Using Reddit News to make stock market trading decisions: difference between predicting price direction and traded volume

Authors
Romain Dumon, Vlad Popa, Stefan Manole

Abstract

In this project, we are using Reddit news headlines to make stock market predictions. For this purpose, we are using four machine learning models: a simple decision tree and an XGBoost algorithm, both combined with a bag of words (with unigrams and bigrams), and a LSTM and bi-LSTM, both using GloVe vectors for word representation. We are forecasting tomorrow′s traded volume and price direction of the Dow Jones Industrial Average index, based on world news from today. Making prediction for both, we have averaged our results and concluded that there is a stronger correlation between news and volume, inferring that the headlines are more relevant for the market as an indicator for traded amounts.

1 Introduction

With the rising notoriety of natural language processing and the increases in computing power, researchers and financial companies are now, more than ever, trying to predict the stock market using NLP techniques. Indeed, NLP is a new set of tools for stock forecasters and applying it to the stock market is a big opportunity to explore the advantages and limitations of the algorithms developed until now. The main motivation behind our research is to apply NLP to the stock market in novel ways. In this paper, we investigate the difference between using Reddit News Article for predicting changes in the price of the Dow Jones index compared to using news to predict changes in volume level. Our findings suggest that volume can be predicted with slightly better accuracy.

To investigate how news headlines can make stock market predictions, we have come up with a particular stock market scenario that we will be investigating.

An investor wants to make a trading decision (buy or sell) when the stock price opens in the morning, and then close his position (sell if he bought/buy if he sold) at the end of the day. If his decision is right and he makes a profit he has won and otherwise he has lost. How should the investor make his decision in the morning to win all the time?

In this project, we are investigating how the investor can use global news headlines from today to help in his trading decision of tomorrow. More importantly, we want to identify what these news headlines predict best between volume traded and changes in stock price for that day. We will then compare our volume and change in stock price predictions to conclude what one day news headlines predict best between the two for our scenario.

Although it is pretty clear why it would be helpful to know the change in stock price in making the trading decision, it might not be clear to the reader why knowing volume is important in trading. Volume is "the total quantity of shares or contracts traded for a specified security (Nickolas, 2018)" and it is important because :

- "Trading volume can help an investor identify momentum in a stock and confirm a trend. If trading volume increases, prices generally move in the same direction. (Nickolas, 2018)."
- "Low trading volume can also signal when an investor should take profits and sell a security due to low activity (Nickolas, 2018)."

2 Related Work

Past years have been crucial in the advancement of Natural Language Processing, especially for those focused on studying financial networks; they have come up with more innovative ways to use the technology in the analysis of those markets, in order to better understand it (Fisher, 2016). The technological evolution in terms of efficiency, accuracy and usage have attracted more and more researchers and investors to work with NLP instruments. One reason is the high expected monetary return for projects combining Machine Learning methods and the financial domain (Bosworth, 1975).

The motivation behind combining the study of financial markets with the analysis of news headline is not new, but has its roots long time ago, at the beginning of trading. As scientists and analysts have early discovered, trying to predict future prices based on historical market data is a naive approach. This is due to the fact that the factors which influence the price movements are only found in real economy (trading volume, political changes, new regulations etc.) (Gidofalvi, 2001). With this in mind and following other researchers in this area, we have focused on discovering how news headlines, a relevant economical factor (Gidofalvi, 2001), can affect the Dow Jones Industrial Average. Furthermore, building on top of those studies, we have also identified how those headlines influence the trading volume, another indicator of stock market movements.

2.1 Price Change

As described in different research papers (Tumarkin, 2001; Qing Li, 2014), the price of stock is influenced by many factors that can give traders relevant information, which they use as a base to decide their next trade. In one such paper, a number of stories have been presented to a multi-agent automated system that would label them as positive or negative, based on how it thinks they would affect the company′s attractiveness to investors (Young-Woo Seo, 2014). Our project uses a similar approach, but having the stock movements labelled with the direction of the price on the next day. This is because we are also trying to assess the problem from a different perspective. While most of the researchers are assuming that financial news are having an immediate impact on the market (Tumarkin, 2001; Jethin Abraham, 2018), we are assessing if the Reddit headlines have an impact on the next trading day. This assumption is made because Reddit is not a traditional tool for professional traders to get information, so the news are not released in real time or selected to be tailored to the financial markets. We believe that this might result in a delay in affecting the market, due to the slow nature of the unprofessional traders that would use those news.

