MSBA 2020 CAPSTONE PROJECT

FINANCIAL INDICATORS OF US STOCK MARKET

Cohort B Team 1:
Xiaoqi Hu, Yanni Lan, Chenhang Niu, Yuyang Shu, Siqi Zhang

AGENDA

TOPIC INTRODUCTION

Topic Introduction and objectives

DATASET INTRODUCTION

Dataset Introduction and Project Workflow

DATASET EXPLORATION

Exploratory Data Analysis

MODEL RESULT

Compare Different Model Results

CONCLUSION

Executive Summary and Finding

01. TOPIC INTRODUCTION

We noticed machine learning is gradually being applied into the finance industry, helping asset management companies to find new investment opportunities and seek for better alpha. As an investor, what would be the most valuable factors from financial reports?

OBJECTIVES

FEATURE SELECTION

Using Different Feature Engineering methods to filter the valuable factors that could apply into the models

PREDICTION

Improving the model accuracy in the prediction of stock volatility

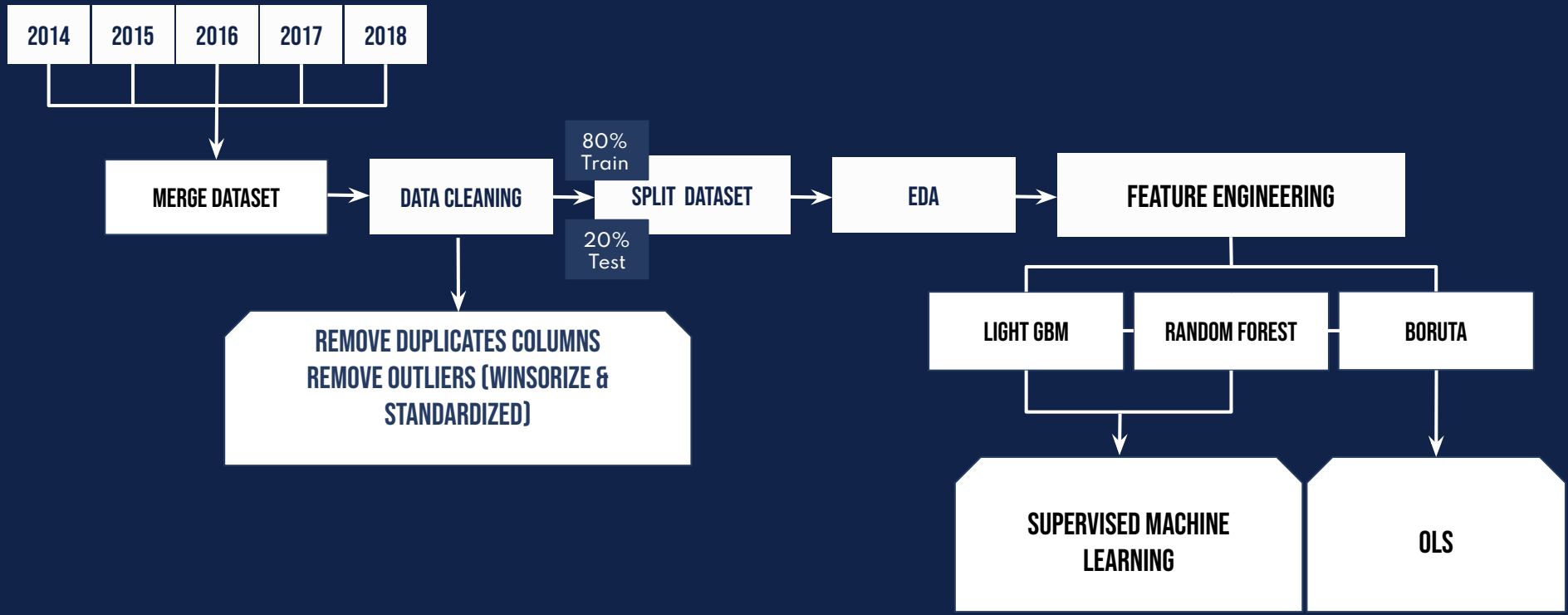
02.
DATASET INTRODUCTION

DATASET INTRODUCTION

DATA SOURCE: KAGGLE 200+ Financial Indicators of US stocks (2014-2018)

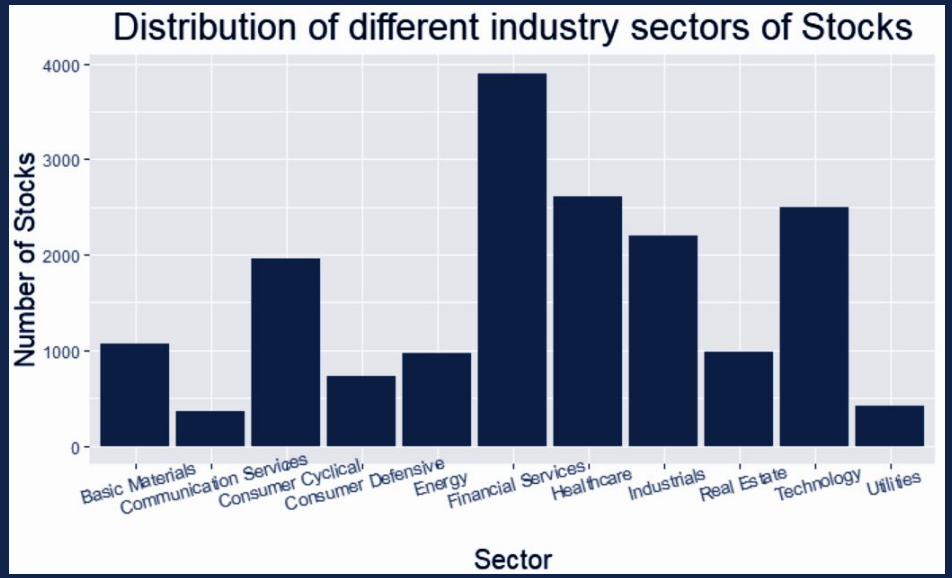
- Five year entry : from 2014 to 2018
- 22,075 Rows and 226 Columns in total
- Prediction Class: **Class 1** (Worth Buying) ; **Class 0** (Not Worth Buying)
- The Classification is price variation compare with last Fiscal year company stock price
- Four of the industrial sectors have over 2000 stocks; Financial Service, Healthcare, Industrials, and Technology

