Financial Indicators for US Stock Market

Xiaoqi Hu, Yanni Lan, Chenhang Niu, Yuyang Shu, Siqi Zhang

Boston University Questrom School of Business, Master of Business Analytics 2020

Github: https://github.com/CNIU1997/BA_888.git

*Blue text with underline represented clickable graphs

INTRODUCTION

Machine Learning is now helping asset management companies to find new investment opportunities and seek better alpha. On the other hand, many IPO companies have invested money on Infrastructure and R&D to raise more capital from the Stock Market. The movement of time series data from financial markets is influenced by a rich mixture of quantitative information from the dynamics of the system. Company operations, as the data we get from annual financial reports, are related to investor's confidence in stock investment.

OBJECTIVES

Regarding our project, We investigate the relationship between financial report indicators and stock market volatility. The information extracted from financial reports better at predicting the direction of underlying asset volatility movement, or its second-order statistics, rather than its direction of price movement. Another main focus of this project is trying different feature engineer methods to filter out similar variables, reduce the coefficient, and improve the model accuracy.

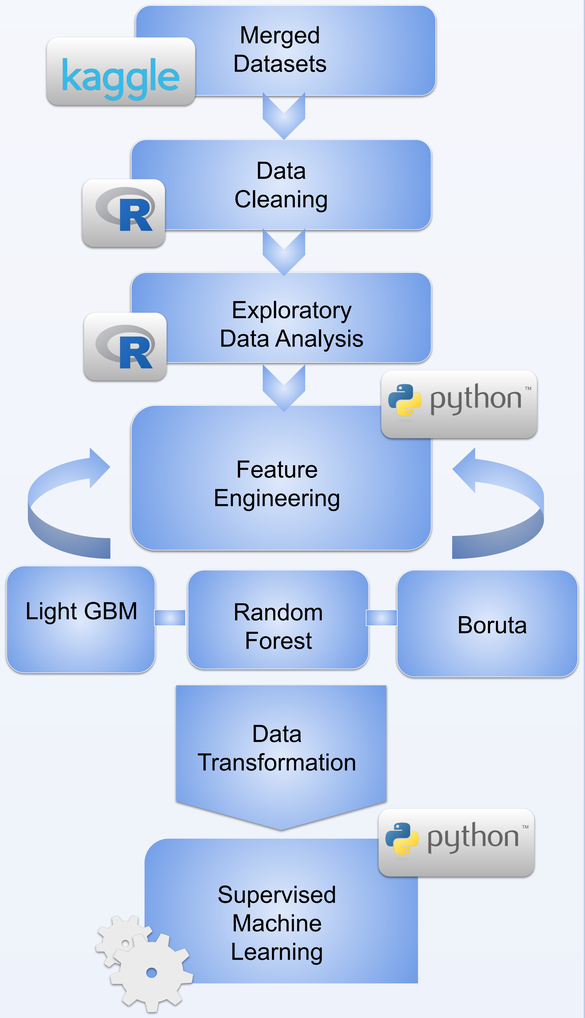
DATASET INTRODUCTION

The datasets we used comes from Kaggle which contains around four thousand stocks of different companies with 226 financial indicators.

Originally we has five-years data stretches from 2014 to 2018. For model training purpose, we combine it into one big dataset with 22,075 observations. All the observation can be divided into 11 industrial sectors. The price variation as the dependent variables indicates price fluctuation compared to the previous fiscal year. The dataset contains variables like Revenue, Cost, Dividend, and all the probable financial indicators from the 10-K filings, which can be used to find the relationship with the stock price variation. Lastly, we randomly divided the combined dataset into train dataset and test dataset following a 80-20 split ratio, with year label taken out. The following is balance checks of both test and train dataset.

Data Type	Percentage of Class 0	Percentage of Class1	Total Number of Observations
Train Data	45.2%	54.8%	17,660
Test Data	44.0%	56.0%	4,415