Projects working on the same dataset with the goal to create a predicting model for the stock market price change are also available to the public. We have read the documentation and identified the similarities and differences between their approach and ours. One project using an unigram and bigram bag of words and linear regression got an accuracy percentage of 57% (Gele, 2017). Another project using the XGBoost algorithm obtained 53% only, but improved it to 57% by scaling the data (Jain, 2017). Analyzing the way those projects have been implemented gave us the idea to use 4 methods to create our models, using a simple decision tree, the XGBoost algorithm, a long short-term memory (LSTM) recurrent neural network and a bi-LSTM network. This would increase our chances to get a better average accuracy, as different methods get different results depending on the format and content of the dataset.

2.2 Traded Volume

As described in *The effect of news and public mood on stock movements*, there seems to be a relation between news and the trading behaviour (volumes traded, price movements etc). In this paper, the authors try to show that after good news were released for a specific company, there was an increase in the volume of tradings. This is because it is assumed that public sentiments cause emotional fluctuations in investors and intervene in their decision making (Qing Li, 2014). Furthermore, in *Financial news predicts stock market volatility better than close price* (Adam Atkins, 2018), the authors conclude that news and price do not have the strongest correlation and that other closer relationships can be found. In our case, we look at traded volume, a feature which was not addressed much in the literature.

In one project that is similar to ours, the authors have scrapped news from various websites in order to assess if they affect the stock market trading volume for particular companies (Soyland, 2015). The news data is obtained from around 6000 online sources, so it is more diverse and rich than ours, and also targeted. Also, the entries were split into 2 categories for the volume, high and low. For each company, the trade volume median was calculated from all trading days within the period. Using a randomized text classification model, an accuracy of 49.6% was obtained. Using a bigram binary representation and a random forest classifier, the authors have obtained a better accuracy of 78.3%. Even though we were expecting a lower accuracy for our project, due to the news dataset that in our case is only obtained from Reddit and it is not targeted at companies, we have tried to replicate and adapt the methods they used.

3 Methods

We are investigating whether news headlines should be used to predict changes in stock price or the volume traded of a stock. In this project, we consider these two prediction tasks as classification problems. Now, to compare the use of news headlines for price change predictions and volume predictions, we use four machine learning models with different encoding mechanisms. The models we use are Simple Decision Tree, XGBoost, LSTM and Bi-LSTM. We train these models to make volume predictions and price change predictions and then keep the best test accuracies of each these models for both metrics. Our aim is to find the highest achievable test accuracies out of all the models for price prediction and volume prediction to then conclude on what news predicts best between the two.

3.1 Metrics

Price Change. We define stock price changes as a binary classification problem: price increase or price decrease. Going back to our scenario in the introduction, this helps the trader make trading decision in the morning because if prices are going up he should decide to buy in the morning and sell in the afternoon, or should decide to do the opposite if prices decrease.

Volume Prediction. We define volume prediction in three classes: low, normal, and high. This is because we know from finance trading theory that it is when the volume is noticeably low or noticeably high that the investor needs to use volume to trade.

Test Accuracy. We define test accuracy as the percentage of correctly classified news headlines into the respective price change and volume category on data not used for training.

3.2 Dataset and Data Cleaning

For this project we used the dataset provided on Kaggle (Aaron7sun, 2017). The first channel of this dataset consists of the most popular 25 Reddit headlines for each day between 08/08/2008 and 01/07/2016. The second part offers information about the Dow Jones Industrial Average (DJIA) and contains the open, high, low, close prices and volume (number of trades) for each day in the specified date range. The two channels of data were combined into a single dataframe in which the first column is "Date", the second one is "Label" and the remaining ones are the reddit news headlines ranked from "Top1" to "Top25". For the price direction predictive model we labelled the data with 0 (if the price in the next day went down) and 1 (if the price in the next day went up), The traded volume was split into 3 categories, -1 being the tag for days with a number below 100 million (25% of data), 1 for those with more than 192 million stocks (50% of data) and 0 for any volumes in between (25% of data). A screenshot of the data format can be seen in Figure 1.

Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...
2016-06-24	0	David Cameron to Resign as PM After EU Referendum	BBC forecasts UK votes to Leave the European U...	Nicola Sturgeon says a second independence ref...	It's official. Britain votes to leave the Euro...	World's Largest Tibetan Buddhist Institute Ord...	Not a single place in Scotland voted to leave ...	Rich Getting Richer at the Expense of the Poor...	Spanish minister calls for Gibraltar to be ret...	...
2016-06-27	1	Barclays and RBS shares suspended from trading...	Pope says Church should ask forgiveness from g...	Poland 'shocked' by xenophobic abuse of Poles ...	There will be no second referendum, cabinet ag...	Scotland welcome to join EU, Merkel ally says	Sterling dips below Friday's 31-year low amid ...	No negative news about South African President...	Surge in Hate Crimes in the U.K. Following U.K...	...
2016-06-28	1	2,500 Scientists To Australia: If You Want To ...	The personal details of 112,000 French police ...	S&P cuts United Kingdom sovereign credit r...	Huge helium deposit found in Africa	CEO of the South African state broadcaster qui...	Brexit cost investors $2 trillion, the worst o...	Hong Kong democracy activists call for return ...	Brexit: Iceland president says UK can join 'tr...	...

Figure 1. Dataset format

Furthermore, we applied basic cleaning methods to the dataset: changing capitalized letters to lower, solving character error format "&amp" to "&" and removing quote error format "b'(quote)'". Punctuation was then removed and by using the Porter Stemmer library we have tokenized the headlines and cleaned the data with the built in functions. The concept of stemming is similar to the normalizing one, removing the tense or other redundant variation from words and keeping them with the dictionary form ("introducing" - "introduce").

3.3 Models

In this section we describe why we use each model, the feature engineering done and how we train them. It is important to note that we implement a variety of models in order to maximize our chances of finding the best achievable test accuracy for volume prediction and price change prediction.

3.3.1 Decision Tree and XG boost

In our project, we have decided to use the XGBoost algorithm because of the number of successful attempts this algorithm had on similar project to ours. This algorithm has showed good results in multiple text-classification tasks like sentiment analysis (Zabocki, 2017) and has the advantageous property of being fast to train. We also decided to train a simple decision tree classifier to have a benchmark for the XGBoost algorithm. Simple Decision tree algorithm is a good benchmark because the XGBoost algorithm is essentially a gradient boosted decision tree.

Decision trees algorithms make predictions based on decision rules learned from prior data. These decision rules will depend on the specific attribution selection measures used. For our project, we have decided to use the Gini Index. The Gini index is a default attribute selection measure for decision trees and is considered to be the standard one. For the frameworks used, the entire training pipeline that we describe below was built using Scikit-learn (scikit-learn developers) module for the Feature engineering part and the Simple Decision Tree model training, while the XGBoost (xgboost developers, 2016) API was used for XGBoost. The entire pipeline is built in Python.

Feature Engineering. The input representation that we feed in our tree models are bag-of-n-gram representation of the news headlines. This bag-of-n-gram representation was implemented using the CountVectorizer from Scikit-learn. First of all the entire cleaned input data corpus is fed into to a CountVectorizer object which builds n-grams for the entire corpus. After splitting the data into training and test sets, the news headlines for each day are converted into bag-of-n-grams representations using the vectorizer object fitted on the entire corpus. As a results for each day in our dataset, we have one bag-of-n-gram representation for the news on that day. We do not specify the n-gram representation here because we perform a parameter search over the different n-gram representations possible. The motivation behind this is explained in the Parameter Search paragraph of this section.

Parameter Search. As mentioned in the Feature Engineering section, we do not fix the n-grams in the method because we carry a parameter search over the best n-gram for each algorithm. Our intuition for the parameter search is that we are not sure what n-gram representations the models can extract more meaning from. For instance, it is possible that the Simple Decision trees model works best with 1-2 gram representation on price predictions assigning meaning to pair of words like United States or Barack Obama, while XGBoost works best with 3-grams representation for volume categories predictions assigning meaning to triplets of words like Barack Obama criticizes. Additionally, we thought that stemming the input data using the PorterStemmer algorithm could also work best for some predictions. This is done using the Porter Stemmer python module (nltk developers) on the entire corpus before any feature engineering is done. Therefore we run this n-gram parameter search with and without Porter Stemming.

Model Implementation. After finding the best n-gram representation for each model for each task in the parameter search, we then fit a simple decision tree and a XGBoost algorithm to these training sets (80%) and test them on a separate testing set (20%). For a similar training set and test set, we get a test accuracy value for the price change prediction and volume prediction for each of the two models. This process is repeated for three different random training and test set in order to see the best accuracies achievable for each of the two models for each classification task.