PROJECT WORKFLOW



DATASET OVERVIEW

Data Type	Number of Class 0	Percentage of Class 0	Number of Class 1	Percentage of Class1	Total Number of Observations
Train Data	7,975	45.2%	9,685	54.8%	17,660
Test Data	1,942	44.0%	2,473	56.0%	4,415



Default Limitation

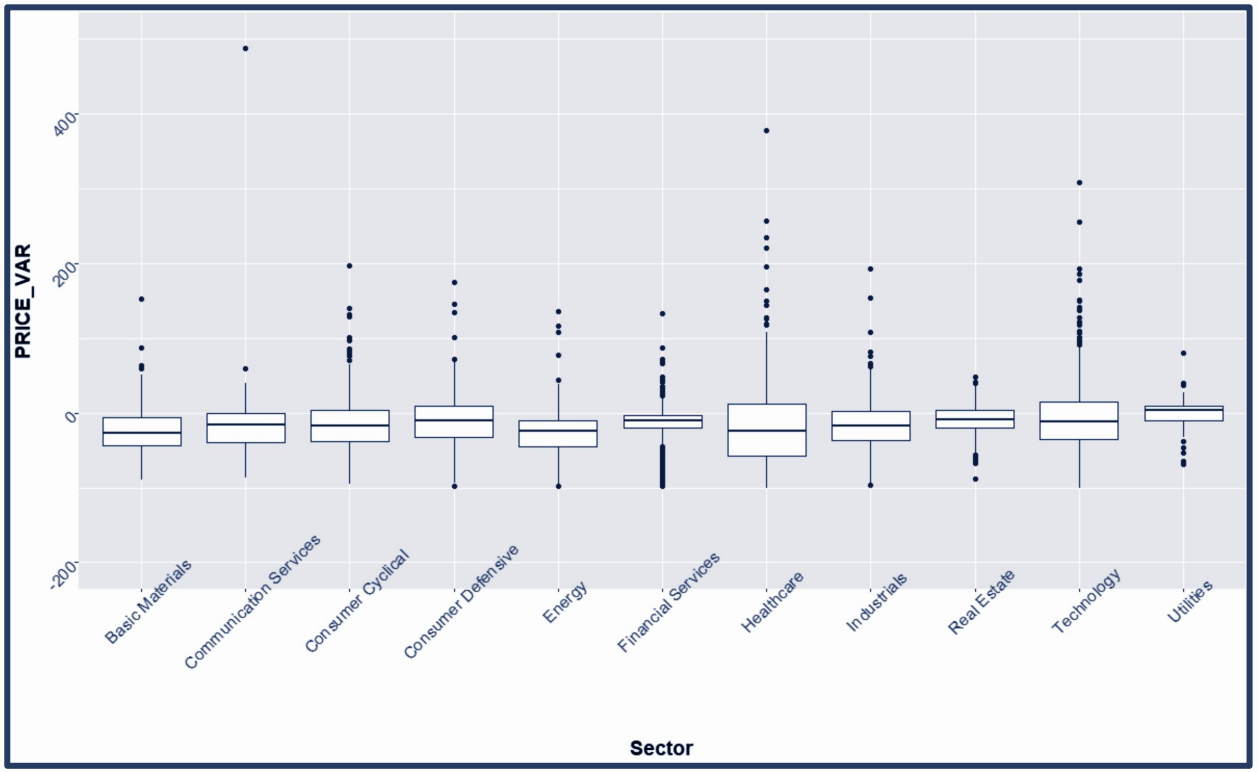
Looking in the summary of our train dataset, we have 22,075 rows of entries and 226 variables. The ratio between data and its variables is 77.91. Have too many variables that might create noise to the model prediction. We will narrow down the number of variables that are most related to the stock performances by applying feature engineering.

03.
DATA EXPLORATION

Exploratory Data Analysis

PRICE VARIATION BY INDUSTRY SECTORS

Compared with different sectors, utilities have the highest median value. The basic materials sector has low median value, it matches today's' market.



VALUABLE STOCKS

Sector	count	meanincrase
Healthcare	25	196.0345
Technology	20	2326.2810
Consumer Cyclical	5	139.4767
Consumer Defensive	4	138.8442
Energy	3	119.7087
Industrials	3	151.4514
Communication Services	2	731.6235
Basic Materials	1	152.6585
Financial Services	1	132.9897

In our dataset, we picked stocks with an increase rate more than 100 percent, and group them by their sectors. We can find Healthcare and Technology has the most counts. That is because Healthcare Sector has the most companies associated with high increase rates, and For Many Tech Companies, they usually have very high gross profit margin

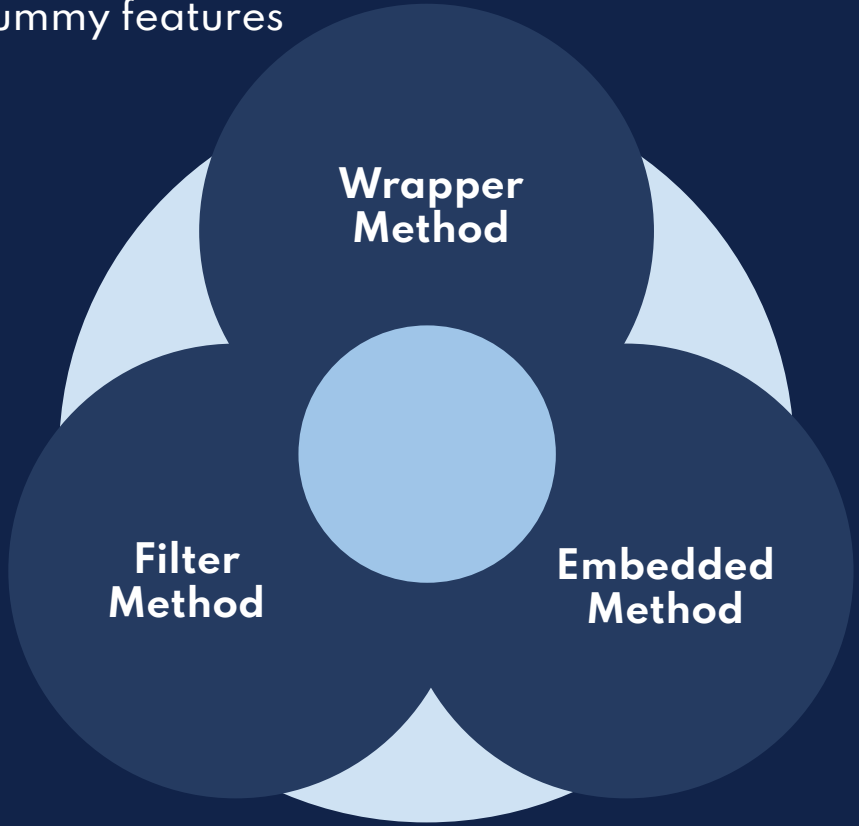
04.

