Forecasting Bitcoin price trends with news

Shanyun Li

99 Jinxiu Avenue, Economic and Technological Development Zone, Hefei, China

lsy927239307@126.com

Keywords: Bitcoin; Prediction; model; R Language; machine Learning

Abstract. The article combines historical news data and historical data of bitcoin and the stock market, performs natural language processing on the news data and reads text information, uses sentiment processing to determine the positive and negative news sentiment, digitizes the news, and combines historical data of the stock market and bitcoin. R language is used for ridge regression, linear regression, logistic regression, random forest, XGBoost and other data analysis to predict the trend and price of Bitcoin.

Introduction

Bitcoin is one of the hottest virtual currency topics today. The price of Bitcoin has fluctuated, and no government or central bank has endorsed it, so no one can provide support for its value. Especially in 2017, the price of bitcoin rose from 1,000 US dollars to nearly 10,000 US dollars. Therefore, it is called the year of bitcoin soaring, and the sharp rise of bitcoin value has attracted the attention of people from all walks of life. This bitcoin boom makes people feel like the latest and crazy speculative bubble, a high-tech field of tulips. During the gradual rise of bitcoin to nearly 10,000 US dollars, it has repeatedly appeared in trouble-disagreement on how to regulate, robbery during transactions, and warnings from regulators. But every time that pundits have warned the bubble is about to burst, the currency has stuttered for a few days and then gone charging higher. [1] Because investors will be affected by various market information and their psychology, resulting in unreasonable investment measures. Many trend-based strategies based on past data have successfully surpassed Bitcoin long and held positions. [2] In order to explain as much as possible the impact of market information on the price of bitcoin, this research collates the news from the time of its appearance to the present and corresponds to the price of bitcoin at the same time. prediction. Of course, the market is complex and efficient. This study believes that this strategy cannot effectively predict the trend of Bitcoin, but the price should be predictable.

Data Sources

This research uses R language to capture the top 25 news data of reddit from September 18, 2014 to December 31, 2019, and collected Bitcoin and Nasdaq and S & P 500 indexes on Yayoo Finance For historical information data, the data before December 10, 2018 is used as a test set to fit the model, and the subsequent data is used as a test set for prediction.

Data preprocessing

The captured data is generally relatively messy and cannot be used directly as the original data resource. It needs to be pre-processed before it can be used. The pre-processing steps are as follows: (1) Processing of Bitcoin data: Converting Bitcoin data to xts format The unified time series operation interface compares today's data with first-order lag data to compare the rise and fall of the closing price. If the closing price of today is greater than or equal to yesterday, it is defined as 1, otherwise it is defined as 0, and then the bitcoin label The data is converted into a data frame format. (2) Processing of news data: Combine 25 daily news together, then convert the date field in the news data into date format, delete all punctuation except the title separator, delete numbers, and

change all content Convert to lowercase letters and remove stop words, as they may have little predictive power. (3) Correlate news data with Bitcoin data and sort them by date.

Experimental results and analysis

Before going into the machine learning algorithm, this study first calculates how different variables affect the price of Bitcoin. The results are as follows:

```
Coefficients:
                  Estimate
(Intercept)     -1.704e+00
sentiment        1.066e-01
Last.label      -1.315e-01
Last.week.label  1.685e-01
Last.month.label -2.474e-02
NDX             -2.082e-04
SP500            1.448e-03
BTC_V1          -3.999e-11
NDX_V1          -6.257e-10
SP500_V1         3.017e-10
```

Judging from the results, positive market sentiment can increase the price of bitcoin. The increase in bitcoin price in the past day and the past month will have a negative impact on the price of bitcoin today. The price of bitcoin has a positive impact. The Nasdaq index and Nasdaq trading volume have a negative impact on the price of bitcoin, and the S & P 500 index and the S & P 500 trading volume have a positive impact on bitcoin.

Predicting the price trend of Bitcoin——unigram model

Use tm to convert the text title to a document term matrix through a corpus object. Each row of the document term matrix will be the combined title of each day, and the column will be the frequency count of the letter combination. [3] Then divide the data into a training set and a test set, use the ridge regression algorithm to fit the glmnet model for prediction on the training set data, put the predicted value and the real value together, use cross-validation to select the best lambda, and then The test set is predicted, and the predicted probability density curve is drawn for the prediction result. From the result (Fig.1), as guessed at the beginning, the probability threshold has no good value, so it is impossible to use lambda to separate these predictions completely.