3.3.2 LSTM / BiLSTM

Lastly, in our project we used Long Short-Term Memory (LSTM) neural networks for both the tasks of predicting the price and volume in the next day. One of the reasons for choosing this technique is because LSTMs are a good way for extracting patterns when the input spans over longer sequences. Moreover, as it contains gates in its architecture, LSTMs can easily manipulate its memory state and deal with time series forecasting.

Feature Engineering. In the words embedding process we firstly tried to train our own vectors by using the Word2Vec model available in the Gensim library. Due to the limited size of the vocabulary, this attempt gave bad results, as the analogies between the similarity of different words proved to be completely wrong. Consequently, we decided to use the words vector representations from the pre-trained Glove algorithm. These embeddings were trained on a one billion words dataset from Wikipedia which contains a vocabulary of 400 thousands words. Glove offers several embedding vector dimensions, such as 50, 100, 200 and 300. We tested all of these versions. After getting the Glove embeddings, we created a weighted matrix for all the words in the training set.

Model Implementation. In order to build and train the model, we firstly cleaned the headlines and convert them into sequences. The data processing implied the cleaning process described above. Furthermore, all the headlines for each day were concatenated into a large string of sentences. We used the Keras tokenizer function to tokenize the text and then created sequences of words with the help of the textstosequences function. At this point, each day contains a an array of integers where each integer represents a word in our vocabulary. In order to have the same input length for each day, we checked the maximum length in our dataset and then padded with zeros the text iputs which were shorter.
Following these steps, we created the Keras model with the Sequential() constructor. The first layer, the embedding layer, of the neural network was added using the weighted matrix with the Glove embeddings presented above. As these embeddings were already learned, we set the trainable attribute of the model to be False. The last layer of the networks was different for each prediction task. When dealing with the price prediction, we added a Dense layer containing the sigmoid activation function and giving a one dimension output. This model was compiled using the binary crossentropy loss. On the other part, in the case of the volume prediction task, we had three different values in the target variable. Consequently, we converted the volume category into a binary class matrix using the $'tocategorical'$ function available in keras. Following this, we added a Dense layer with 3 units as output dimensionality and a softmax activation function. This model was compiled using the categorical crossentropy loss.
In addition, we also used a bidirectional LSTM (BiLSTM). By using the bidirectional neural network, the inputs are processed in two different ways: from future to past and from past to future. In this scenario, the model can get a better picture of the context because it can keep information from both future and past at any point in time. The BiLSTM network implementation follows the same steps as the ones described above.

4 Experiments

For running the experiments, the dataset was divided into random training and testing using a 80-20 percentage split. For the LSTM training model, we also created a validation set which was 20% of training subset. In all models, we used accuracy as the main metric for measuring performance. Moreover, the binary crossentropy and categorical crossentropy loss functions were used when training the LSTM model for predicting the price direction and volume category respectively.

4.1 N-gram Parameter Search for Decision Trees

The bag of words data used to train the decision tree and XGBoost models for price change and volume category can be set up in different ways as mentioned in the method section. In order to decide the best N-Gram and stemming combination, we did a search in the style of a parameter search of a neural network among the possible combinations. The table of results of this experiment can be seen in Appendix I. From this experiment, the XGBoost algorithm gave us the best accuracy for Volume prediction. To increase the accuracy on this task, we tuned the learning rate of XGBoost [0.3,0.03,0.003,0.5,0.7] and the max tree depth [2,4,6,12,24] for the volume prediction task.

4.2 LSTM and Bi-LSTM

For the LSTM and BiLSTM we firstly tried to use and compare all the available dimension sizes offered by Glove. Secondly, we ran multiple experiments using a different value, between 1 and 50, for the number of epochs/iterations. By doing this, it was noticed that the model encounters overfitting as the validation accuracy and loss was always getting worse after the first 5 iterations. In order to reduce overfitting we added a Dropout layer and tested several rates between 20% and 50%. Additionally, for improving the model performance, we varied the number of neurons in the network and used the SGD Keras library in order to tune parameters such as learning rate, momentum or decay in the optimization algorithm.

5 Results

The table below outlines the best test accuracies obtained for each of the models. In this section, we report the best results that we obtain and discuss our findings.