METHODOLOGY

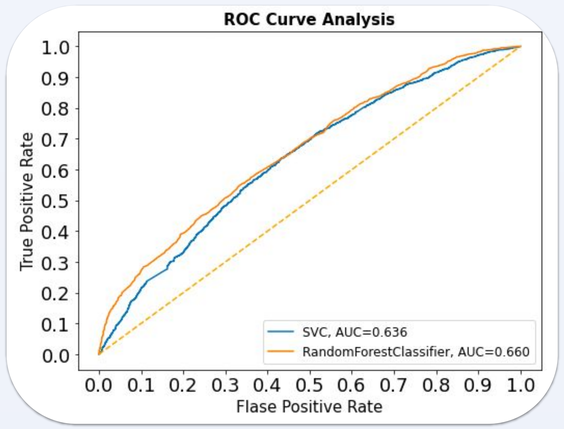


The implement of the project follows the below procedures. The initial step started with merging five datasets into one consolidated data frame. We further implemented initial data cleaning and conducted an exploratory analysis in the R environment. Additionally, the next crucial step is implementing features selection. Regard to our feature engineering, we have tried our various methods but narrowed down to three, Light Gradient Boosted Machines (GBM), Random Forest, and Boruta. We further accomplished supervised machine learning methods on selected features.

RESULT

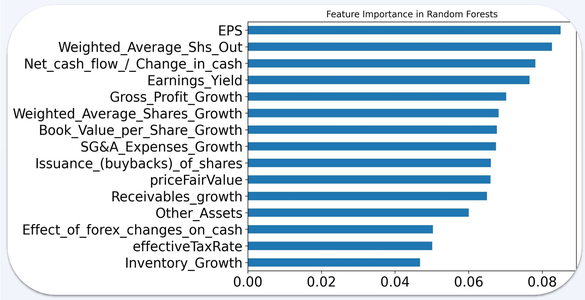
Our project displays included **the result of feature engineering** and the **modeling with selected features**. For each feature method, we have selected a subset of features out of 206 variables and further applied to various models.

The first approach of feature engineering is Light GBM. Light GBM grows tree leaf-wise while other algorithms grow level-wise. In other words, the leaf-wise algorithm can reduce more loss when compared to a level-wise algorithm.137 features reaching 90% of cumulative importance, but we only choose the top 15 features in fitting the models. Later, we tuned the hyper-parameter based on the results. SVM and random forest are the best two models. After finding the best hyper-parameter, random forest indicates the best results with accuracy 61.3% and the value of AUC 66%.



The second approach of feature engineering focuses on Random Forest. As an ensemble model made up of decision trees, Random Forests are often used for feature selection. After averaging and sorting the importance of features, we selected the first 20 variables and applied them to the models. As for modeling, various methods were applied and reported. Additionally, we further tuned hyper-parameters for the best two models, Random Forest and XGBoost, and improved the accuracy slightly from 60.14% to 61.33%, 60.14% to 60.20%, respectively. The AUC for XGboost decreased slightly from 0.642 to 0. 614, but the AUC improved from 0.638 to 0.646 for Random Forest.

The last approach is Boruta which is a random forest method used in R. Apply boruta that will help us randomly shuffle the dataset and compare it to the original dataset to find feature importance. After 200 iterations, the model confirmed 131 variables and reject 46 variables.
According to feature importances ranking we selected different combinations of variables to put into our logistic regression model using our test dataset, after many runs, we finally picked top 42 variables and reached an accuracy of 69.4% and ROC of 0.628 as our highest accuracy record with this approach.



After applying different features into various models, the first feature engineering approach Light GBM achieved the highest accuracy and AUC value, with 61.3% and 66%, respectively. The plot shown above is the feature importance under the best model random forest, which features EPS, weighted average shares out, net cash flow, and other 13 features attribute the most among the total 205 features.

CONCLUSIONS

We advise potential feature investors to make their decisions of picking stocks based on features from the best model result and lay their eyesights on features with top importance scores since the majority of these features reflect either the ability to generate the profits or the change in cash to identify value stocks and make wise decisions.

REFERENCES

Kaggle Original Dataset
https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018/kernels