Then use the ROCR package to evaluate the effect, draw the ROC curve, and get the result AUC of 0.5265. The result shows that the effect is not good. This result is the same as random guessing and has no predictive value.

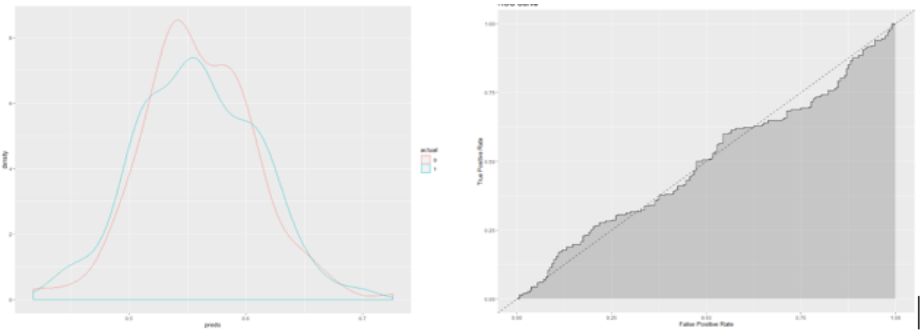

Fig. 1 Probability density curve and ROC curve

Predicting the price trend of Bitcoin——bigram model

Use the same method for reading news data through the tm corpus as the above model, but using a tuple instead of a single word, use the NGramTokenizer in the RWeka package to build a bigram token generator, and then enter it into the control list In setting limits, only binary tokens that exist between 20 and 50 are used.

A ridge regression algorithm was used to fit the glmnet model on the training set data, and then

the test set was predicted. Based on the results, a predicted probability density curve was drawn to determine the best lambda value, and a bigram ROC curve was also drawn.

From the result (Fig. 2), the lambda value cannot be distinguished. The accuracy rate AUC of the ROC curve in Figure 4 is 0.4965, which is worse than the first attempt and worse than random guessing!

Judging from the results of the unary and binary models of natural language processing, the trend of Bitcoin cannot be effectively predicted by the news.

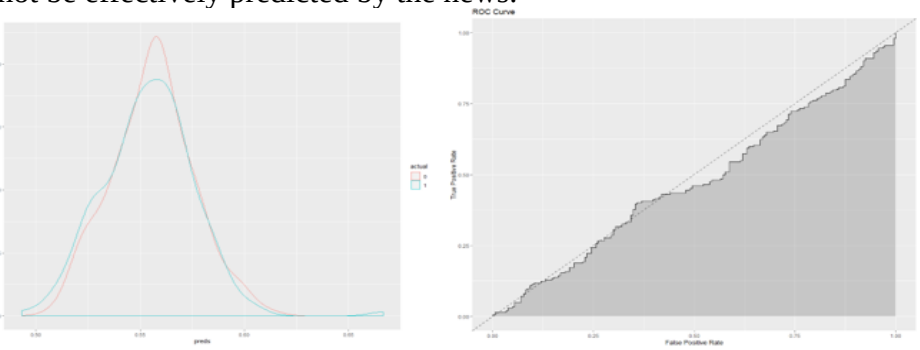

Fig. 2 Probability density curve and ROC curve

Predicting the price trend of Bitcoin——Linear regression

The news data is processed with sentimentr package [4], and then the sentiment results in the news sentences are used as new data. In addition, the following new data is added to the data set: the trend of bitcoin in the past day, the past week The trend of Bitcoin, the trend of Bitcoin in the past month, the historical data of the NASDAQ index, the historical data of the S & P 500 index, and the historical trading volume of Bitcoin and the two indices. Let Bitcoin's current day as the dependent variable, sentiment data and newly added data as independent variables.

Perform a linear regression analysis on the training data: "reg = lm (Label ~ sentiment + Last.label + Last.week.label + Last.month.label + NDX + SP500 + BTC_V1 + NDX_V1 +, SP500_V1,, data = Data_train) ", Make predictions on the test set, compare the prediction results with real values, and the accuracy of the prediction results is 0.53367. This result is not very good, so it is suspected to be a problem of collinearity among dependent variables. [5] The correlation was analyzed with the corrplot package, and the following results were obtained (Fig. 3).

