Predicting the Price of Cryptocurrency using the Statistical Learning

Yusi Li[a]

[a]*Department of Financial Mathematics, University of Southern California, Los Angeles, USA*

* Yusi Li: yusili@usc.com

Predicting the Price of Cryptocurrency using the Statistical Learning

The goal of this report is to ascertain with what accuracy the direction of cryptocurrency price in USD can be predicted. By collecting the data of top 20 ranking cryptocurrencies from Apr. 28th 2013 to June 5th 2018 together, the report uses the statistical learning methods to predict the price trend and price of cryptocurrency next day. The best method to predict the next day trend of cryptocurrency is quadratic discriminant analysis (QDA), which produces the probability that the prediction is correct to be 54.13%. The best method to predict the next day price of cryptocurrency is random forests model, which produces the MSE to be 14566.2.

Keywords: cryptocurrency; statistical learning

Subject classification codes: G17, C15

I. Introduction

There are disadvantages of the traditional physical cash and digital cash. Buyers and sellers have to be physically present at the same location in order to trade. Digital cash, however, can be copied any number of times at negligible cost. If cash data files can be copied and the duplicates used as currency, they cannot serve as a payment instrument, which refer to the "double spending problem". To solve this problem, centralized payment systems appears, but they require trust, which increase the cost. Furthermore, centralized systems are vulnerable to hacker attacks, technical failures, and malicious governments that can easily interfere and confiscate funds. (Berentsen & Schar 2018)

The creators' original motivation behind Bitcoin was to develop a cash-like payment system that permitted electronic transactions but that also included may of the advantageous characteristics of physical cash. The purely peer-to-peer version of electronic cash would allow online payments to be sent directly from one party to another without going through a financial institution. (Nakamoto 2013) This novel technology allows us to store and transfer a monetary unit without the need for a central authority, similar to cash by inventing the Bitcoin Blockchain. Its other fundamental characteristics are:

being decentralized, and having a fixed total number of coins: 21 million, with more than 16 million already in circulation. Bitcoin is still very young, especially considering the fact that the last coin is to be mined around year 2140.

Bitcoin, as one of the largest market capitalization, uses peer-to-peer technology to operate with no central authority or banks; managing transactions and the issuing of bitcoins is carried out collectively by the network. Bitcoin is open-source; its design is public, nobody owns or controls Bitcoin and everyone can take part. Through many of its unique properties, Bitcoin allows exciting uses that could not be covered by any previous payment system. However, price volatility and scaling issues frequently raise concerns about the suitability of Bitcoin as a payment instrument.

The past research of predicting the price of cryptocurrency including the Bayesian optimized recurrent neural network (RNN), Long Short Term Memory network, time series ARIMA model and Monte Carlo approach. According to the past research, the LSTM achieves the highest classification accuracy of 52% which outperforming ARIMA. (McNally, Roche & Caton 2018) The accuracy of statistical inferences of Monte Carlo simulation with 104 geometric fractional Brownian motion is 10%. (Tarnopolski 2017) Both deep learning models are benchmarked on both a GPU and a CPU with the training time on the GPU outperforming the CPU implementation by 67.7%. (McNally, Roche & Caton 2018) This report use the statistical learning method to predict the trend and price of cryptocurrency. For the prediction of price trend of cryptocurrency, logistic regression, linear discriminant analysis, quadratic discriminant analysis. The methods of price prediction are ridge regression, the lasso regression, principal components regression, partial least squares and decision tree model.

The article is organized in the following manner. In section 2, the data set is described. In section 3, the methods and analyst result are discussed. The section 4 presents the conclusion and discussion.

II. Data

In this section, the article describes and summarizes the descriptive statistics of the data used in our study. The data set of this article comes from https://www.kaggle.com. The time range from Apr. 28th 2013 to June 5th 2018. There are around 1645 different kinds of cryptocurrencies in the market. This article selects top 20 cryptocurrencies in the market, which are Bitcoin, Ethereum, Ripple, Bitcoin Cash, EOS, Litecoin, Stellar, Cardano, IOTA, TRON, NEO, Monero, Dash, Tether, NEM, VeChain, Binance Coin, Ethereum Classic, Ontology, Qtum.

According to the available data, the characteristics of cryptocurrencies include date, the rank, the open price, the close day, the highest price of the day, the lowest price of the day, the volume, the market capitalization, the spread of USD. In particular, the closed ratio of the price is get by using the difference of highest price and lowest price divided by the difference of closed price and open price. The classification response is the price trend If the price of next day increase, then the price trend is 1. If the price of next day increase or stay static, the price trend is 0. The regression response is the price of next day.