CONTACT

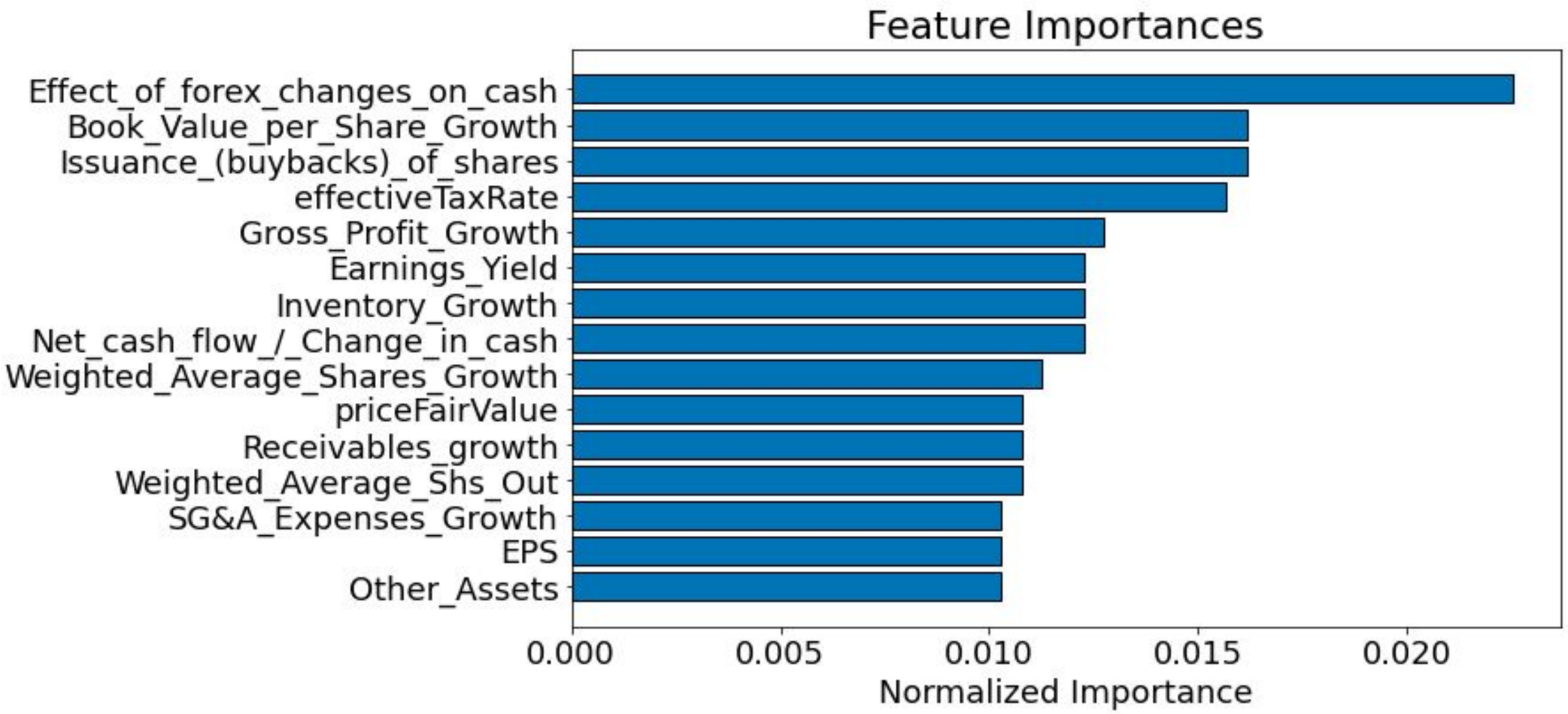
Boston University, Questrom School of Business, MSBA 2020 Cohort B Team 1

