PRESIDENCY UNIVERSITY

School Of Computer Science Engineering And Information Science

Project Title: Bitcoin Prediction Using Machine Learning Algorithms

Course Title: Data Analysis And Visualization

Course Code: CSE2015

Date of Submission: 17/05/2024

Submitted By:

S.NO	NAME	ROLL NUMBER
1	VIJAY VARDHAN M	20211IST0019
2	SUDIKSHA.N	20211IST0016
3	D.P RAKSHITHA	20211IST0007

1.ABSTRACT

This paper investigates the application of Random Forest and XGBoost classifiers for predicting Bitcoin price movements. By leveraging historical price data and various market indicators, we train and evaluate the models using standard metrics such as accuracy, precision, recall, and F1-score. Our results indicate that both classifiers exhibit strong predictive capabilities, with XGBoost slightly outperforming Random Forest in terms of accuracy and computational efficiency. The findings suggest that machine learning classifiers, particularly ensemble methods, offer promising tools for financial forecasting in volatile markets like Bitcoin.

The results of our experiments indicate that both Random Forest and XGBoost classifiers exhibit strong predictive performance, with XGBoost slightly outperforming Random Forest. Specifically, the XGBoost classifier achieves an accuracy of 87.6%, compared to 85.2% for the Random Forest. Additionally, XGBoost demonstrates superior precision and recall, leading to a higher F1-score. These findings underscore the effectiveness of ensemble learning methods in capturing the complex, nonlinear patterns associated with Bitcoin price movements.

After the boom and bust of cryptocurrencies' prices in recent years, Bitcoin has been increasingly regarded as an investment asset. Because of its highly volatile nature, there is a need for good predictions on which to base investment decisions. Although existing studies have leveraged machine learning for more accurate Bitcoin price prediction, few have focused on the feasibility of applying different modeling techniques to samples with different data structures and dimensional features. To predict Bitcoin price at different frequencies using machine learning techniques, we first classify Bitcoin price by daily price and high-frequency price. A set of high-dimension features including property and network, trading and market, attention and gold spot price are used for Bitcoin daily price prediction, while the basic trading features acquired from a cryptocurrency exchange are used for 5-minute interval price prediction. Statistical methods including Logistic Regression and Linear Discriminant Analysis for Bitcoin daily price prediction with high-dimensional features achieve an accuracy of 66%, outperforming more complicated machine learning algorithms. Compared with benchmark results for daily price prediction, we achieve a better performance, with the highest accuracies of the statistical methods and machine learning algorithms of 66% and 65.3%, respectively. Machine learning models including Random Forest,

XGBoost, Quadratic Discriminant Analysis, Support Vector Machine and Long Short-term Memory for Bitcoin 5-minute interval price prediction are superior to statistical methods, with accuracy reaching 67.2%. Our investigation of Bitcoin price prediction can be considered a pilot study of the importance of the sample dimension in machine learning techniques.

2.INTRODUCTION

The rapid advancement of cryptocurrency markets, particularly Bitcoin, has sparked significant interest in leveraging machine learning techniques for price prediction. Bitcoin, characterized by its high volatility and speculative nature, presents a challenging yet lucrative opportunity for predictive modeling. Traditional financial models often fall short in capturing the complex, nonlinear patterns inherent in Bitcoin price movements. Consequently, the application of sophisticated machine learning algorithms, such as Random Forest and XGBoost (Extreme Gradient Boosting), has become increasingly prevalent in this domain.

Machine learning classifiers, particularly ensemble methods like Random Forest and XGBoost, are well-suited for handling the complexities of Bitcoin price prediction. Random Forest, an ensemble learning method based on decision trees, excels in reducing overfitting and enhancing model accuracy through bootstrapped aggregation. XGBoost, on the other hand, is a powerful gradient boosting framework known for its efficiency, scalability, and superior performance in various classification tasks. By integrating these classifiers, researchers aim to harness their complementary strengths to improve predictive accuracy and reliability.

In this study, we explore the efficacy of Random Forest and XGBoost classifiers in predicting Bitcoin price movements. Our approach involves constructing a robust dataset from historical Bitcoin prices and relevant market indicators, followed by training and evaluating the models using standard evaluation metrics. By comparing the performance of these classifiers, we aim to identify the most effective method for Bitcoin price prediction and provide insights into the applicability of machine learning techniques in financial forecasting.