	Price Acc	Volume Acc
S.D. Tree	55.89 %	52.63 %
XGBoost	54.88 %	**63.90**%
LSTM	55.37 %	53.69%
Bi-LSTM	**57.67** %	54.73

5.1 Decision tree and XGBoost

The best accuracy was obtained on Volume Category prediction using XGBoost. The feature engineering set up which led to this result was bag-of-1-2-grams representation without the Porter Stemmer Algorithm. The XGBoost parameters for this accuracy were learning rate: 0.3 and max tree depth: 6.

	Best Acc.	Feature Eng.
S.D.T Price	55.89 %	3-gram Porter
S.D.T Volume	52.63 %	2-gram
XGB Price	54.88 %	1-2 gram
XGB Volume	63.90 %	1-2 gram

This table shows what Feature Engineering bag-of-n-grams/Porter Stemmer combination was used to obtain the best accuracy results for each model.

5.2 LSTM/BiLSTM

The best results when using an LSTM/BiLSTM to predict the price direction and volume category can be seen in Figure x. We obtained these results by conducting multiple runs and tuning several parameters. For the words embedding process, the model performed best when using the 100 Glove dimension vectors trained on Wikipedia. Furthermore, for the price direction task, the best configuration had the following parameters: learning rate-0.01, momentum-0.8, number of neurons-100, while for the volume category task we had: learning rate- 0.1, momentum-1 and number of neurons-100.

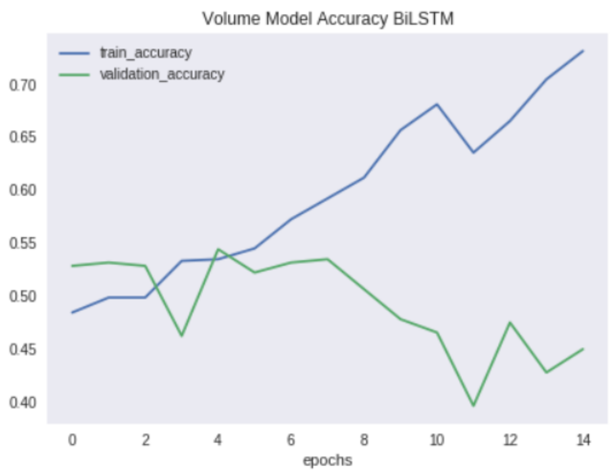

Figure 2. Model Accuracy

Figure 2 illustrates the training and validation accuracy over 15 epochs. Even if the training accuracy increases with every iteration, the validation one reaches its peak in the first epochs and then it gets worse. Consequently, we decided to run the model with only 5 epochs and added a Dropout layer with a 20% dropout rate. This was the final configuration of the network and the one that performed the best.

5.3 Discussion

It is worth mentioning that all the implemented models performed better than the null accuracy of the testing subset which was 50%. This aspect assures us that there is predictive power in using financial news for stock market prediction. For predicting the price direction the best accuracy was obtained with the Bi-LSTM network, which performed better than the normal LSTM in all our experiments. On the other part, in the volume category prediction task the best performing model was the XGBoost. Our final accuracies are in line with the ones obtained in the literature on the same prediction task and when working with the same dataset.

However, even if the volume prediction performance was better, we consider that the final results do not represent a cornerstone in deciding which of the two prediction tasks is more correlated to the financial news. This is mainly due to the lack of data and variety related to the vocabulary. Moreover, it was noticed that there was no specific words embedding or parameter configuration that performed best in all of the models. Additionally, for every compilation the LSTM model gives different results due to the weights assignment and general stochastic optimization. Following these inconsistencies and limitations, the results obtained are not considered reliable enough to reach a final conclusion.

6 Conclusion

In conclusion, it seems from our result that news headlines predict volume categories slightly better than price changes and we would advise the trader from our trading scenario to follow results from our research. A plausible explanation for this is that the total number of stocks which are traded in one day is more affected by news, rather than the price which is affected by demand, supply and many other factors. For instance, if Bloomberg News mention, out of the blue, how listed Company A had bad quarterly revenue growth compared to expectations, investors will immediately react. As a result of the news, many investors will start selling or buying shares of this company which should increase the volume of the stock. On the other hand, if there is no news about Company A, volume should be low because the there is no information for traders to trade on. The relationship between news and price is less clear. The stock price depends on the demand and supply of a stock and it is less easy to interpret how a particular piece of news might impact a company. For instance, taking the example just above, although bad quarterly revenue growth sounds like it would decrease the stock price, maybe it was expected because Company A went through a restructuring phase and, as a result, the stock price does not change. Thus, price changes would need more context than volume to be predicted with news.