Financial Indicators for US Stock Market

Xiaoqi Hu, Yanni Lan, Chenhang Niu, Yuyang Shu, Siqi Zhang

Boston University Questrom School of Business, Master of Business Analytics 2020

RESULT PLOT



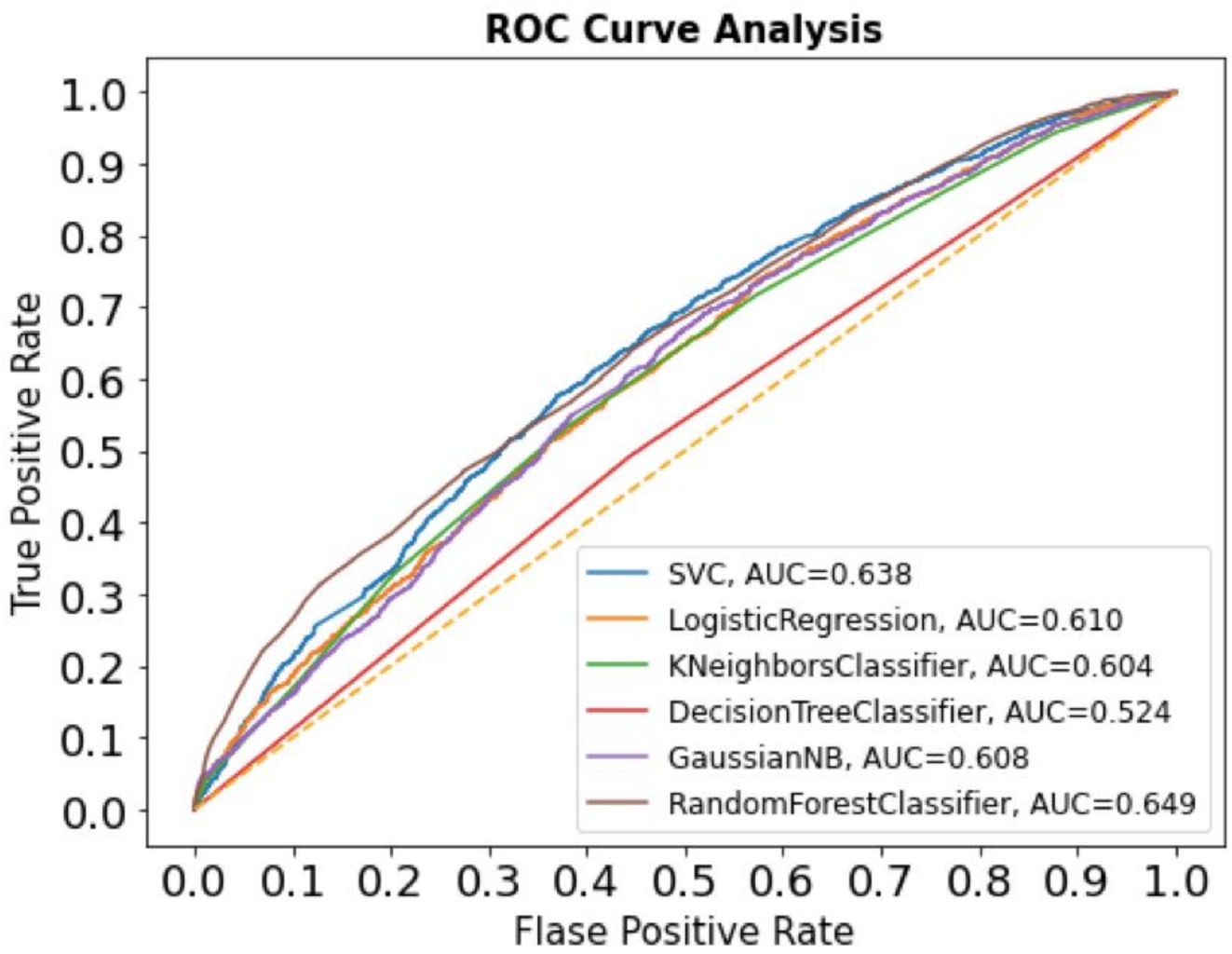
Back to the main table

Financial Indicators for US Stock Market

Xiaoqi Hu, Yanni Lan, Chenhang Niu, Yuyang Shu, Siqi Zhang

Boston University Questrom School of Business, Master of Business Analytics 2020

RESULT PLOT



Model	Accuracy
SVM	61.3%
Logistic Regression	59.8%
Decision Tree	52.0%
KNN	58.1%
Naive Bayes	59.7%
Random Forest	60.5%