Based on 17,254 sample, around 45.99% of the price decrease or stay static and 54.01% of the price increase. The price range is from 0 to 19497.4, the mean price is 256.714. The range of trade volume is from 0 to $2.384*10^{+10}$, more than 25% of the daily trade volume is over $2.029*10^{+15}$. The market capitalization is from 0 to $3.260*10^{+11}$, more than 25% of the market capitalization is over $3.640*10^{+7}$ and more than 75% of the market capitalization is over $3.373*10^{+9}$. The range of spread of

USD is from 0 to 4110.40, more than 75% of the spread of USD is less than 2.62. The descriptive statistics is in Figure 1, Figure 2 and Figure 3.

III. Empirical results

Prediction the price trend - Classification

The first part of this section is to predict the price trend of the cryptocurrency. The response is trend of next day. The predictors are closed ratio, volume, market capitalization, current rank and spread of USD. The data set are randomly divided by two equal length part to be the training data and testing data.

By using the logistic regression, the probability that the prediction is correct is around 46.76%. The predictors close ratio and the current rank is 100% statistically significant, volume is 99% statistically significant. Then considering the Linear Discriminant Analysis Method, we model the distribution of the predictors next day trend separately in each of the response classes, and then use Bayes' theorem to flip these around into estimates for $Pr(Y = k|X = x)$. When these distributions are assumed to be normal, it turns out that the model is very similar in form to logistic regression. The probability that the prediction is correct is around 53.94%. Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. However, QDA assumes that each class has its own covariance matrix and the quantity x appears as quadratic function. The probability that the prediction is correct is around 54.13%. Then we use R=100 bootstrap to calculate the estimated standard errors of logistic regression model and LDA model. The results are shown in Table 1 and Table 2.

Prediction the price - Regression

The second part of this section is to predict the price of cryptocurrency. The response is price of next day. The predictors are closed ratio, volume, market capitalization, current rank and spread of USD. The data set are randomly divided by two equal length part to be the training data and testing data. Mean square estimate, which is a risk function measures the average of the squares of the errors, that is, the average squared difference between the estimated values and what is estimated.

By using the linear regression, the mean square estimate is 171067.7. Then we use R=100 bootstrap to calculate the estimated standard errors of linear regression model. The result is shown in Table 3.

By adding the slightly different quantity to minimize we can estimate the coefficients by using the Ridge Regression and Lasso Model. The Ridge Regression by choosing the best lamda as 0.4977024 produces the mean square estimate as 171020. The Lasso Model produce the mean square estimate to be 1115238. Principal Components Analysis is a popular approach for deriving a low-dimensional set of features from a large set of variables.

Another method is to use the dimension reduction technique for regression. The principal components regression (PCR) approach involves constructing the first M principal components, and then using these components as the predictors in a linear regression model that is fit using least squares. We plot the cross-validation MSE as Figure 4. From the graph, we can choose M=3. Under this condition, The PCR Model produce the mean square estimate to be 225379.5. The PCR approach involves identifying linear combinations, or directions, that best represent the predictors. These directions are identified in an unsupervised way, since the response is not used to help determine the principal component directions. Consequently, PCR suffers from a drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response, which is the reason for large MSE. As a result, we use the partial least squares (PLS)

approach to find directions that help explain both the response and the predictors. We plot the cross-validation MSE as Figure 5. From the graph, we can choose M=3. Under this condition, The PLS Model produce the mean square estimate to be 180224.9.

The last part of this section describes the tree-based methods. These involve stratifying or segmenting the predictor space into a number of simple regions. In order to make a prediction for a given observation, we typically use the mean or the mode of the training observations in the region to which it belongs. The result of decision tree model is in Figure 6. The Decision Tree Model produce the mean square estimate to be 125205.1. Then we use the bootstrap to improve the decision tree model. To apply bagging to regression trees, we simply construct B regression trees using B bootstrapped training sets, and average the resulting predictions. These trees are grown deep, and are not pruned. Hence each individual tree has high variance, but low bias. Averaging these B trees reduces the variance. The Bagging Approach produce the mean square estimate to be 15179.16. The importance of the predictors from high to low are market capitalization, current rank, USD spread, closed ratio, volume as in Figure 7. Random forests provide an improvement over bagged trees by way of a random forest small tweak that decorrelates the trees. In building a random forest, at each split in the tree, the algorithm is not even allowed to consider a majority of the available predictors. The Random Forests Model produce the mean square estimate to be 14566.2. The importance of the predictors from high to low are market capitalization, current rank, USD spread, closed ratio, volume as in Figure 8.