RESULT

Introduce our feature selection and models

FEATURE ENGINEERING

Doing feature selection based on the cleaned data:
206 features, included 11 dummy features

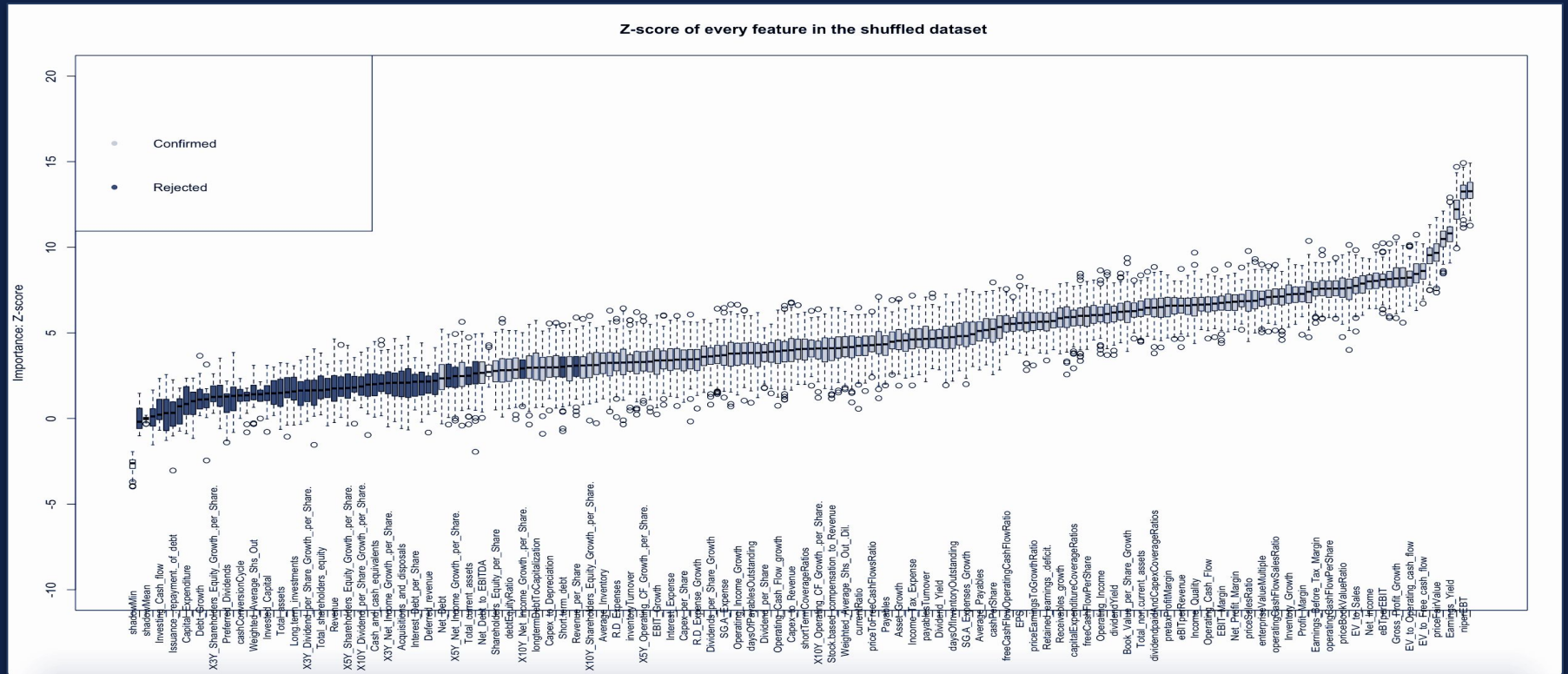


- Distance metrics
- Correlation
- Mutual Information
- Consistency metrics

- Decision Tree
- LASSO & Ridge

FEATURE ENGINEERING I

Based on the Boruta Package in R, we selected out 131 Variables out of 206 Variables base on their importance.



