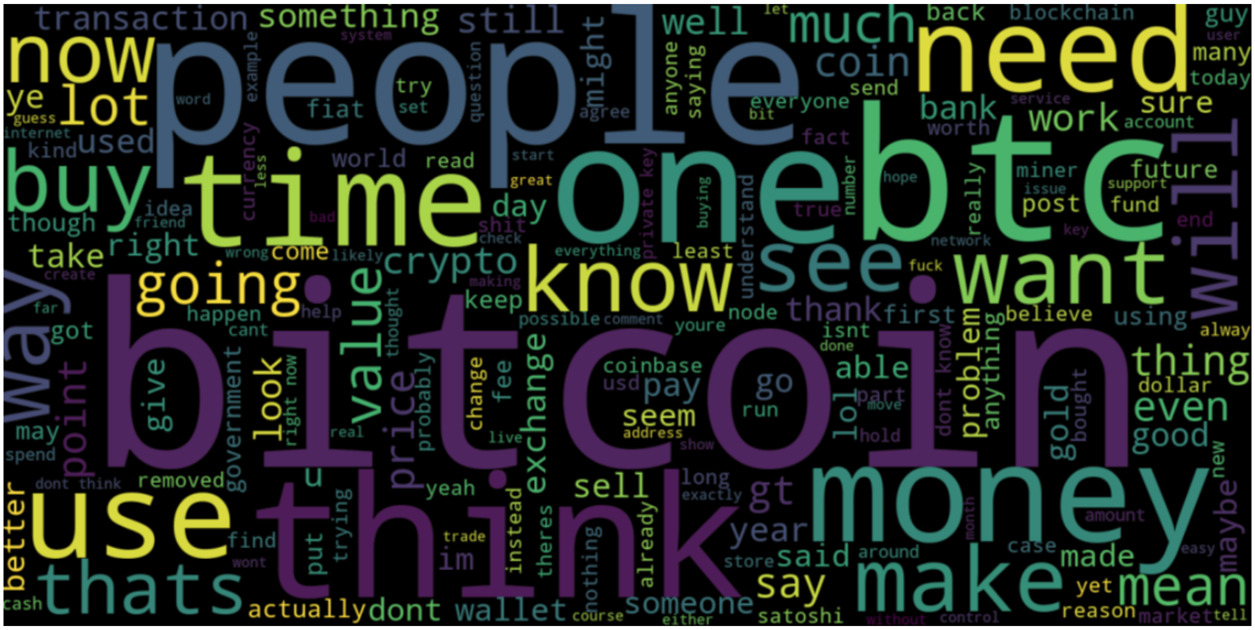


Cryptocurrency Market Prediction

By Derek Plemons

Introduction

There are many benefits to being able to predict whether the market will increase or decrease for a stock or in this case, the cryptocurrency token Bitcoin. What factors play into this market fluctuation? How could one make accurate predictions of the market? How much data is needed? Can social media be used to make such a prediction? There are just some questions I am posed to answer in this project.

The most commonly used trading related statistic on the internet is that 95% of all traders fail. While there is no research to validate this claim, the truth might not be much better. There are numerous factors that play into the success of a trader such as: How quick to make a decision? When to cut one's losses? Holding on to winners and selling losers? I won't be attempting to solve these problems.

Business Problem

Using social media comments, can one make better than random predictions on whether the market will increase or decrease for the following day with better than 50% accuracy?

H_0 : The model predicts whether the price of bitcoin will go up or down for the following day better than 50%. $p = 0.50$

H_a : The model does not predict whether the price of bitcoin will go up or down for the following day better than 50%. $p \neq 0.50$

Target Audience

This problem would be useful for cryptocurrency traders and investors. While this project is focused on predicting market trends for Bitcoin, one could apply the same methodology to other cryptocurrencies.

Data Requirements

1. r/bitcoin subreddit comments from 12/31/2019 to present
    a. ~2500 comments per day were collected with pushshift
2. Historical bitcoin price from 12/31/2019 to present

Methodology

Using the pushshift reddit api, comments from 12/31/2019 to present are pulled utilizing python. The data is broken down by day and stored in an individual csv file. Historical bitcoin cryptocurrency data is collected from the same time frame and also stored in a csv file. After gathering the data, comments from the pushshift api will be cleaned to filter out comments that are not useful to performing a sentiment calculation of each comment. The comments which have extraneous information such as: website urls, “deleted”, “removed” or other such text that does not lend value to sentiment analysis were also removed from the dataset.

