lagged returns and the lagged standard deviation of BTC returns. Random Forests give us a much better fitting result, with a train RMSE of 2.3359 and test RMSE of 3.4657. Although the test error is much

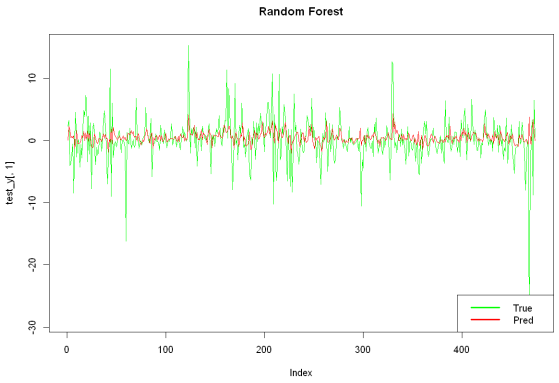

Figure 6: Performance of Random Forest on Test Set

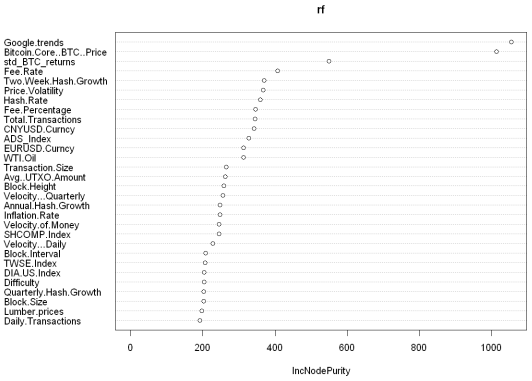

Figure 7: Variable Importance in Random Forest

higher than the train error, this is not due to overfitting, but due to the sudden investment interest during the time duration of test data. Figure 6 shows the forecasted returns using Random Forest, compared to true returns.

Next, we try boosting method which is another ensemble method using weak decision trees. Here, we use XGBoost which is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. We see that the lagged BTC returns, Google trends and lagged Hash rate are the most important features. Results: The test RMSE is 3.4646 which similar to Random Forest and the train RMSE is 3.2198. Again, RMSE performance results are shown in Figure 8.

5 Models for Classification

Having looked at price prediction, we now look at predicting the up or down movement in returns. For this we use Logistic Regression, SVM and K Nearest Neighbors for our analysis. We label up movement as '1' and down movement as '-1'.

First, we fit Logistic Regression on our train dataset as it is easy to implement. It will try to learn the relationship of the predictors to the given set of labeled data. We set threshold of the probability of up movement as 0.3 and

observed the test accuracy of 0.4936 which is lesser than 0.5.

Second, we fit SVM with linear kernel on our train data. We hope that the algorithm will find the best data separating hyperplane used as the decision boundary. We see that the train accuracy is 0.57988 and the test accuracy is 0.4978. Both the models did not perform well.

Next, we use K nearest neighbors to classify the daily movements. The algorithm will classify new data based on its neighboring data points. The test accuracy is 0.9873 which is the best among all the classifiers.

6 Summary and Conclusions

The tables summarizes the results of Regression models. As we can see, Elastic Net performs the best on test data among all Linear models while Random Forest performs the best overall.

Models	Train RMSE	Test RMSE
Linear Models		
Ordinary Least Squares	3.3252	4.0689
Ridge Regression	3.4067	3.4723
Lasso Regression	3.3904	3.4629
Elastic Net Regression	3.3848	3.4587
Non Linear Models		
Random Forest	2.3359	3.4657
XGBoost	3.2198	3.4646

Figure 8: Performance Comparison for Regression Models

All of these models barring the ordinary least squares model performed better than the baseline model of ARIMA (0,1) GARCH(1,1). The test RMSE obtained from the baseline model is 3.4784 which is greater than the other models and can be seen from Figure 8.

Models	Train Accuracy	Test Accuracy
Logistic Regression (with threshold=0.3)	0.5695	0.4936
SVM	0.5798	0.4978
K- nearest neighbor	-	0.9873

Figure 9: Performance Comparison for Classification Models

We see that the bitcoin daily returns are ill explained by macroeconomic factors but explained well by bitcoin related features like hash rate and lagged returns. Also, the Random Forest model along with XGBoost suggested that Google trends was a major factor in the price movement. Additionally, in absence of missing values, we expect the sentiment score to be a valuable predictor in predicting bitcoin returns.

7 Future Works

This study only considered Bitcoin due to lack of availability of data for other cryptocurrencies. As an extension, we can try to analyse the correlation between different currencies so that we can use the predicted price for Bitcoin scaled by the correlation factor as a proxy for prices of other cryptocurrencies. Another improvement could be

using sliding validation window for hyperparameters tuning rather than fixed validation split. This would only be feasible if we have more data that is only after a few years, when enough data is available for crytpocurrencies market. We can also build regime based models to account for high volatility.

Bibliography

1. Shahbazi, Zeinab, and Yung-Cheol Byun. 2022. "Knowledge Discovery on Cryptocurrency Exchange Rate Prediction Using Machine Learning Pipelines" Sensors 22, no. 5: 1740.

2. Cortez, K.; Rodríguez-García, M.d.P.; Mongrut, S. Exchange Market Liquidity Prediction with the K-Nearest Neighbor Approach: Crypto vs. Fiat Currencies. Mathematics 2021, 9, 56.

3. Vo, A.; Yost-Bremm, C. A high-frequency algorithmic trading strategy for cryptocurrency. J. Comput. Inf. Syst. 2020, 60, 555–568.

4. Tan, X.; Kashef, R. Predicting the closing price of cryptocurrencies: A comparative study. In Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems, Dubai, United Arab Emirates, 2–5 December 2019; pp. 1–5.

5. Azari, A. Bitcoin price prediction: An ARIMA approach.

6. Sun Yin, H.H.; Langenheldt, K.; Harlev, M.; Mukkamala, R.R.; Vatrapu, R. Regulating cryptocurrencies: A supervised machine learning approach to de-anonymizing the bitcoin blockchain. J. Manag. Inf. Syst. 2019, 36, 37–73.

7. 41. Saad, M.; Choi, J.; Nyang, D.; Kim, J.; Mohaisen, A. Toward characterizing blockchain-based cryptocurrencies for highly accurate predictions. IEEE Syst. J. 2019, 14, 321–332.

8. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of
the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Franisco, CA, USA, 13–17 August 20 pp. 785–794.

9. Mallqui, D.C.; Fernandes, R.A. Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques. Appl. Soft Comput. 2019, 75, 596–606.