MODEL RESULT

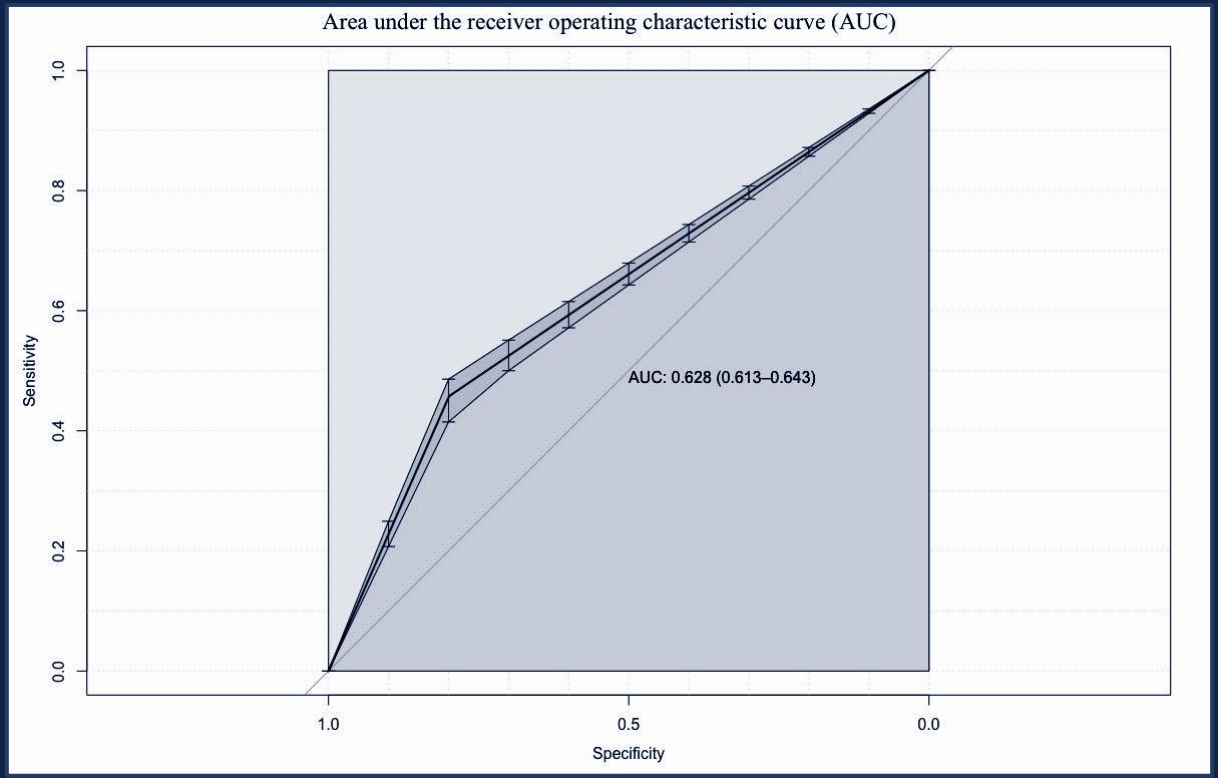
And out of 131 variables we selected 42 variables according to the model performance result. The following are the 42 selected variables' names:

- 'nIperEBT',
- 'Effect_of_forex_changes_on_cash',
- 'Earnings_Yield',
- 'effectiveTaxRate',
- 'priceFairValue',
- 'SG.A_to_Revenue',
- 'EV_to_Free_cash_flow',
- 'Weighted_Average_Shares_Growth',
- 'Gross_Profit_Growth',
- 'assetTurnover',
- 'eBTperEBIT',
- 'Net_Income',
- 'Net_Income_Com',
- 'EV_to_Sales',
- 'priceToOperatingCashFlowsRatio',
- 'priceToBookRatio',
- 'priceBookValueRatio',
- 'Enterprise_Value_over_EBITDA',
- 'operatingCashFlowPerShare',
- 'Earnings_Before_Tax_Margin',
- 'Revenue_Growth',
- 'Profit_Margin',
- 'Inventory_Growth',
- 'Free_Cash_Flow_Yield',
- 'operatingCashFlowSalesRatio',
- 'grossProfitMargin',
- 'Earnings_before_Tax',
- 'enterpriseValueMultiple',
- 'Gross_Margin',
- 'Net_Income_per_Share',
- 'priceSalesRatio',
- 'Net_Profit_Margin',
- 'netProfitMargin',
- 'EBIT_Margin',
- 'EBIT',
- 'Operating_Cash_Flow',
- 'eBITperRevenue',
- 'Free_Cash_Flow_margin',
- 'Income_Quality',
- 'pretaxProfitMargin',
- 'EBITDA_Margin',
- 'Sector' (Categorical Variables)

MODEL RESULT

With top 42 variables we get an prediction accuracy of 0.694 in test dataset, and we also get an AUC of 0.628 which is meaningful than randomly select stocks from the market.

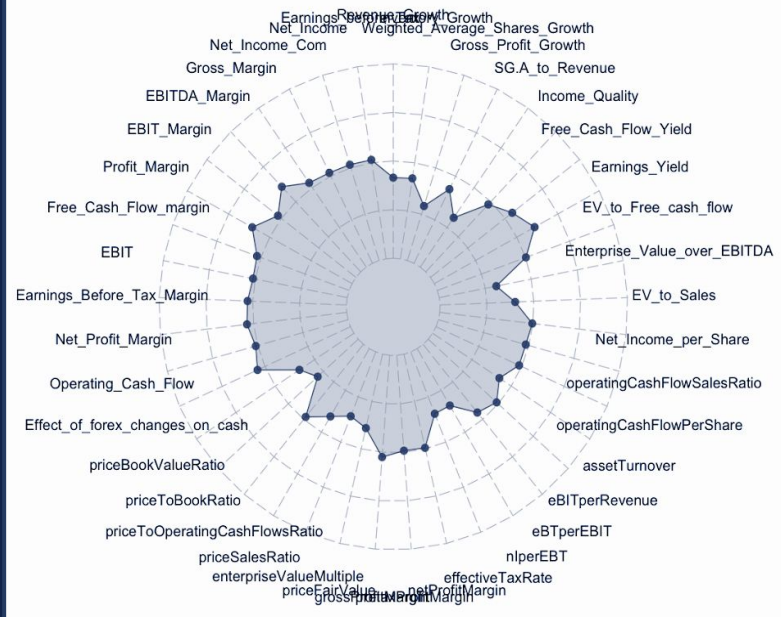
```
Confusion Matrix and Statistics

          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
```