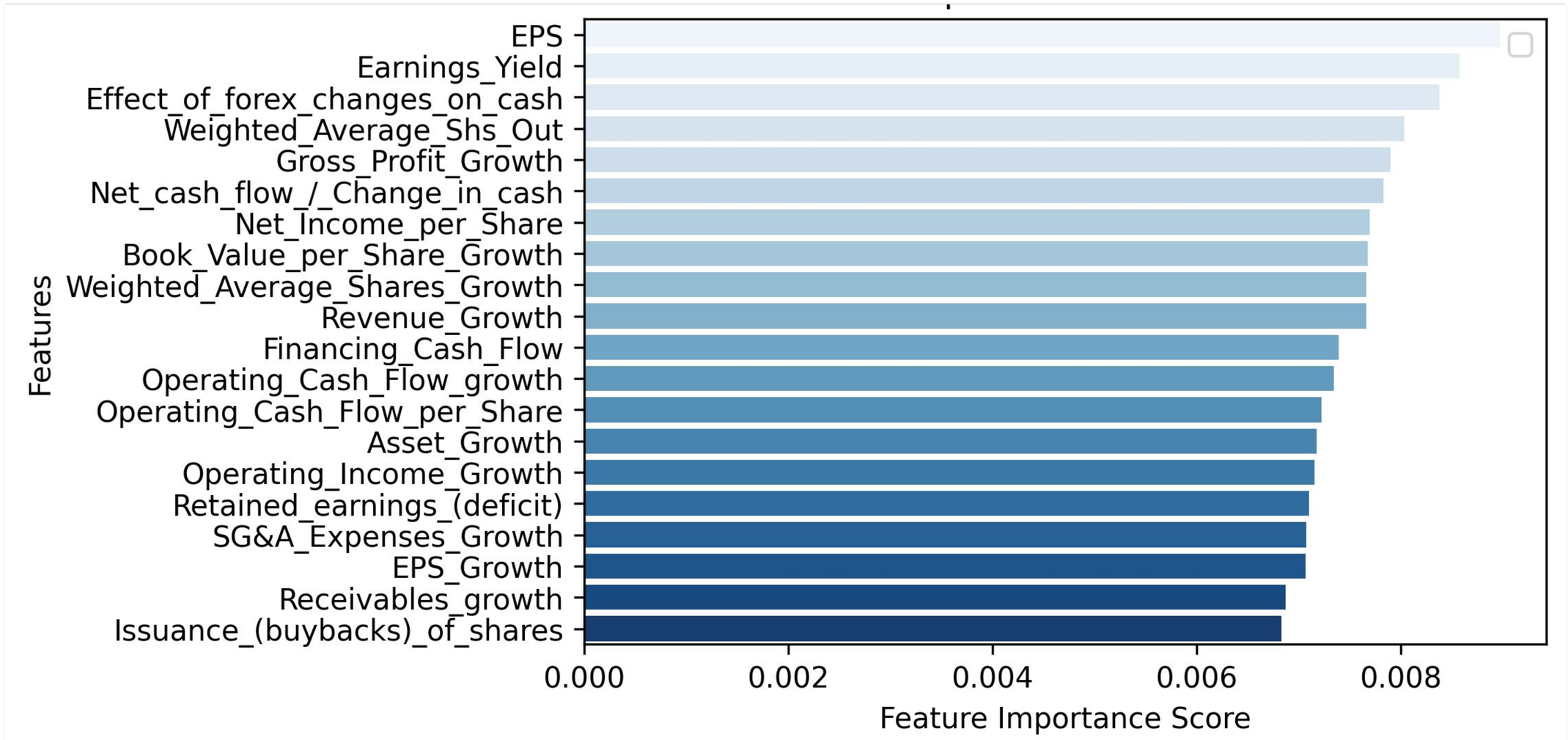
Back to the main table

Financial Indicators for US Stock Market

Xiaoqi Hu, Yanni Lan, Chenhang Niu, Yuyang Shu, Siqi Zhang

Boston University Questrom School of Business, Master of Business Analytics 2020

RESULT PLOT



Back to the main table

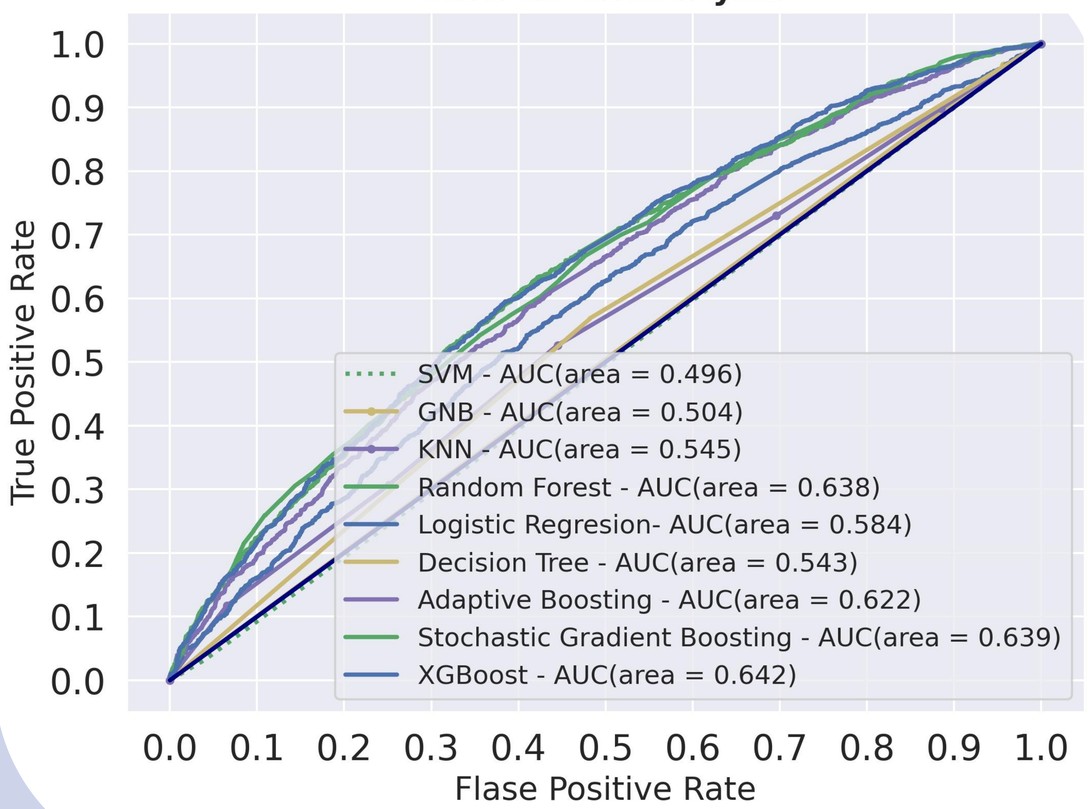
Financial Indicators for US Stock Market

