project\_proposal\_EECS545

Method: Gradient Boost Trees

**Model Features**

The implementation of the model supports the features of the scikit-learn and R implementations, with new additions like regularization. Three main forms of gradient boosting are supported:

Gradient Boosting algorithm also called gradient boosting machine including the learning rate.

Stochastic Gradient Boosting with sub-sampling at the row, column and column per split levels.

Regularized Gradient Boosting with both L1 and L2 regularization.

**What Algorithm Does XGBoost Use?**

The XGBoost library implements the gradient boosting decision tree algorithm.

This algorithm goes by lots of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines.

Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. A popular example is the AdaBoost algorithm that weights data points that are hard to predict.

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

This approach supports both regression and classification predictive modeling problems.

**Data:** Behavior Data of Retail Investors | **Format:** .xlsx | **Size:** 1.3GB

**General Description:** This data set is about users’ behavior on a stock data app. It mainly includes click numbers, favorite percentages, numbers of news and these numbers grouped by different types of investors. Market data retrieved from financial data terminal will also be used and serve as labels in exploration but the variables in the following table are major characters in this project.

We are interested in turning the data into information and decisions. Therefore, analysis of the data is essential.

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Types | Description | Locations(Local and part of the data are uploaded to my Github) |
| Positive & Negative News Numbers | int | For stock A on day D, the positive news numbers and negative news numbers | https://github.com/RickYuankangHung/Quantitative-Finance/blob/master/PosAndNegNewsNumber.xlsx |
| Favorite Percentage | float | For stock A on day D, Favorite Percentage =(number of users who add A to their favorite list)/(number of users) | https://github.com/RickYuankangHung/Quantitative-Finance/blob/master/FavoritePercentage.xlsx |
| Click & Percentage | Int;  float | For stock A, Click =(number of users who click A for information)Percentage is Click/Total Click | https://github.com/RickYuankangHung/Quantitative-Finance/blob/master/ClickAndPercentage.xlsx |
| News Click Percentage | float | For stock A on day D,  =(click number of news )/(total click number of news) | https://github.com/RickYuankangHung/Quantitative-Finance/blob/master/NewsClickPercentage.xlsx |
| Favorite Percentage Group by Asset | float | Similar to Favorite Percentage but are grouped by Client Type | https://github.com/RickYuankangHung/Quantitative-Finance/blob/master/FavoritePercentageGroupbyAsset.xlsx |
| Favorite Percentage Group by Experience | float | Similar to Favorite Percentage but are grouped by Client Experience | https://github.com/RickYuankangHung/Quantitative-Finance/blob/master/FavoritePercentageGroupbyExperience.xlsx |

**Four interesting exploratory questions:**

1. What is the relationship between investor behavior and stock return? Is that deterministic? Can they be correlated in part of the data?
2. Is there any relationship between news number and investor behavior? Which is the leading indicator?

**Data Analysis Methods, Visualization Methods, tools or Machine Learning Methods**

In conclusion, our results suggest that by analyzing Twitter sentiment and trading volume, an Extreme Gradient

Boosting Regression Tree Model serves as a viable means of predicting price fluctuations within the ZClassic cryptocurrency

market. As such, given the complete lack of research within this academic sphere, our model serves as a

proof of concept that social media platforms such as twitter can be used to capture investor sentiment, and that this

sentiment is an early signal to future price fluctuations in alternative cryptocurrencies. Of particular interest is seeing

whether this approach produces similarly strong results when applied to other alternative cryptocurrencies such as

ZCash and Bitcoin Private. However, this discovery sheds light to the possibility of arbitrage opportunities that utilize

social media platform sentiment to predict future cryptocurrency prices.

Our pricing model could be further improved by factoring in other social media platforms or data, such as Google

Search results, Facebook posts, and Reddit Posts. Moreover, the dictionary that we have used in our model could

be also be aided by adding investment-specific terms that indicate positive and negative sentiment such as “bull”

and “bear” respectively. As seen from our manual vs. algorithm cross-analysis, the algorithm’s greatest weakness

is in classifying tweets that should otherwise be characterized as “negative” as “positive.” After careful review it

is evident that such inaccurate characterizations are due to the algorithm’s inability to detect sarcasm, a pervasive

language schema in popular social media platforms. As such, further research to enhance our algorithm to detect

sarcasm would increase the reliability of the sentiment analysis, and thereby potentially improve the accuracy of our

prediction to retail driven price changes.

Lastly, it would be interesting to further train and test our model over a longer time period. Given the confines of

the date of our cryptocurrency’s fork and our computational capacity, our study was restricted to a data set that covered

a time frame of 3.5 weeks. However, our results suggest a necessity to devote further resources and investments that

would enable us to implement study our pricing model under a longer time frame and with other cryptocurrencies.

**Papers:**

**https://arxiv.org/pdf/1805.00558.pdf**