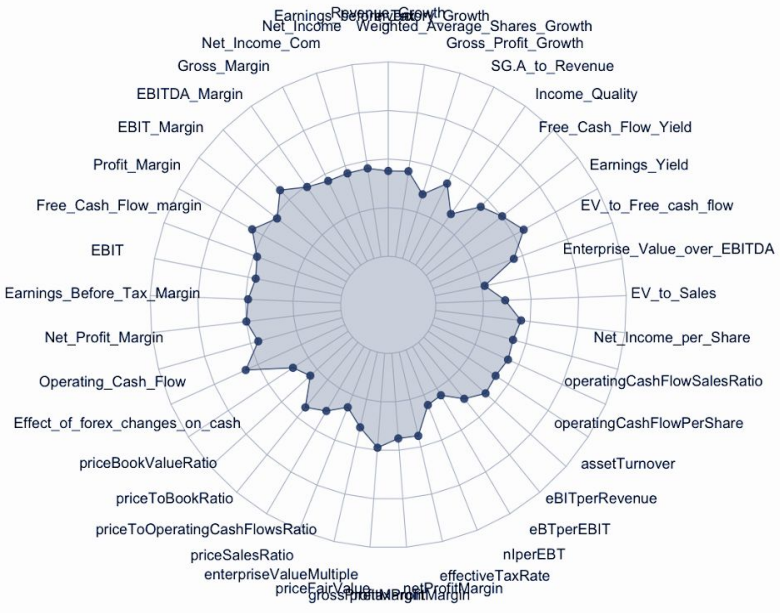
MODEL RESULT

This graph shows the characteristics of two different class base on the 41 numerical variables.