After cleaning the data, a sentiment calculation will be performed on each comment for everyday using the sentiment analysis tool Vader. Vader is “a parsimonious rule-based model for sentiment analysis of social media text.” Once sentiment analysis is performed on each day of r/bitcoin comments, the mean of each sentiment score: Positive, Neutral, Negative and Compound will be taken from each csv file and appended to a new dataset with all the means of each day of r/bitcoin comments.

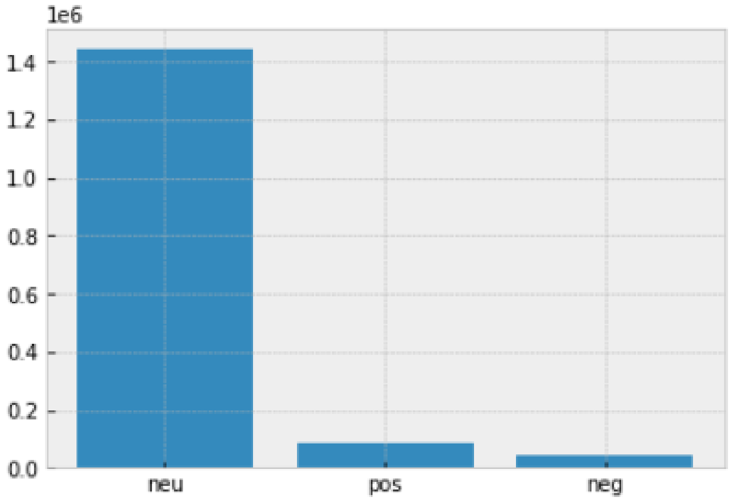
Once this dataset was created, it was merged with the dataset of historical bitcoin closing price. Then, a new column was created with binary values where 1 would equal an increase in bitcoin price from the following day and 0 would equal a decrease in price. A secondary dataset was created where these values would fall on the previous day in order to see if the comment sentiment on the same day could be used for prediction.

Using the cleaned sentiment score and bitcoin price dataset, scikit-learn machine learning library was used for classification: Logistic Regression, K Nearest Neighbors, Support Vector Machine, Random Forest and Gradient Boosting.

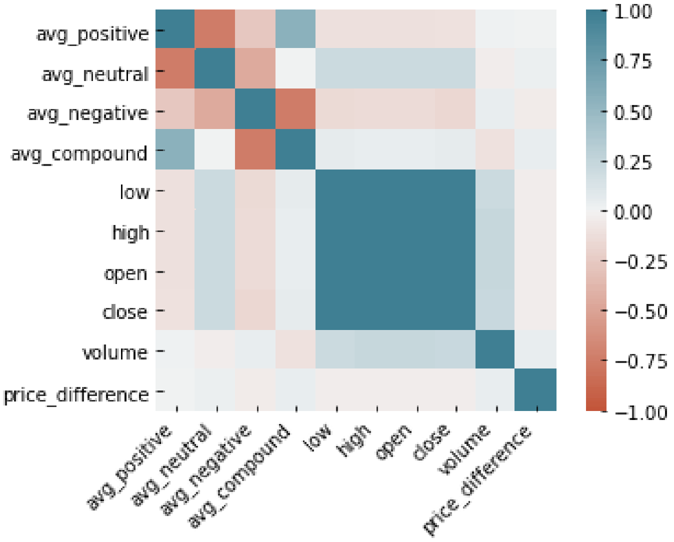
Training and Test datasets were created using scikit-learn and modeled appropriately for each type of classification model. The accuracy, balanced accuracy, precision score and Recall score were calculated for each model. Then a classification report and confusion matrix were used to validate the previous metrics. Finally a ROC and Precision Recall curves were plotted to validate the model visually.