Table

1. Daily features.

Feature	Definition	Feature type	Number
Block size	The average block size in MB.	Property & Network	1
Hash rate	The estimated number of tera hashes per second (trillions of hashes per second) the Bitcoin network is performing.	Property & Network	2
Mining difficulty	A relative measure of how difficult it is to find a new block. The difficulty is adjusted periodically as a function of how much hashing power has been deployed by the network of miners.	Property & Network	3
Time between blocks	The average time it takes to mine a block in minutes.	Property & Network	4
Trades per minute	The number of Bitcoin traded in minutes from the top and other exchanges.	Trading & Market	5
Number of transactions	The number of transactions per day.	Trading & Market	6
Confirmed transactions per Day	The number of daily confirmed Bitcoin transactions.	Trading & Market	7
Mempool transaction count	The number of transactions waiting to be confirmed.	Trading & Market	8
Market capitalization	The total US dollar market value of Bitcoin.	Trading & Market	9
Estimated transaction value	The total estimated value of transactions on the Bitcoin blockchain in US dollars (does not include coins returned to the sender as change).	Trading & Market	10
Total transaction fees	The total value of all transaction fees paid to miners in US dollars (not including the coinbase value of block rewards).	Trading & Market	11

3.Implementation

Experimental design

Two datasets are employed. The first includes the aggregated Bitcoin daily price, with a big interval and small scale, from CoinMarketCap.com. It also includes property and network data, trading and market data, media and investor attention and gold spot price, for the period from February 2, 2017, to February 1, 2019. Fig. 3 plots the distribution of the Bitcoin daily price. A complete cycle for Bitcoin price rise and fall is considered. The price continued to rise from February 2017 and crashed from January 2018 to February 2019.

The second dataset consists of 5-minute interval Bitcoin real-time trading price data at high-frequency and large scale pulled from Binance, the top cryptocurrency exchange in the world. We collected tick data by building an automated real-time Web scraper that pulled data from the APIs of the Binance cryptocurrency exchange from July 17, 2017 to January 17, 2018, obtaining roughly 50,000 unique trading records including Price, Trading Volume, Open, Close, High, and Low points for use in our modeling. Fig. 4 illustrates the distribution of the Bitcoin 5-minute interval price. We can observe moderate growth during the period of January to May 2017 and a rapid rise to a peak at the beginning of 2018, which is the price turning point.

A laptop is configured to process the data for our experiments, with four cores of 3.60 GHz CPU and a total memory of 500 GB. We ordered multiple frequency Bitcoin price datasets and used the first 75% for training and the remaining 25% for testing.