IV. Conclusion

Statistical learning model can be used to predict both the trend and the price of cryptocurrency. Using the data set of top 20 cryptocurrency, the predictors include close ratio, volume, market capitalization, current rank, USD spread. To predict the trend of cryptocurrency, we use the logistic regression model,

linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), while the probability that the prediction is correct are 46.76%, 53.94%, 54.13%, which means that quadratic discriminant analysis (QDA) is the best model to predict the trend of cryptocurrency. However, the logistic regression model with the classification accuracy of 46.76% perform more poorly than the random guess. Compared with the past research results, both the linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA) outperform the best exist method, which is the LSTM with the 52% classification accuracy. To predict the price of cryptocurrency, we use the linear regression model, ridge regression model, lasso regression model, principal components regression (PCR), partial least squares (PLS), decision tree model, bagging tree model and random forests model. The MSE of these model are 171067.7, 171020, 1115238, 225379.5, 180224.9, 125205.1, 15179.16, 14566.2, which means the random forests model is the best. The prediction accuracy is shown in Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13.

V. Appendices

Table 1. Logistic Regression for Classification

	Estimate	Bootstrap bias	Bootstrap Std. Error
Intercept	-1.487e-01	3.302173e-03	4.657891e-02
Close ratio	3.273e-01	-3.739775e-03	5.003139e-02
Volume	1.070e-10	2.638400e-12	3.247059e-11
Market Capitalization	-2.089e-12	-9.725326e-14	1.827746e-12
Current Rank	-1.864e-02	-4.469762e-05	3.176184e-03
USD Spread	-4.124e-04	-3.682174e-06	2.401879e-04

Table 2. LDA Model for Classification

	Coefficients of Linear Discriminants	Bootstrap bias	Bootstrap Std. Error
Close ratio	2.043367e+00	-2.970730e-02	2.446229e-01
Volume	6.624087e-10	-1.099571e-11	2.194395e-10
Market Capitalization	-1.285158e-11	1.589362e-12	1.019458e-11
Current Rank	-1.162771e-01	3.652694e-03	1.332535e-02
USD Spread	-2.572756e-03	-2.770535e-04	1.317868e-03

Table 3. Linear Regression for Regression

	Estimate	Bootstrap bias	Bootstrap Std. Error
Intercept	-1.040e+02	-9.987036e-02	8.437514e+00
Close ratio	4.134e+01	-6.901082e-01	9.370032e+00
Volume	-1.690e-07	-2.168969e-09	2.354788e-08
Market Capitalization	4.513e-08	5.459902e-11	1.723465e-09
Current Rank	4.859e+00	4.605555e-02	5.513802e-01
USD Spread	2.849e+00	2.862915e-02	4.170947e-01

Figure 1. Data Description

```
    ranknow          volume             market          close_ratio        spread
 Min.   : 1.000   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.0000   Min.   :   0.00
 1st Qu.: 3.000   1st Qu.:2.029e+05   1st Qu.:3.640e+07   1st Qu.:0.2143   1st Qu.:   0.01
 Median : 8.000   Median :6.667e+06   Median :3.080e+08   Median :0.4762   Median :   0.14
 Mean   : 8.801   Mean   :2.933e+08   Mean   :6.286e+09   Mean   :0.4837   Mean   :  21.26
 3rd Qu.:13.000   3rd Qu.:9.358e+07   3rd Qu.:3.373e+09   3rd Qu.:0.7587   3rd Qu.:   2.62
 Max.   :20.000   Max.   :2.384e+10   Max.   :3.260e+11   Max.   :1.0000   Max.   :4110.40
 nextday_trend    nextday_close
 Min.   :0.0000   Min.   :    0.000
 1st Qu.:0.0000   1st Qu.:    0.223
 Median :0.0000   Median :    2.320
 Mean   :0.4599   Mean   :  256.714
 3rd Qu.:1.0000   3rd Qu.:   27.510
 Max.   :1.0000   Max.   :19497.400
```