CLASS 1 RADAR CHART

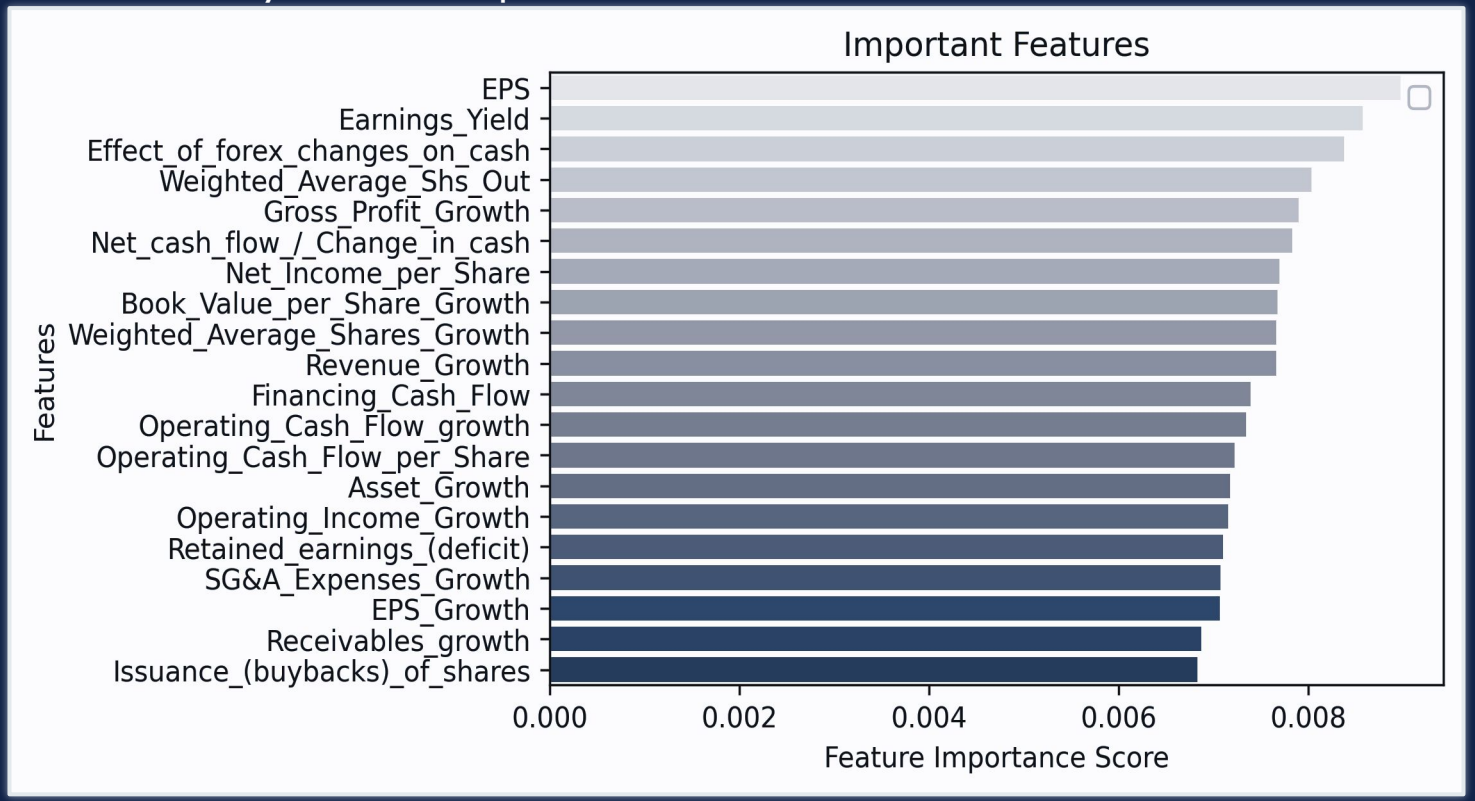


CLASS 0 RADAR CHART



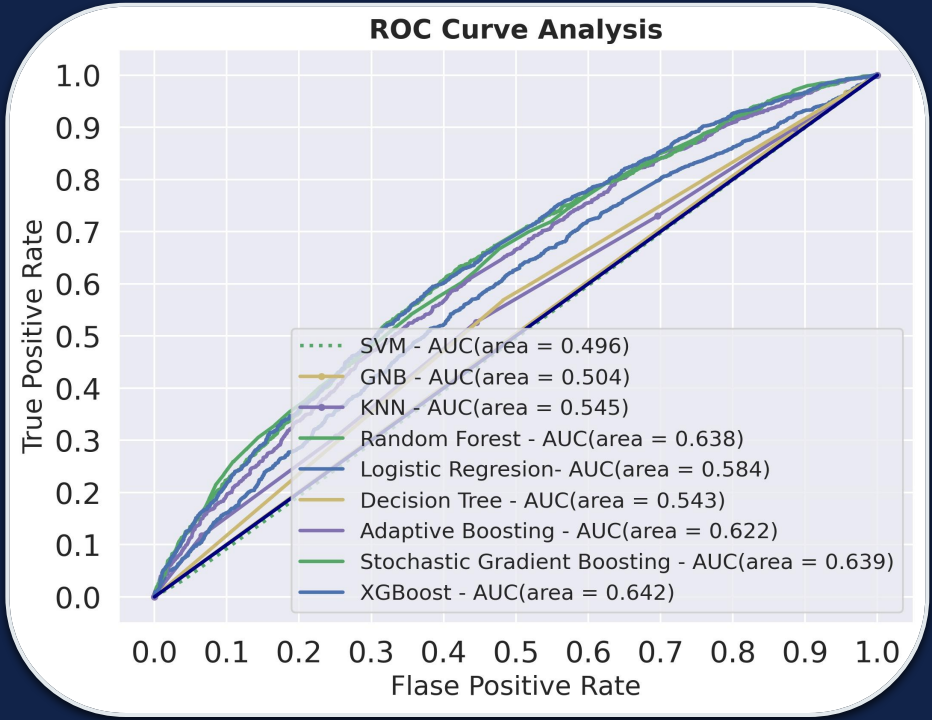
FEATURE ENGINEERING II

Using Random Forest, we filter 20 variables out of 206 variables and order them by feature importances.



MODEL RESULT

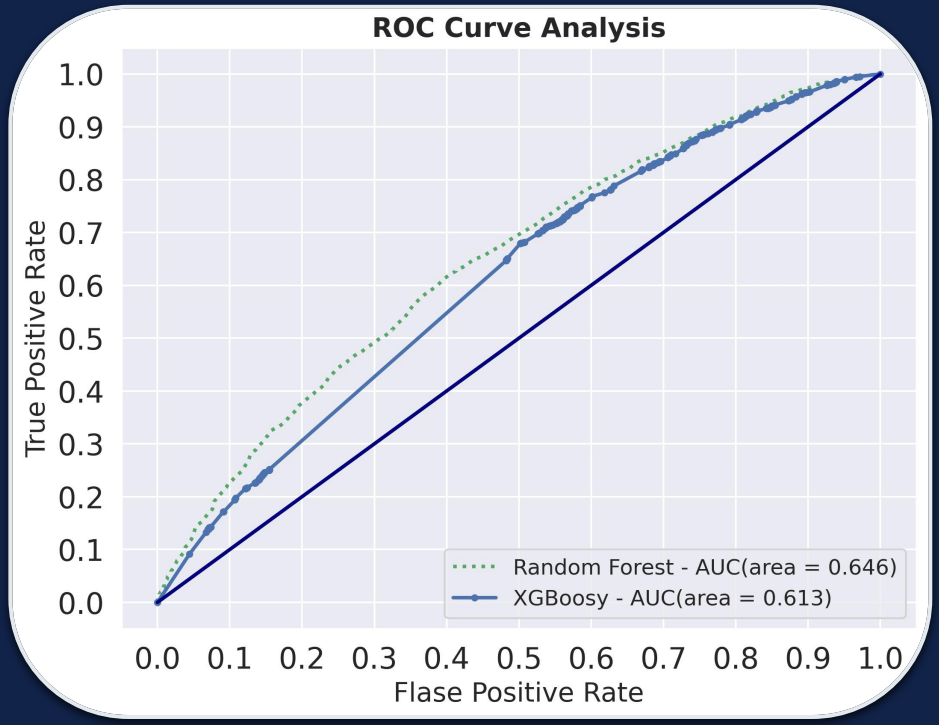
Building on the 20 variables, we further implement supervised machine learning methods.



Model	Accuracy
Random Forest ☑	60.1%
XGBoost ☑	60.1%
Stochastic Gradient Boosting	60.1%
KNN	58.8%
Logistic Regression	58.1%
Naive Bayes	57.9%
Decision Tree	54.7%

MODEL RESULT

Random Forest surpasses than XGBoost on the measure of accuracy after hyperparameter tuning.



Random Forest

After tuning the hyper-parameters, accuracy increased slightly, from **60.14%** to **61.33%**.

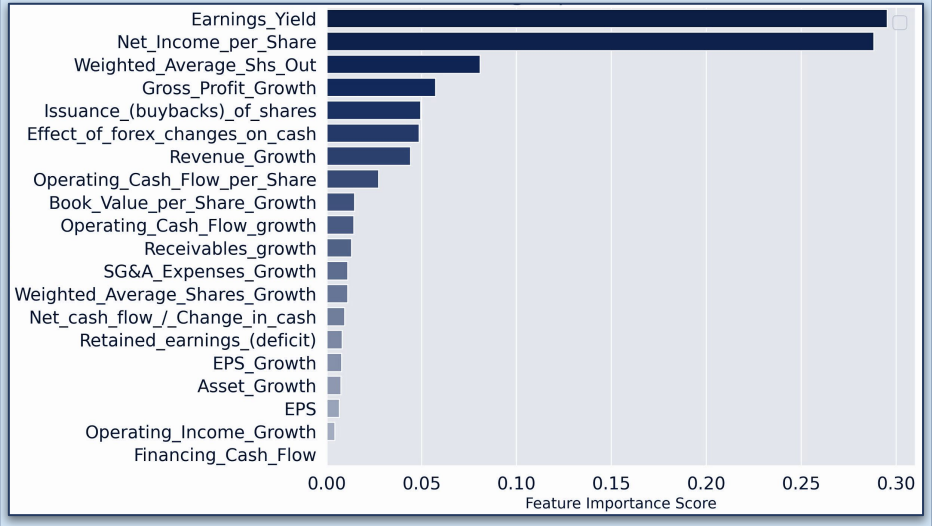
XGBoost

After tuning the hyper-parameters, accuracy increased slightly from **60.14%** to **60.20%**.