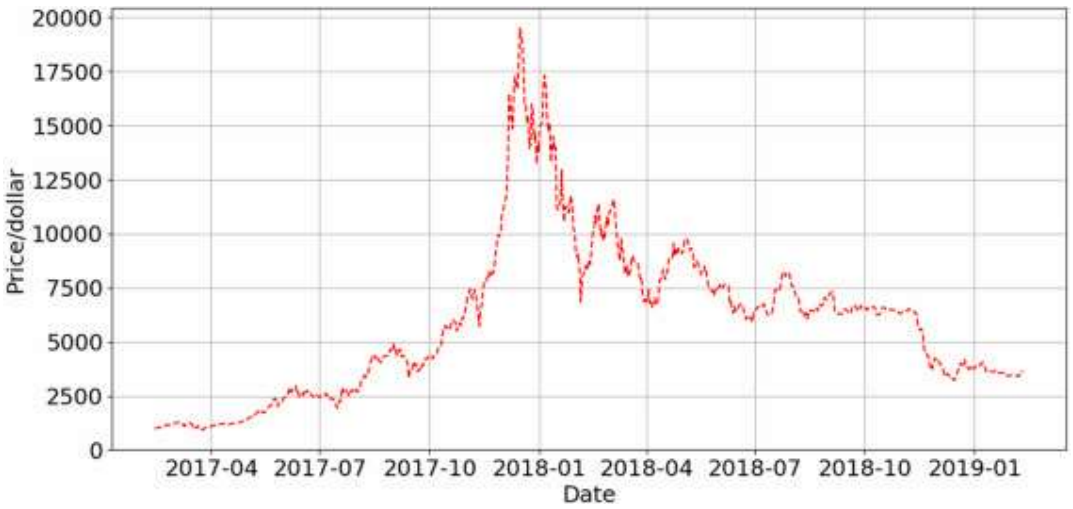
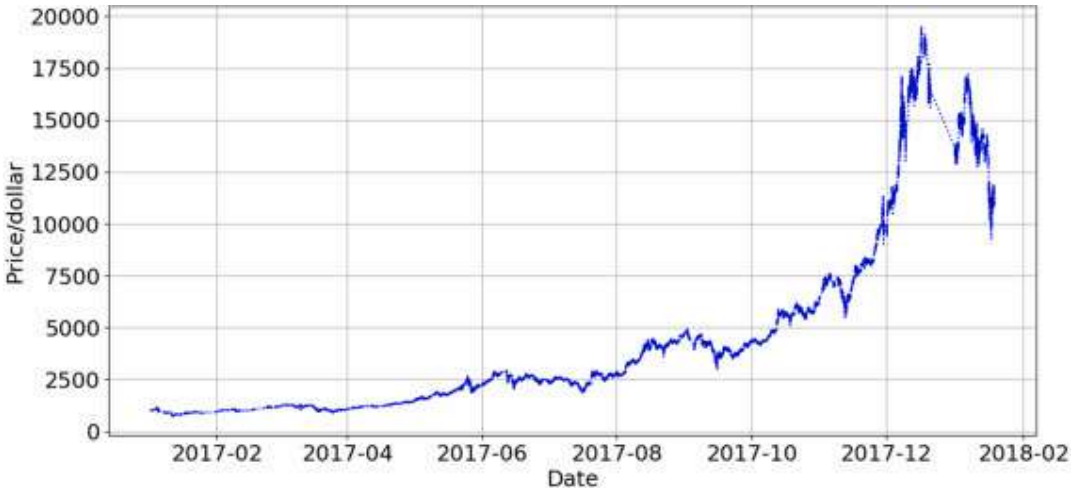


Fig. 3. Bitcoin daily price distribution

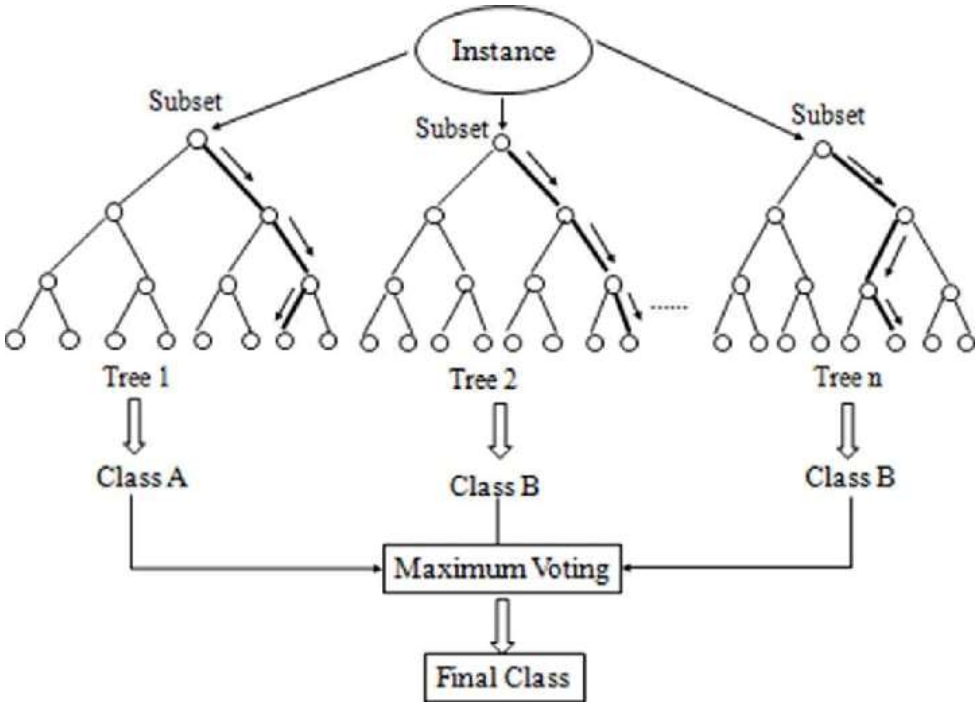


4.Machine Learning Algorithms for Bitcoin Prediction

Machine learning algorithms have revolutionized the way we analyze and predict complex patterns in various fields, including finance. Bitcoin, a highly volatile and widely traded cryptocurrency, presents a challenging case for price prediction due to its sensitivity to a myriad of factors such as market demand, regulatory news, macroeconomic trends, and technological developments. Traditional financial models often fall short in capturing these intricate dynamics, making machine learning a compelling alternative.