Figure 2. Correlation matrix of data

	ranknow	volume	market	close_ratio	spread	nextday_trend	nextday_close
ranknow	1.00	-0.18	-0.30	-0.07	-0.18	-0.05	-0.25
volume	-0.18	1.00	0.89	0.05	0.82	0.02	0.84
market	-0.30	0.89	1.00	0.04	0.79	0.03	0.92
close_ratio	-0.07	0.05	0.04	1.00	0.03	0.05	0.05
spread	-0.18	0.82	0.79	0.03	1.00	0.02	0.84
nextday_trend	-0.05	0.02	0.03	0.05	0.02	1.00	0.04
nextday_close	-0.25	0.84	0.92	0.05	0.84	0.04	1.00

Figure 3. Scatterplot matrix of data

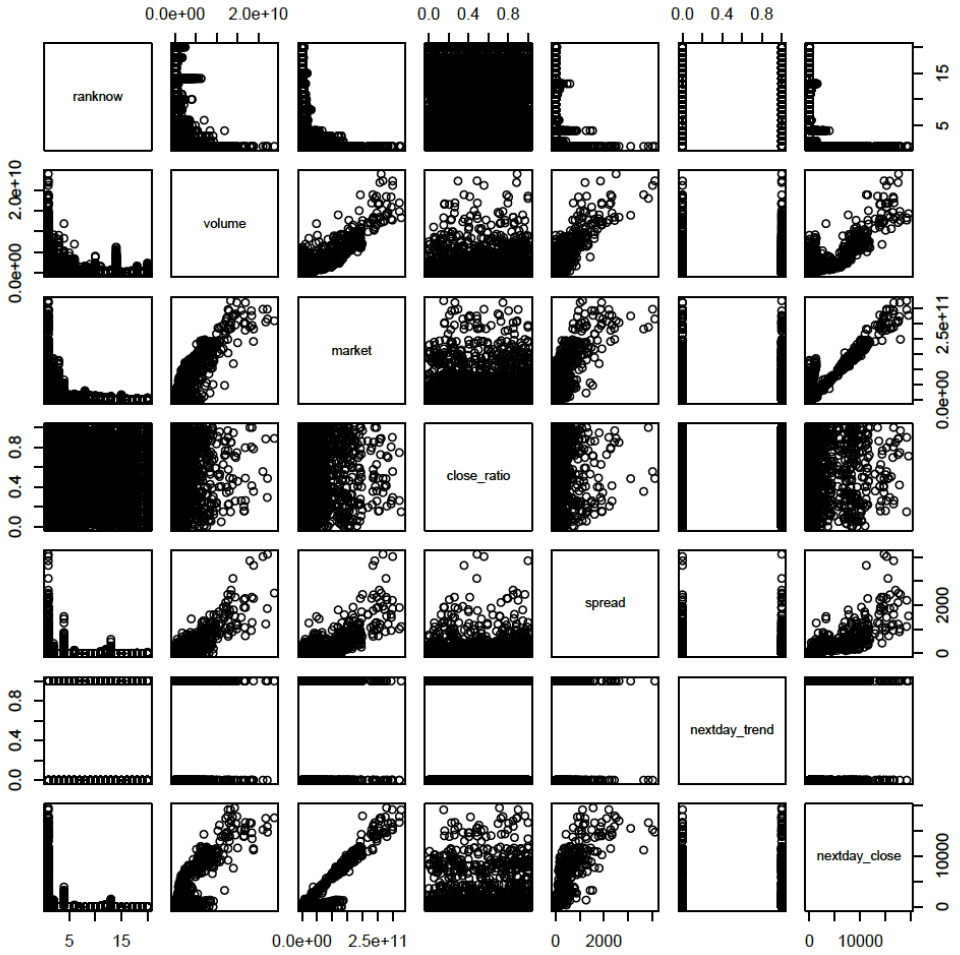


Figure 4. Cross-Validation MSE of PCR

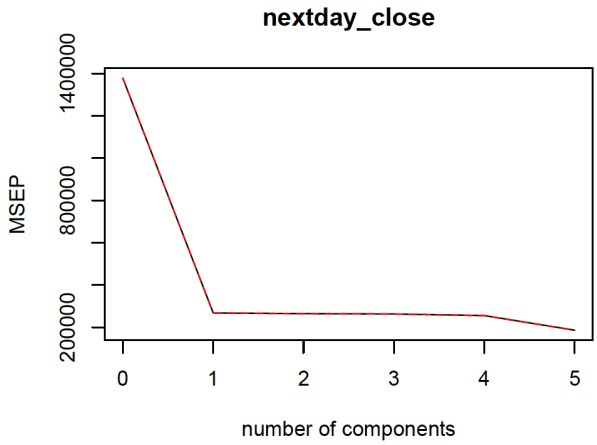


Figure 5. Cross-Validation MSE of PLS

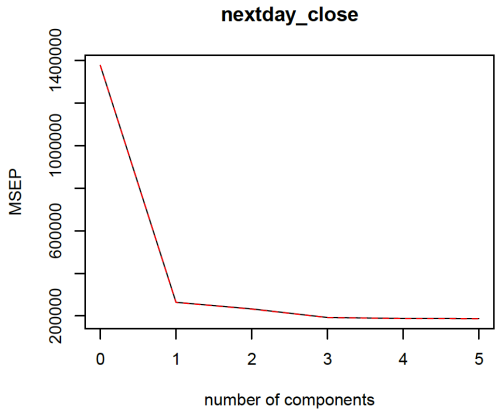


Figure 6. Decision Tree Model