Results

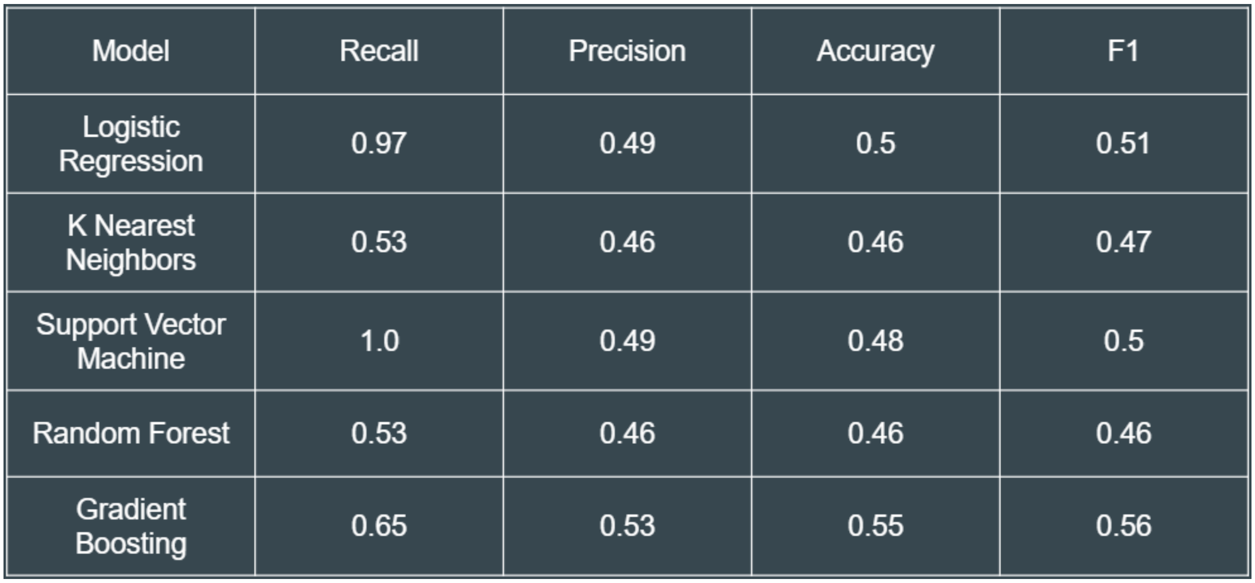
During exploratory analysis, it was found that the vast majority of the comments were neutral.



The below correlation heatmap shows almost no correlation between sentiment scores and the price of Bitcoin. However, the highest correlation between the average neutral sentiment score and the price of Bitcoin presented some opportunity.



The classification report, confusion matrix, ROC and Precision Recall curves were used to determine which model was most effective at predicting whether or not the price of bitcoin would increase or decrease. Below you can see the table showing the Recall, Precision, Accuracy and F1 score for each model.



Model	Recall	Precision	Accuracy	F1
Logistic Regression	0.97	0.49	0.5	0.51
K Nearest Neighbors	0.53	0.46	0.46	0.47
Support Vector Machine	1.0	0.49	0.48	0.5
Random Forest	0.53	0.46	0.46	0.46
Gradient Boosting	0.65	0.53	0.55	0.56

Discussion

Gradient boosting presented the best model for predicting whether the price of Bitcoin would go up or down for the following day. Given that the dataset is imbalanced, the F1 score presented the best metric for determining the best model.

The goal of this predictive model was to be able to predict whether the price of Bitcoin would go up or down better than 50%. In this case, a F1 score of 56% with an accuracy of 55% allows us to fail to reject the null hypothesis that we are able to predict whether the cryptocurrency market will go up or down for the following day. And allow us to reject the alternative hypothesis that the model will not be able to predict the market better than 50%.

Limitations and Suggestions for Further Research

1. Forecast out predictions for longer duration of market trends. In other words, not predicting if the market will increase or decrease for the following day, but in the next week, month or year.
2. Collect more complete comment dataset from reddit for better predictions
3. Use Twitter and Google trends for added features
4. Use Bag of Words method for prediction
5. Access computer with more ram for faster modeling

Conclusion

While the model was able to do better than 50% accuracy and fail to reject the null hypothesis, this was with only 10-15% of the total comments per day. With access to more comments this model may be able to improve on the predictions. However, with a model that predicts the trends of Bitcoin for the following day better than 50%, this presents an opportunity to use for cryptocurrency trading.

References

1. https://vantagepointtrading.com/whats-the-day-trading-success-rate-the-thorough-answer/
2. https://tradeciety.com/24-statistics-why-most-traders-lose-money/
3. https://www.liberatedstocktrader.com/stock-market-statistics/
4. https://pypi.org/project/vaderSentiment/