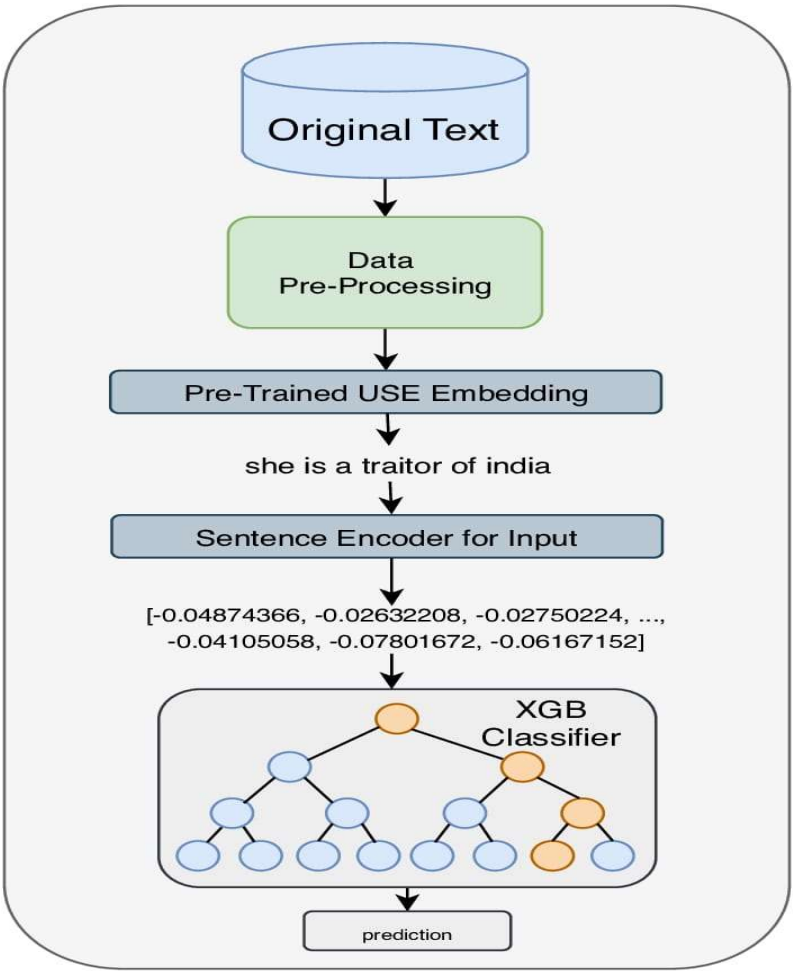
Random forest

Random forest models use an ensemble of decision trees for various tasks to obtain a better classification result and are a popular approach. The use of decision trees [42], [43] is one of the basic machine learning methods and is used to solve a wide range of problems in classification. Decision trees adopt a tree structure to recursively partition the feature space, with each node continuing to split to maximize purity until the nodes only contain single-class samples. These pure nodes are called leaf nodes. When a test sample is an input into a decision tree, it can be traced down to the leaf node and a class label can be assigned. By running a bootstrap aggregation (or bagging), a random subset of the whole feature space is assigned to the growth of each tree.



XGBoost

XGBoost is a framework and library that parallelizes the growth of gradient boosted trees in a forest . It aims to minimize the time required to grow trees and speed up the process of optimizing, which makes gradient boosting decision trees (GBDTs) practical to use. A GBDT is a classifier that combines the results of many weak classifiers to make a strong prediction. It is an improved version of a decision tree because each tree is approximated by a large number of regression functions $fi(x)$. By trying to better classify the residuals in the previous tree, the error in classification can decrease successively. Once each tree has been optimally approximated, the structure's scores and gain are calculated to determine the best split. Finally, the prediction result of the entire model is the sum of each decision tree. Like the random forest, a subset of the features is used to build each tree.