Xiaoqi Hu, Yanni Lan, Chenhang Niu, Yuyang Shu, Siqi Zhang

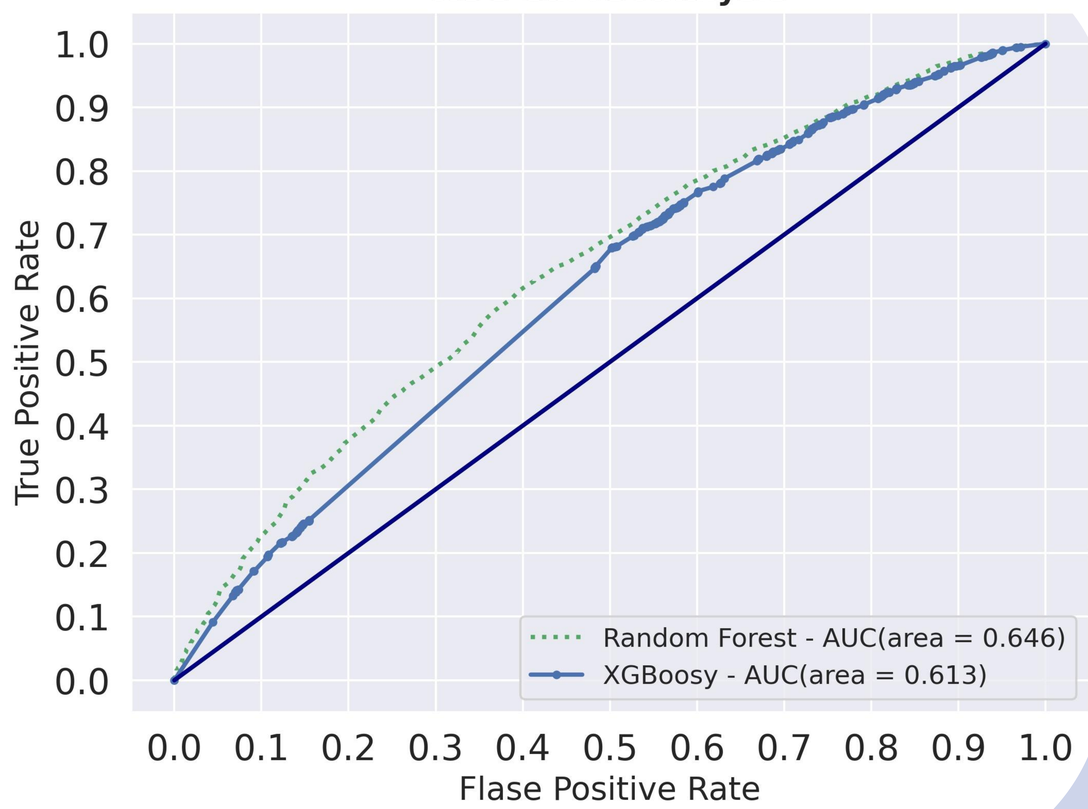
Boston University Questrom School of Business, Master of Business Analytics 2020

