IEOR242 Applications in Data Analysis

Estimating Bitcoin Price Trends based on Tweets Attitude

Group Members

Yuehe Wen, yuehe98@berkeley.edu, ID: 3037343565

Yujia Song, yujia-song@berkeley.edu, ID: 3036337615

Chenyao Zhu, cyzhu@berkeley.edu, ID: 3037466211

Kaifeng Lu, kaifeng_lu@berkeley.edu, ID: 3037391304

Yuntian Shen, ytshen@berkeley.edu, ID: 3035183230

1. Motivation

1.1 Background

Bitcoin(BTC) is gaining popularity in recent years, and it's one of the most important decentralized cryptocurrencies based on blockchain technology. Despite its popularity, the price of bitcoin it's not stable. The Bitcoin market cap had grown over 1,000 billion USD In April 2021 and declined to 600 billion U.S. dollars in June 2021. This dramatic situation made us curious about the reason behind this fluctuation.

1.2 Assumption

Without assurance from companies or central banks, the price of bitcoin can fluctuate significantly and it's hard to consider all features that may affect its price. In the project, our team assumed that public opinion toward bitcoin is one the most important factors that affect the Bitcoin price. Since Twitter is a platform where people can express their opinions freely, we used tweets related to bitcoin to represent the public's attitude towards Bitcoin.

1.3 Goal

Through machine learning methods and analysis, we want to verify our assumption that there is a relationship between public attitude and bitcoin price and use tweets to predict the trend of the price of bitcoin of the day. Furthermore, we will give some insights about the bitcoin's market, analyze the current public sentiment of the market, warn potential buyers about the possible risks of bitcoin's market, and give some recommendations for investment in cryptocurrencies like bitcoin.

2. Data Preprocessing

2.1 Data Collection

We collected our dataset from two sources: "bitcoin tweets" from Kaggle (Kash, 2021), which contains tweets from February 5th, 2021 to November 26th, 2021, and BTC price history in USD from Yahoo Finance (2021) during the same period.

2.2 Labeling

According to our assumption, BTC price increases when the market attitude is positive and vice versa. Thus, we labeled each tweet into three categories: "Positive", "Negative", and "Neutral" based on the change in price, the difference between close price and open price, on a corresponding day. When the change ratio did not exceed 1%, we labeled the tweets as "Neutral"; otherwise, tweets were labeled "Positive" and "Negative" based on positive and negative changes respectively. To reduce the risk of having spam and meaningless tweets, we selected the tweets from users with more than 100k followers.

2.3 Text Cleaning

We cleaned the text of the tweets with string library and nltk library: we removed punctuation, digit, non-English characters, and English stopwords and finally processed word-stemming before tokenizing the words. In all models below, the default setting has min_df = 0.005 in tokenizing.

3. Model Implication

After the preprocessing of data, we randomly assign 70% of them to be the training set and 30% be the test set. The training set has 5633 “Positive”, 4154 “Negative” and 2231 “Neutral”. The test set has 2423 “Positive”, 1781“Negative” and 947 “Neutral”. The barcharts for the sentiment distributions are shown in Figure1 and Figure2 in appendix. We can observe that they look similar in relative proportions. And we get the baseline model accuracy at 0.470.

3.1 Logistic regression

We first try logistic regression to build the classifier. Logistic regression here is used as a non-linear model and works to the classification algorithm. Using the LogisticRegression module from sklearn, we can build a linear regression for the training set. To check the accuracy of the logistic regression model, we set a default threshold of 0.5, which means assigning each row a probability of bringing positive and then making a prediction for each row where that probability is greater than or equal to 0.5. The accuracy of the logistic regression model is 0.470, which does not show an improvement with the baseline model.

3.2 Linear Discriminant Analysis

Then we consider using Linear Discriminant Analysis. LDA can be used to increase accuracy if there exists a homogeneity of variance/covariance problem compared with the linear regression model. Using the LinearDiscriminantAnalysis module from sklearn, we can build the model. Then we use the accuracy_score module to check the accuracy of LDA and get the value is 0.506, which seems better.

3.3 CART

Classification and regression trees (CART) model splits the nodes to subnodes by searching for best homogenous subgroups on the basis of Gini Index (Dutta, 2021). In our case, the CART model is expected to be able to find the homogeneity between tweets.

We first directly train the dataset with selected ccp_alpha, the result model has accuracy rate at 0.471. Almost the same as the baseline accuracy. By observing the classification tree, we discover that some common words like ‘btc’ are used in analyzing the sentiment (Figure 3). So we remove words like ‘btc’, ‘bitcoin’, ‘crypto’, and ‘cryptocurr’ to try further analysis. But the updated model with selected ccp_alpha has accuracy rate at 0.475, which does not improve much from the previous model. We also try different characteristics in tokenizing: use bi_gram, and set min_df=5, which means all words (1 or 2 word length) that show up more than 5 times are included. However, the highest accuracy we can achieve is 0.479. Still, no obvious improvement from the baseline model. In addition, keywords used for classification like olympicgram, cleannft, long bybit (Figure 4) do not show a clear sentiment for the tweets. Hence, we conclude

that CART might not be a good model for this classification.