MODEL RESULT

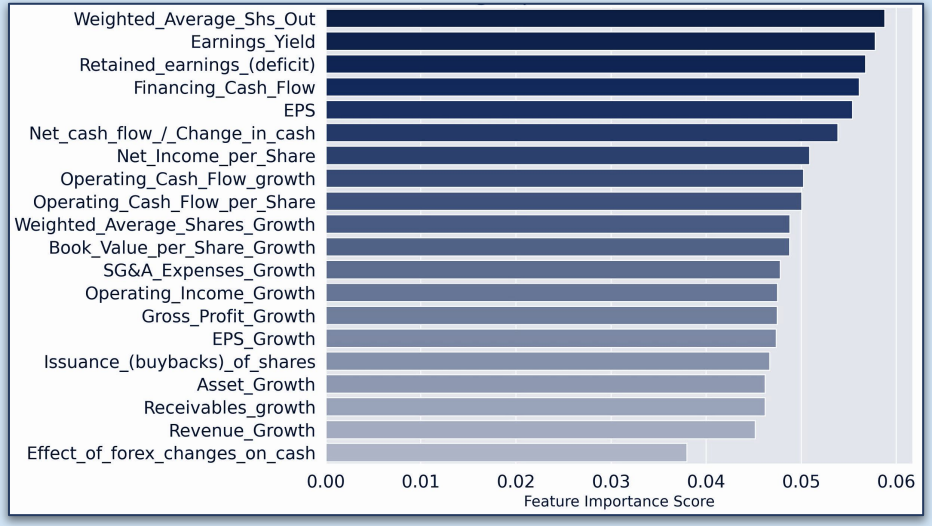
The influence of earnings yield and weight share outstanding remain strong after hyperparameter tuning.

FEATURE IMPORTANCE VIA XGBOOST



Earnings Yield and **Net Income per Share** stand out as the two most important features

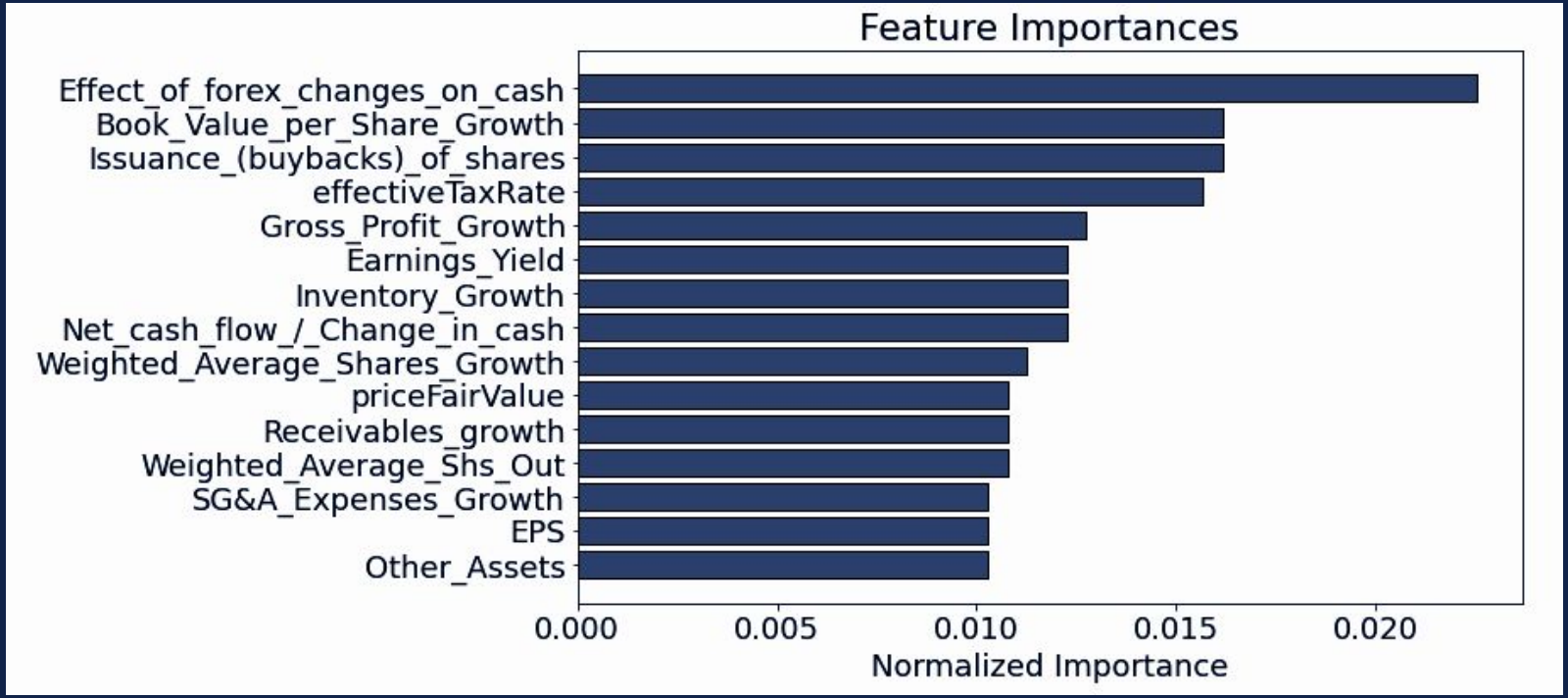
FEATURE IMPORTANCE VIA RANDOM FOREST



Weighted Average Share Outstanding, **Return Earnings(Deficit)** also reveal their importance

FEATURE ENGINEERING III

Using Light GBM, we filter 15 variables out of 206 variables.