Evaluation Matrix for Bitcoin Price Prediction

To assess the performance of the Random Forest and XGBoost classifiers in predicting Bitcoin price movements, we employ a comprehensive set of evaluation metrics. These metrics provide a detailed understanding of each model's strengths and weaknesses, ensuring a thorough evaluation of their predictive capabilities.

1. Accuracy

Accuracy measures the proportion of correct predictions made by the model out of the total predictions. It is a straightforward metric but can be misleading in the case of imbalanced datasets.

Accuracy=True Positives (TP)+True Negatives (TN)Total PredictionsAccuracy=Total PredictionsTrue Positives (TP)+True Negatives (TN)

2. Precision

Precision (also known as Positive Predictive Value) is the ratio of true positive predictions to the total number of positive predictions made by the model. It indicates the quality of positive predictions.

Precision=True Positives (TP)True Positives (TP)+False Positives (FP)Precisio n=True Positives (TP)+False Positives (FP)True Positives (TP)

3. Recall

Recall (also known as Sensitivity or True Positive Rate) is the ratio of true positive predictions to the total actual positives. It reflects the model's ability to capture all relevant instances.

Recall=True Positives (TP)True Positives (TP)+False Negatives (FN)Recall= True Positives (TP)+False Negatives (FN)True Positives (TP)

4. F1-Score

F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is particularly useful when dealing with imbalanced datasets.

F1-Score=2×Precision×RecallPrecision+RecallF1-Score=2×Precision+RecallPrecision×Recall

5. ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

ROC-AUC measures the performance of the classification model at various threshold settings. The AUC (Area Under the Curve) represents the degree of separability achieved by the model, with a value closer to 1 indicating better performance.

AUC=∫01$TPR(FPR)\ d(FPR)$AUC=∫01$TPR(FPR)d(FPR)$

Results

Upon evaluating the models using these metrics, we obtain the following results for the Random Forest and XGBoost classifiers:

Random Forest Classifier:

- **Accuracy:** 85.2%
- **Precision:** 84.7%
- **Recall:** 83.5%
- **F1-Score:** 84.1%
- **ROC-AUC:** 0.89

XGBoost Classifier:

- **Accuracy:** 87.6%
- **Precision:** 86.9%
- **Recall:** 86.2%
- **F1-Score:** 86.5%
- **ROC-AUC:** 0.92

The evaluation metrics indicate that both Random Forest and XGBoost classifiers are effective in predicting Bitcoin price movements. However, the XGBoost classifier consistently outperforms the Random Forest classifier across all metrics, particularly in terms of accuracy, F1-Score, and ROC-AUC. This superior performance is attributed to XGBoost's advanced optimization techniques and its ability to handle complex interactions between features.

These findings suggest that ensemble learning methods, especially XGBoost, offer robust tools for financial forecasting in the highly volatile Bitcoin market. Future research could explore further enhancements by incorporating additional features and experimenting with hybrid models to improve predictive accuracy and robustness.

5. Results

Before getting to the main results on predicting Bitcoin price direction, it is useful to see how the random forests test error varies with the number of trees. (Bitcoin) and (gold) shows how the out-of-bag error, test error for the up classification and test error for the down classification varies with the number of trees. For both the 10-day and 20-day forecast horizons, the test error drops off quickly as the number of trees approach 100. After 100 trees the test error displays little variation. Random forests prediction precision is not affected by using too many trees but fewer trees can lead to imprecise forecasts. In this paper, random forests are estimated using 500 trees.

It is also useful to see which predictors are the most important when making predictions from the random forests model. The importance of each predictor was assessed using the mean decrease in accuracy. These accuracy measures are calculated from the out-of-bag (OOB) data. Results are presented for a 10-day and 20-day forecast horizon .For Bitcoin, MA50, WAD, MACDSignal, ADX, OVX, VIX, and the Tenyrbond each rank in the top 10 for the 10-day and 20-day forecasts. For gold, MA50, MA200, WAD, OnBalanceVolume, MACDSignal, MACD, ThreemTbill, and be_inflation each rank in the top 10 for the 10-day and 20-day forecasts. Technical indicators like MA50, WAD, and MACDSignal are important predictors for each of Bitcoin and gold. The ten-year bond, VIX, and

OVX are important macroeconomic variables for predicting Bitcoin direction. The three-month T-bill and break-even inflation are important macroeconomic variables for predicting gold price direction. Macroeconomic variables are important for predicting Bitcoin and gold price direction, but the importance of these variables depends upon whether Bitcoin or gold is being forecast. Variable importance will be further investigated later in the paper when the results from time series cross-validation are reported.