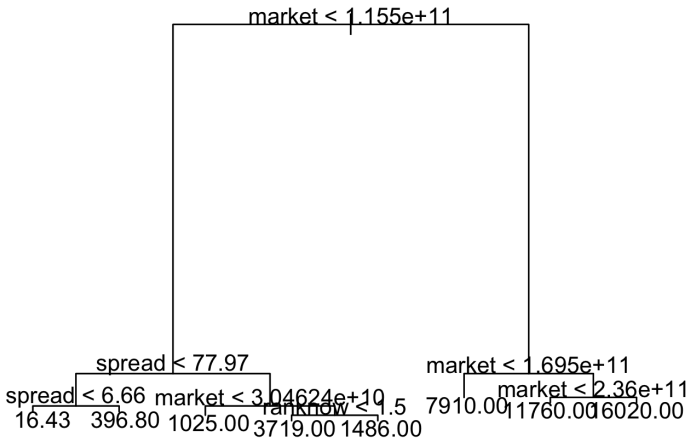


Figure 7. Importance of variables in Bagging Approach

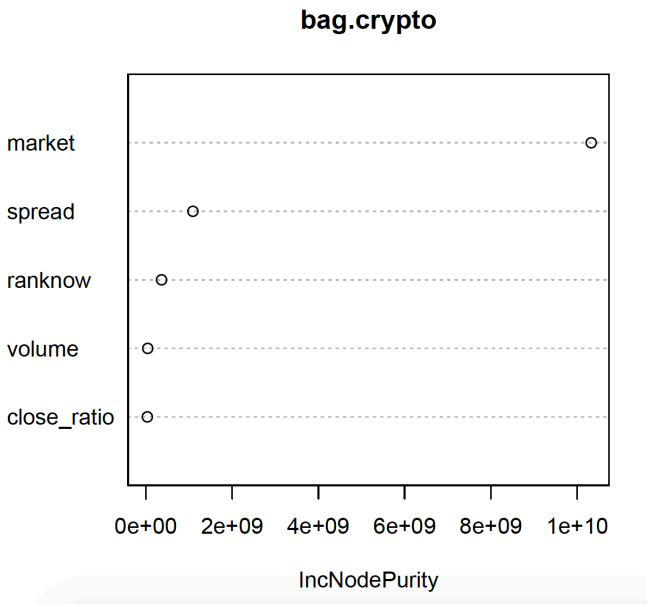


Figure 8. Importance of variables in Random Forest Tree Model

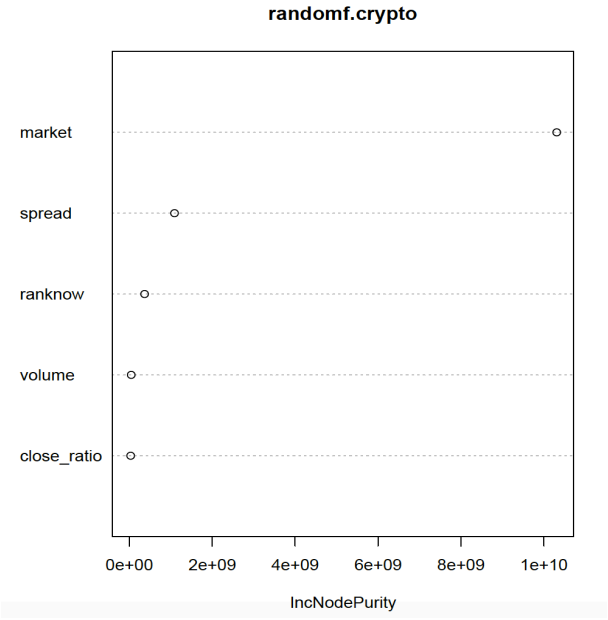


Figure 9. Prediction accuracy of Linear Regression

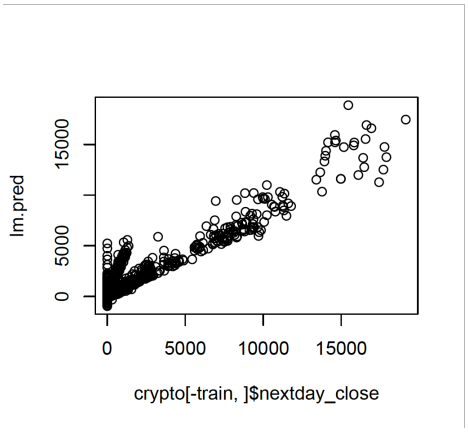


Figure 10. Prediction accuracy of Ridge Regression

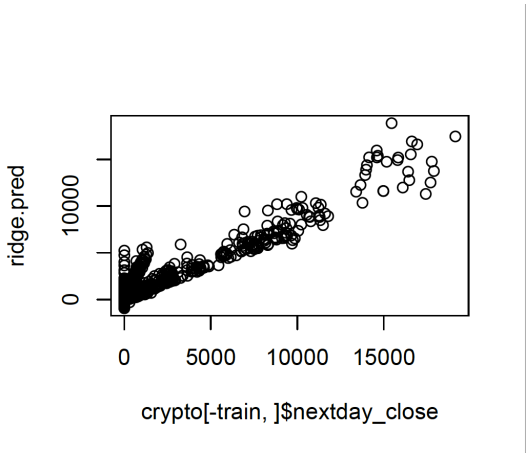


Figure 11. Prediction accuracy of Decision Tree Regression

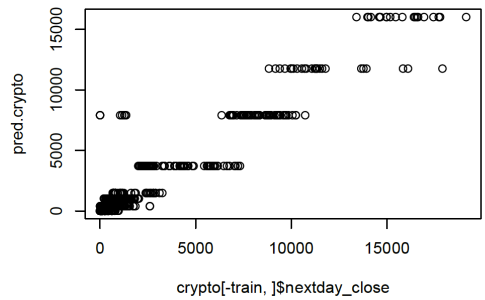


Figure 12. Prediction accuracy of Bagging Tree Model

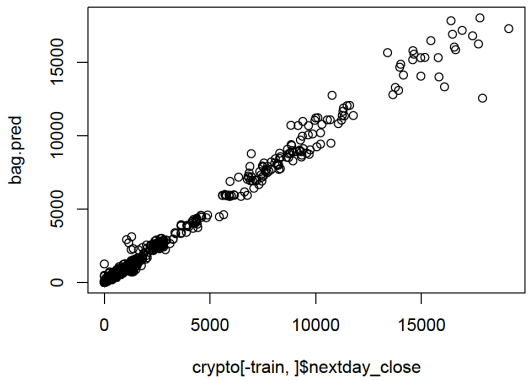
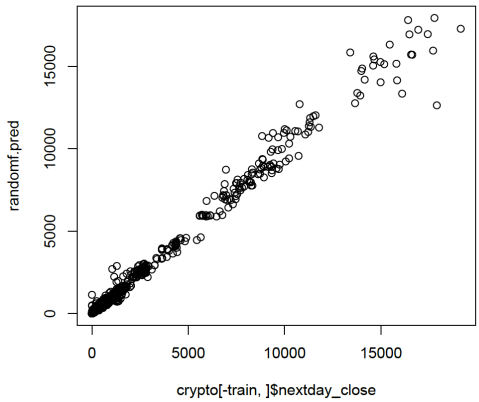


Figure 13. Prediction accuracy of Random Forests Model



References

Satoshi Nakamoto, "Bitcoin: A peer-to-peer Electronic Cash System" www.bitcoin.org.

Sean McNally, Jason Roche, Simon Caton, "Predicting the price of Bitcoin Using Machine Learning" 26th Euromicro International Conference on Parallel.

Mariusz Tamopolski, " Modeling the price of Bitcoin with fractional Brownian motion: a Monte Carlo approach" www.researchgate.net/publication/318392462.

Aleksander Berentsen, Fabian Schar, "A Short Introduction to the World of Cryptocurrencies" Federal Reserve Bank of St. Louis Review, First Quarter 2018, 100(1), pp. 1-16.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, "An Introduction to Statistical Learning with Applications in R" February 11th, 2013.