Nevertheless, it is important to realize that there are some limitations to our findings as mentioned in the discussion section and the findings of this report need more work to be confirmed. In the next section, we discuss possibilities for improvement.

6.1 Future Work

It is clear that having a larger dataset would be the first step for trying to get more accurate results. Apart from Reddit news, one could use additional information such as tweets or financial news taken from online newspapers. Additionally, a future direction would be to filter the news such that we only focus on the companies that were at some point part of the Dow Jones Industrial Average. By gathering more data and having a larger vocabulary, we would be able to train and fine tune the embedding vectors for this particular prediction task.

Moreover, an attention mechanism added to the LSTM network would be good future step. This way, the network would be able to make correlations between the output and certain sections in the input sequence. Another approach, already considered in the literature (Y. Zhai and Halgamuge, 2007), was to combine news and technical indicators to predict the stock market trends. In terms of the methodology used, one could try implementing Convolutional Neural Networks or Support Vector Machines in order to get better results. Lastly, the classification task could be switched to a regression by trying to predict the percentage change in volume and price.

$$X' = X - X_{min} \overline{X_{max} - X_{min}}$$

References

Aaron7sun. 2017. Daily news for stock market prediction. Last accessed 17 March 2019.

Enrico Gerding Adam Atkins, Mahesan Niranjan. 2018. Financial news predicts stock market volatility better than close price. Last accessed 17 March 2019.

Barry Bosworth. 1975. The stock market and the economy. Last accessed 17 March 2019.

scikit-learn developers. scikit-learn. Last accessed 17 March 2019.

nltk developers. Stemming words with nltk. Last accessed 17 March 2019.

xgboost developers. 2016. Xgboost. Last accessed 17 March 2019.

Ingrid E. Fisher. 2016. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Last accessed 17 March 2019.

Andrew Gele. 2017. Omg! nlp with the djia and reddit! Last accessed 17 March 2019.

Gyozo Gidofalvi. 2001. Using news articles to predict stock price movements. Last accessed 17 March 2019.

Shreyams Jain. 2017. Stock price prediction - 94% xgboost. Last accessed 17 March 2019.

Daniel Higdon Juan Ibarra Jethin Abraham, John Nelson. 2018. Cryptocurrency price prediction using tweet volumes and sentiment analysis. Last accessed 17 March 2019.

Steven Nickolas. 2018. Using trading volume to understand investment activity. Last accessed 17 March 2019.

Ping Li Ling Liu Qixu Gonga Yuanzhu Chenb Qing Li, TieJun Wang. 2014. The effect of news and public mood on stock movements. Last accessed 17 March 2019.

Christian Soyland. 2015. Interday news-based prediction of stock prices and trading volume. Last accessed 17 March 2019.

Robert Tumarkin. 2001. News or noise? internet postings and stock prices. Last accessed 17 March 2019.

A. Hsu Y. Zhai and S. Halgamuge. 2007. Combining news and technical indicators in daily stock price trends prediction. Last accessed 17 march 2019.

Katia Sycara Young-Woo Seo, Joseph Giampapa. 2014. Financial news analysis for intelligent portfolio management. Last accessed 17 March 2019.

Marcin Zabocki. 2017. Sentiment analysis of tweets with python, nltk, word2vec scikit-learn. Last accessed 17 March 2019.

A Appendix I

Best Accuracies out of 3 runs

N-grams	Steller	Price Accuracy Decision tree	Price Accuracy XGBoost	Volume Accuracy Decision tree	Volume Accuracy XGBoost
1,1	True	50.38	51.13	48.12	56.89
1,1	False	49.87	53.13	49.62	59.14
2,2	True	54.64	53.63	48.87	54.39
2,2	False	51.38	**54.88**	49.87	55.63
3,3	True	**55.89**	52.88	47.61	48.87
3,3	False	53.13	53.38	46.87	49.62
1,2	True	52.88	50.37	49.37	56.64
1,2	False	55.13	51.63	**52.63**	**59.89**
1,3	True	51.63	50.37	48.37	57.14
1,3	False	53.88	51.62	51.12	59.40
2,3	True	55.13	53.13	49.62	54.89
2,3	False	49.87	53.63	48.37	55.14

XGBoost and Decision Tree