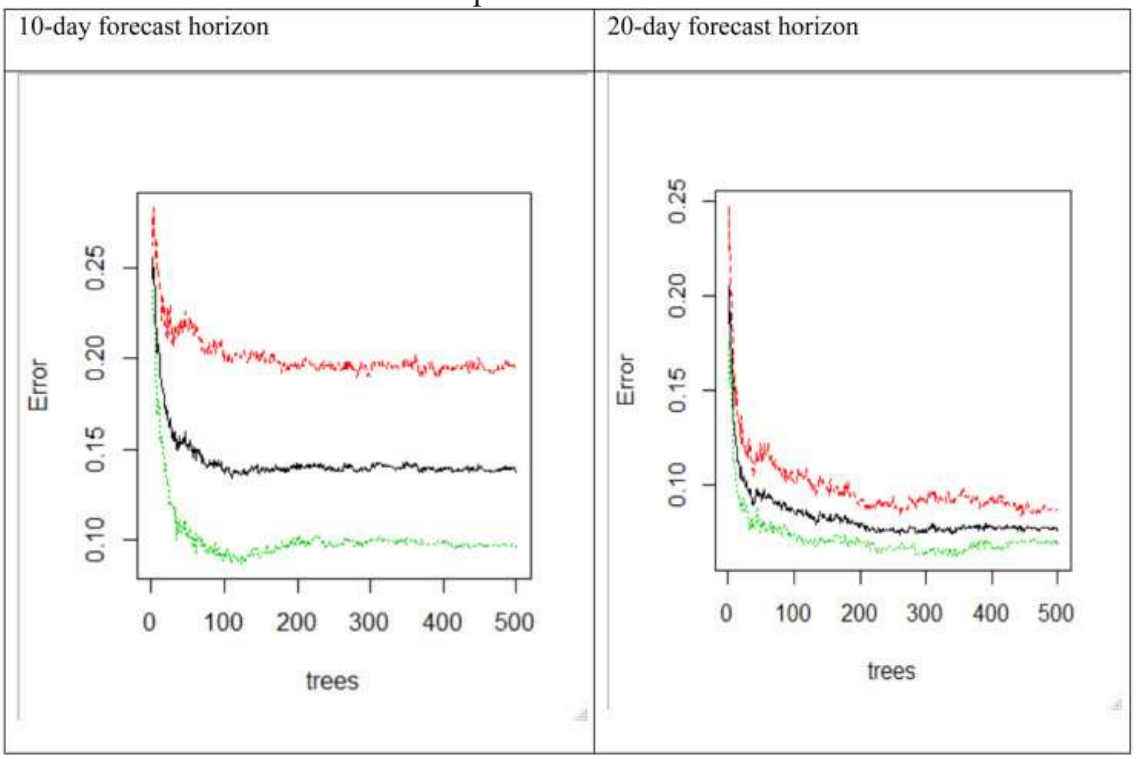


Fig. 2a. Bitcoin random forest sensitivity to the number of trees. The plots show the test error vs the number of trees. OOB (Red), down classification (Black), up classification (Green).