The results indicate that there is a collinearity problem between the variables, so logistic regression was attempted to solve this problem.

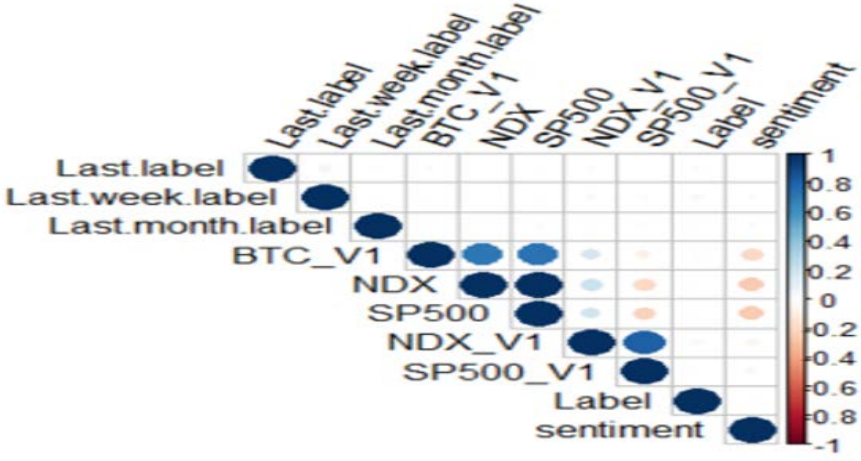

Fig. 3 Correlation between variables

Predicting the price trend of Bitcoin——Logistic regression

The dependent and independent variables are the same as the linear regression, the news is

treated the same, and the training set is analyzed by logistic regression: reg_label = glm (Label ~ sentiment + Last.label + Last.week.label + Last.month.label + NDX + SP500 + BTC_V1 + NDX_V1 + SP500_V1,, data = Data_train, family = 'binomial'), make predictions on the training set, and compare the prediction results with the true values. The accuracy of the prediction results is 0.53368, which is similar to the linear regression results. Linearity is not a problem that leads to low prediction accuracy. In order to make better conclusions, we continue to use random forest [6] and XGBoost [7] for prediction.

Predicting the price trend of Bitcoin——random forest

First, analyze the variables and remove the factors that are not good for the model. The results are as follows (Fig. 4)

From the results, remove the negative variables and then use a for loop to select the number of trees and questions with the highest accuracy, 200 and 5, respectively. Analyze and fit the training set with a random forest model, and test the set. Prediction, the result is 0.5311.

	MeanDecreaseAccuracy
sentiment	2.2340951
Last.label	-2.1911973
Last.week.label	-0.8438493
Last.month.label	-2.3855094
NDX	0.6957796
SP500	0.3481935
BTC_V1	2.4639125
NDX_V1	-5.5157479
SP500_V1	-6.4822282

Fig. 4

Predicting the price trend of Bitcoin——XGBoost

The variables are the same as those in the random forest model. The training set is fitted with the XGBoost model, the test set is predicted, and the prediction result is compared with the real value to obtain the following result(Fig.5).The prediction accuracy is close to 0.56.

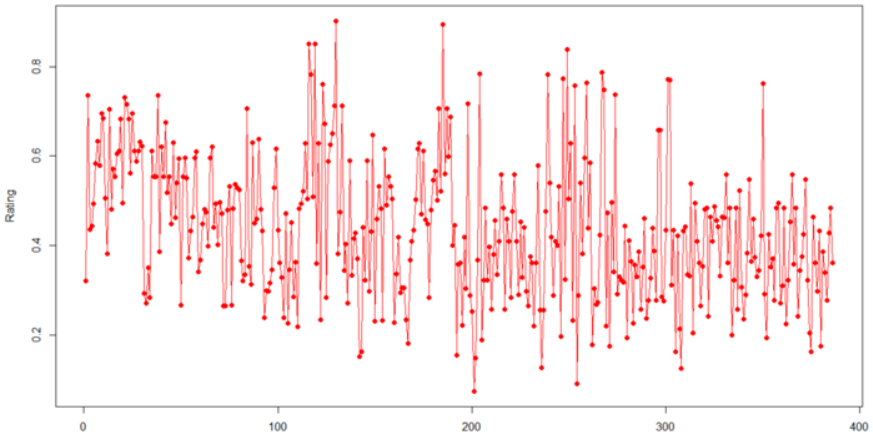

Fig. 5

Predicting the price of Bitcoin——Linear regression

Similar to the linear regression when predicting the trend of Bitcoin, the dependent variable is changed to the price of Bitcoin, the independent variable is unchanged, the linear regression fitting is performed on the data of the training set, and the data of the test set is predicted to obtain the following Results (Fig. 6)