3.4 Random Forest

Next we try the random forest. We set up our parameters based on 5 fold cross-validation. And we got an accuracy of 0.475 of the initial random forest model, which is very close to the result of the baseline model, which is 0.470.

We narrowed the initial features to features whose document frequency was higher than 1% and updated our model. And the accuracy of this updated model is 0.477, which is slightly higher than the original model. But still, the result is not satisfactory so we conclude that Random Forest is not the best choice for this classification.

3.5 K Nearest Neighbors

Another model we use is the K Nearest Neighbors to classify whether the Bitcoin price is positive, neutral, or negative. First, we determine the number of clusters by examining the sum of squares within the clusters (WCSS). We plot the Scree plot (Figure 5) and it can be seen from the graph that the sum of squares are all small, so we choose a proper number of clusters, which is 50. The out-of-sample accuracy is 0.4655. To improve the accuracy, we try two ways of feature selection. One way is to use the Neighborhood Component Analysis, which is a method of selecting features by learning a quadratic distance metric with the goal of maximizing the prediction accuracy of classification. The result has an accuracy of 0.467, which is improved slightly. The other method is to manually select the first 200 most frequent features in the training and test set, and the result gives an accuracy of 0.454. Since both of the results are dissatisfactory, we decide not to use KNN model to be our final model.

3.6 Recurrent Neural Network with LSTM

Recurrent neural networks with Long Short-Term Memory are commonly used to process sequential data, such as tweets. LSTM enables adding or removing information to node states through three gates: input, forget, and output, which helps the model understand the dependency of each word on previous words in the text.

Tweets were vectorized into a vector space before being fed into the model. Our RNN model consisted of an embedding layer, a LSTM layer, an internal layer, a dropout layer, and an output layer. The loss was set to “sparse_categorical_crossentropy” as a multi-classification problem, and the model was trained with Adam optimizer.

The basic LSTM models were trained with an accuracy of 0.48 in the test set. To further improve its performance, we checked the word frequency in the tweets and found topic words such as ‘btc’, ‘bitcoin’, were common in most positive, negative, and neutral tweets. Therefore, those words were less significant to classifying the tweets and needed to be removed to enable the model to take other words as inputs. After we removed topic words and adjusted the coefficients, the best LSTM model, with an embedding layer of dimension 100, a LSTM layer of 60 nodes, a ReLU layer of 256 nodes, a dropout layer of rate 0.3, and an output layer, got an accuracy of 0.59 in the test set.

4. Results

The performance metric we choose is accuracy rate. This is because the "cost" of incorrect estimation for each sentiment is relatively equal. We would like to minimize this "cost". In reality, the investors may even require more than the sentiment to make optimal investment decisions. The accuracy rate for each model is:

	Logistic	**LDA**	**CART**	**RF**	**KNN**	**LSTM**
ACC	0.470	0.506	0.479	0.477	0.467	0.59

The performance of our models on this dataset were similar. In this multi-classification problem, the logistic regression, Linear Discriminant Analysis, CART, Random Forest, and K Nearest Neighbors models had an accuracy of around 0.5 on the test set. The Recurrent Neural Network with Long Short-Term Memory performed the best in the test set, with an accuracy of 0.59, which proved the advantage of LSTM to handle sequential data.

5. Future work

The LSTM model had the best performance on this dataset with an accuracy of 0.59, which was not very ideal. We think there are two possible reasons. First, the tweets in this Kaggle dataset were very short; there were too few words as features for each tweet to train the models. Thus, we plan to test our models on other datasets that contain longer texts, such as posts and articles. In addition, our assumption was likely to be inadequate: there might exist other factors, such as government subsidies or the inflation rate, that encourage people to invest in BTC and affect BTC price. Therefore, we can adjust our assumption and add more features that might affect BTC price.

References

Dutta, B. (2021, July 26). *A Classification and Regression Tree (CART) Algorithm | Analytics Steps*. Analyticsteps. Retrieved December 16, 2021, from https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm

Kash. (2021, November 28). *Bitcoin Tweets*. Kaggle. Retrieved December 16, 2021, from https://www.kaggle.com/kaushiksuresh147/bitcoin-tweets

Yahoo Finance. (2021, December 6.). *Bitcoin USD (BTC-USD)*. Retrieved December 16, 2021, from https://finance.yahoo.com/quote/BTC-USD/history/