RESULT PLOT

ROC Curve: All Model



ROC Curve: After Tuning Hyper-parameter



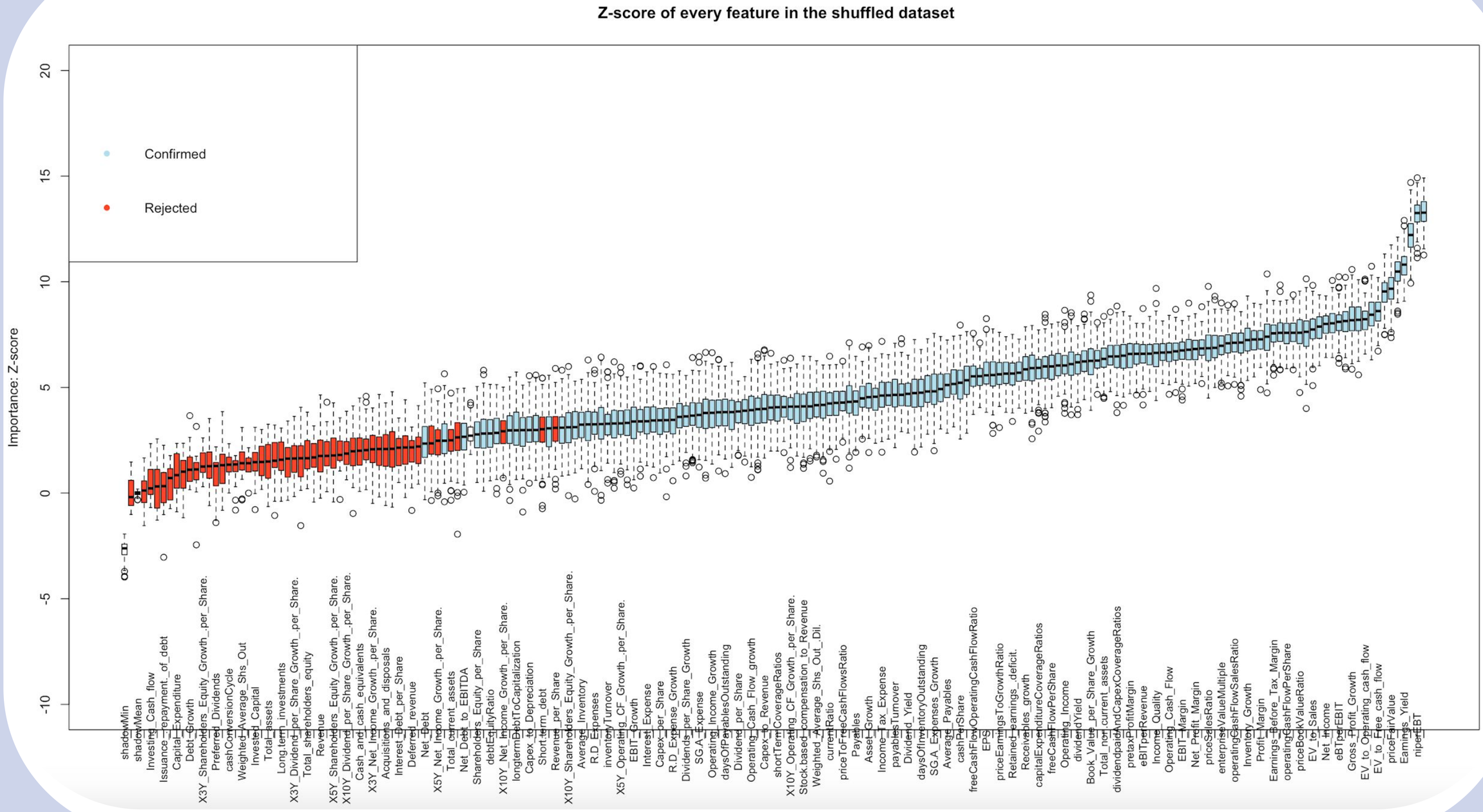
Back to the main table

Financial Indicators for US Stock Market

Xiaoqi Hu, Yanni Lan, Chenhang Niu, Yuyang Shu, Siqi Zhang

Boston University Questrom School of Business, Master of Business Analytics 2020

RESULT PLOT



Back to the main table

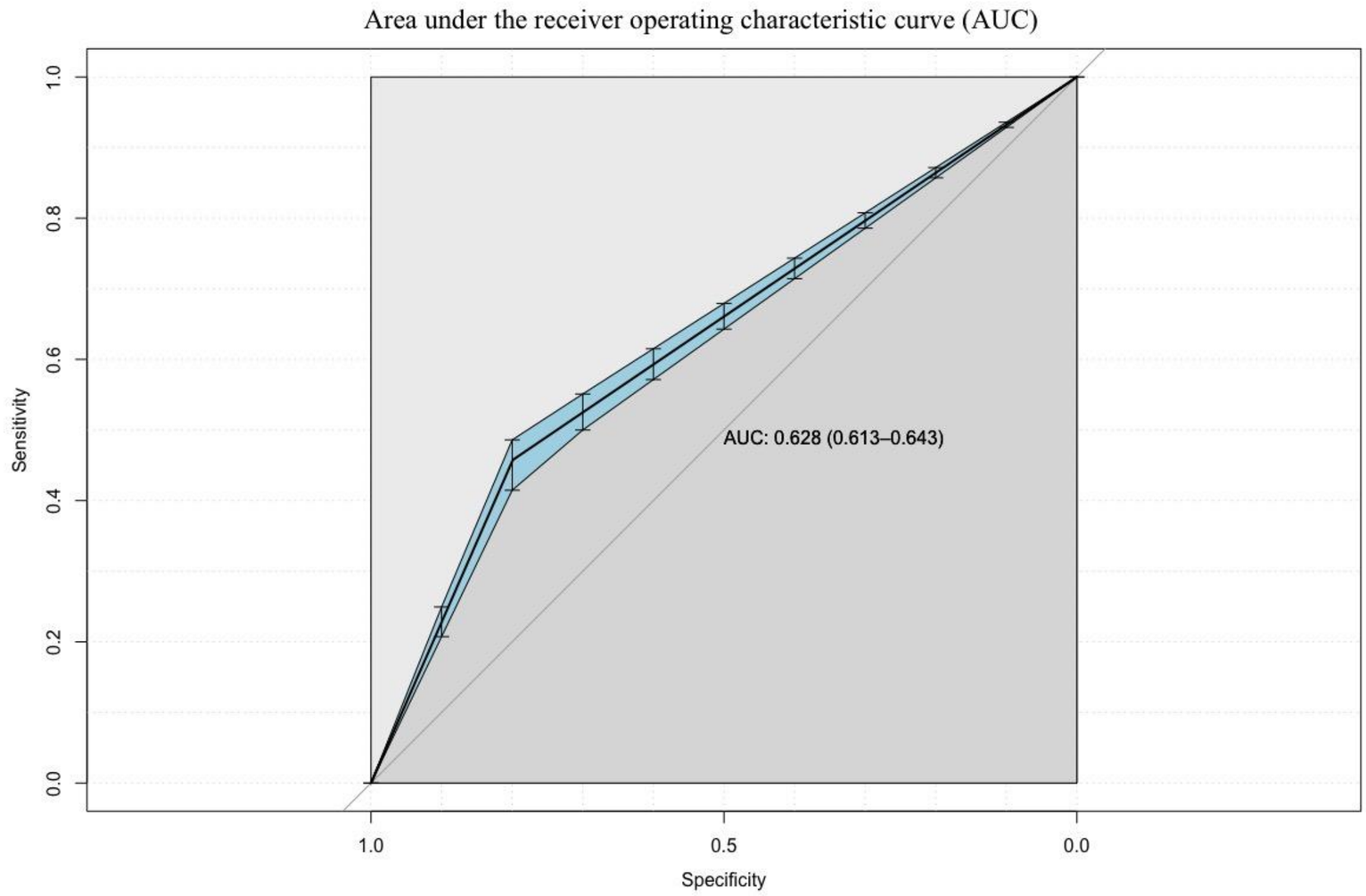
Financial Indicators for US Stock Market

Xiaoqi Hu, Yanni Lan, Chenhang Niu, Yuyang Shu, Siqi Zhang

Boston University Questrom School of Business, Master of Business Analytics 2020

RESULT PLOT

```
Confusion Matrix and Statistics

          Reference
Prediction    1    0
         1 2434  730
         0  612  616

               Accuracy : 0.694
                 95% CI : (0.681, 0.708)
    No Information Rate : 0.694
    P-Value [Acc > NIR] : 0.4552

                  Kappa : 0.263

 Mcnemar's Test P-Value : 0.0014

            Sensitivity : 0.799
            Specificity : 0.458
         Pos Pred Value : 0.769
         Neg Pred Value : 0.502
             Prevalence : 0.694
         Detection Rate : 0.554
   Detection Prevalence : 0.720
      Balanced Accuracy : 0.628

       'Positive' Class : 1
```



Back to the main table