The results show that the average value of the deviation between the predicted bitcoin price and the real bitcoin price in the real bitcoin price is about 0.01631. The price of the coin.

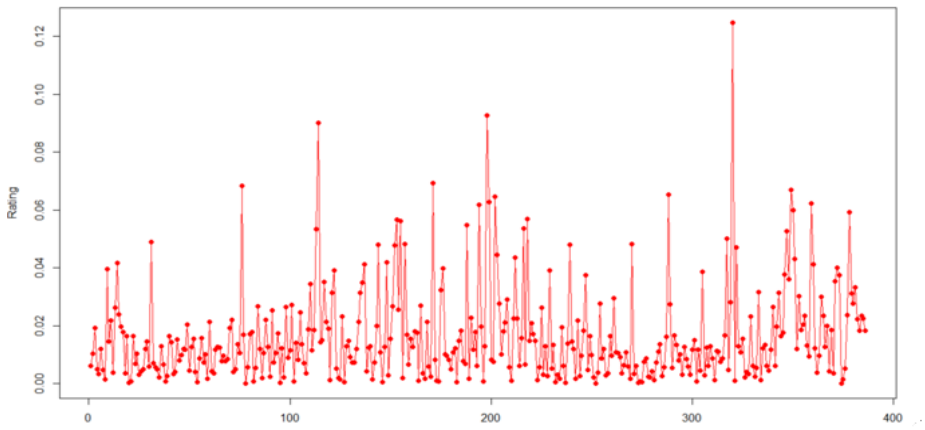

Fig. 6

Conclusion

From the results of the above model, the accuracy rate of predicting the trend of Bitcoin is about 54%, which is not a good result, indicating that it is difficult to predict the trend of Bitcoin through news and historical data of the stock market, but it can be very accurate Predict the price of Bitcoin. The above models have been tested based on past data. Past data does not guarantee that the market will continue to behave in a similar manner in the future. The market, especially the emerging markets, is an evolving entity controlled by participants. These participants are diverse and complex. Therefore, it is a difficult and difficult problem to predict the price trend of Bitcoin.

References

[1] Rory Cellan-Jones. Bitcoin risky bubble or the future? [J / OL]. https://www.bbc.com/news/technology-42138370,2017-11-27/2020-03-16.

[2] John. The Nature of Cryptocurrency Markets[J / OL]. https://adaptiveanalysis.io/the-nature-of-cryptocurrency-markets/,2019-07-08/2020-3-16.

[3] Troy Walters. Stock Prediction with R glmnet and tm packages [DB/OL]. https://www.kaggle.com/captcalculator/stock-prediction-with-r-glmnet-and-tm-packages,2016-11-01/2020-03-16.

[4] Peng Chen, Zhongqian Sun. Recurrent Attention Network on Memory for Aspect Sentiment Analysis[A]. Martha Palmer. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing[C]. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 452–461.

[5] Dongbo Zhao. Study on Multicollinearity in Linear Regression Model[D]. Bohai: Bo Hai university, 2017.

[6] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.

[7] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System[A]. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C].2016.785-794.

[8] Bamert, T., Decker, C., Elsen, L., Wattenhofer, R., Welten, S.: Have a snack, pay with Bitcoins. In: Peer-to-Peer Computing (P2P), 2013 IEEE Thirteenth International Conference on IEEE, 1–5 (2013)

[9] Bergstra, J.A., de Leeuw, K.: Bitcoin and beyond: exclusively informational monies. arXiv preprint arXiv:1304.4758. (2013)

[10] Ateniese, G., Faonio, A., Magri, B., De Medeiros, B.: Certified bitcoins. In: International Conference on Applied Cryptography and Network Security, Springer, Cham, 80–96 (2014)