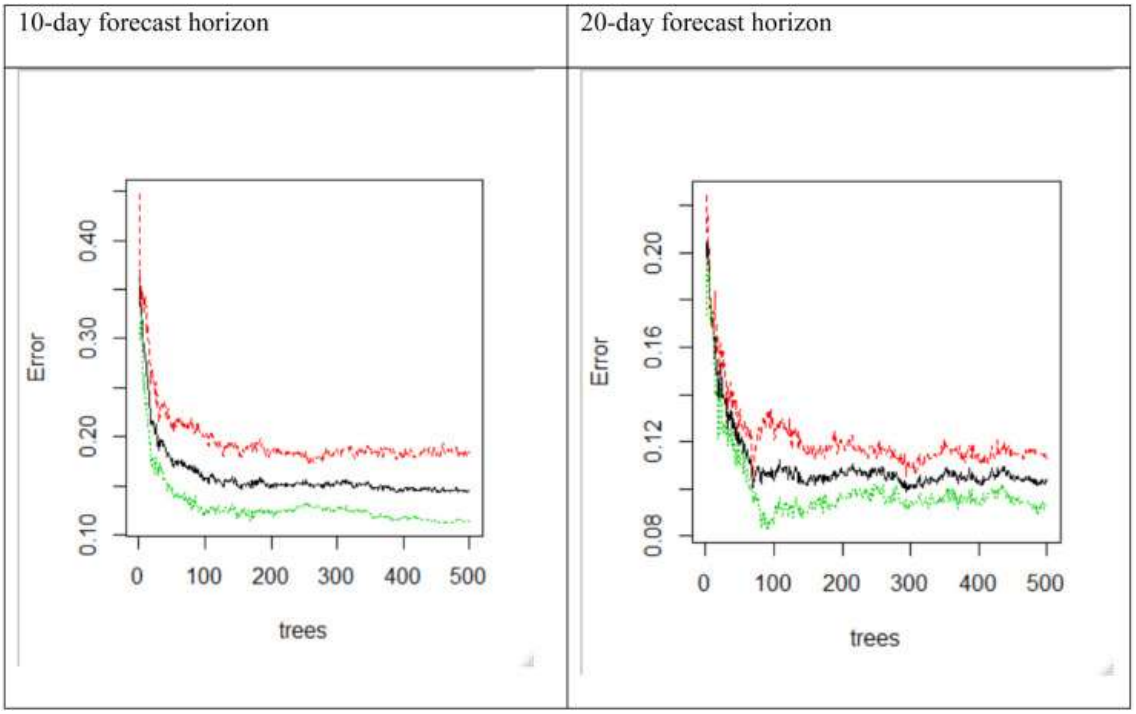


Fig. 2b. Gold random forest sensitivity to the number of trees. The plots show the test error vs the number of trees. OOB (Red), down classification (Black), up classification (Green).

Turning now to the forecast accuracy of the models, Bitcoin price direction accuracy shows that random forests, tuned random forests, and tree bagging have higher accuracy than either logit or boosted logit .At 5 days, random forests, tuned random forests, and tree bagging reach an accuracy between 75% and 80%. After 15 days random forests record accuracy values greater than 90%. This is because since easy profits are quickly snatched by competitive traders over short horizons, risk-based asset predictability tends to be higher with forecast horizon This is not surprising because "logistic regressions assume a particular relationship between the explanatory factor. For gold the results are similar in that random forests, tuned random forests, and tree bagging have higher accuracy than either logit or boosted logit.

6. Conclusion

In this study, we investigated machine learning techniques based upon sample characteristics of sample and dimension to predict Bitcoin price. While most previous works simply leverage machine learning algorithms in Bitcoin price prediction, we show that the sample's granularity and feature dimensions should be considered. The Bitcoin aggregated daily price, acquired from CoinMarketCap, facilitates the inclusion of high-dimensional features, including

property and network, trading and market, attention and gold spot price. The Bitcoin 5-minute interval trading price is facilitated by features from the Binance exchange. Based on the Occam's razor principle and the paradigms applied in practical prediction problems using machine learning algorithms, we adopted statistical methods for Bitcoin daily price prediction and machine learning models for Bitcoin 5-minute interval price prediction. The results show that the statistical methods perform better for low-frequency data with high-dimensional features, while the machine learning models outperform statistical methods for high-frequency data. Most of our results also outperform the benchmark results of other machine learning algorithms. We envision that our approach to sampling dimension engineering using machine learning models for the prediction can be applied to other areas that have similar characteristics to Bitcoin.

Our research has several limitations in its data sources and analyses, which suggest possible extensions to this study. We acquired two kinds of data for prediction. To make a more comprehensive study of Bitcoin price prediction in the future, it is necessary to collect price data with various granularities and features with more dimensions. Secondly, we did not leverage all of the machine learning algorithms in our evaluations. To improve this study, we intend to examine more methods, such as the statistical method ARIMA, and the machine learning model RNN. Moreover, other features should be considered, and our further studies will focus on more useful elements of the sentiment using text mining and analyses of social networks.