MODEL RESULT

UNDER DEFAULT SETTING

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

AFTER TUNING HYPER-PARAMETER

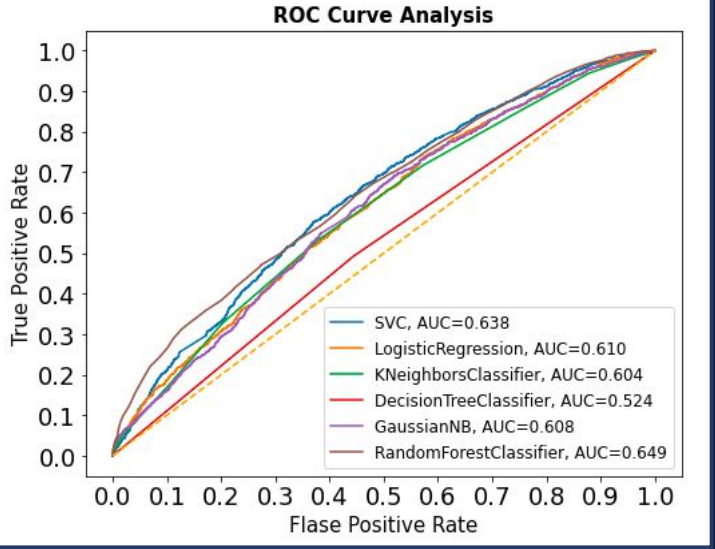
Model	Accuracy
SVM	61.3%
Random Forest	61.3%

The accuracy of SVM **doesn't change**

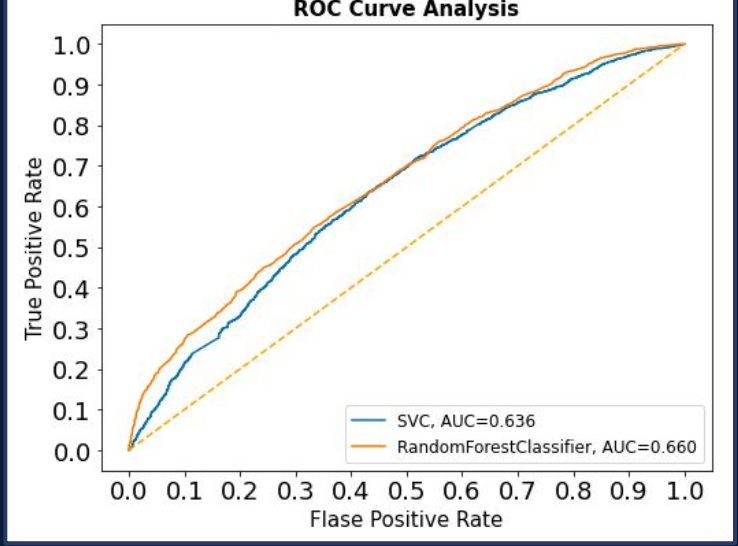
The accuracy of random forest **slightly increased**

MODEL RESULT

UNDER DEFAULT SETTING



UNDER TUNING HYPER-PARAMETER



The AUC of Random Forest increased over 1%. Compared with the best two models SVM and Random Forest, Random Forest will be the best choice under Light GBM feature selection.

RESULTS COMPARISON

LIGHT GBM

- Model: Random Forest
- Accuracy : 61.3%
- AUC: 66.0%

RANDOM FOREST

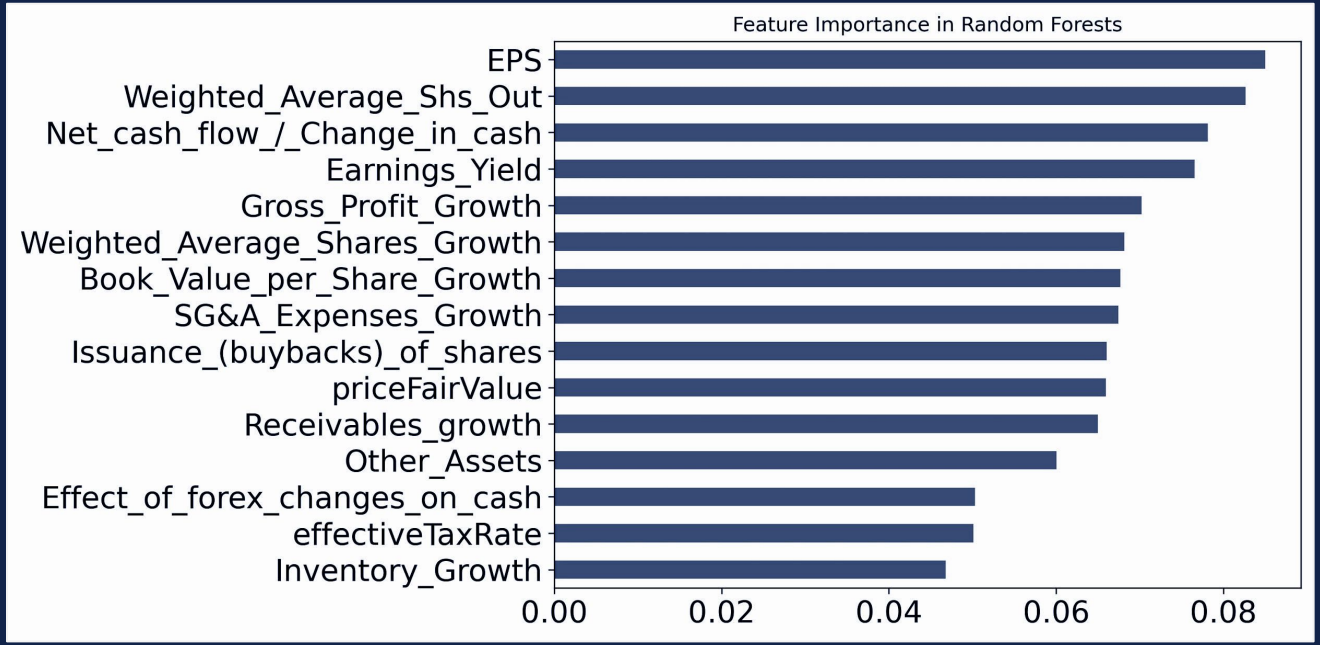
- Model: Random Forest
- Accuracy : 61.3%
- AUC: 64.6%

BORUTA

- Model: OLS
- Accuracy : 69.4%
- AUC: 62.8%

FEATURE IMPORTANCE

Feature importance in Random Forests with best hyper-parameter.



EPS, **Weighted Average Shares Outstanding**, **Net Cash Flow** generated most in the model

05.

CONCLUSION

SUMMARY BEHIND THE RESULTS

EARNING PER SHARE

EPS is positively affected the stock price. It represent the amount of income that company generated for each share of stock.

WEIGHTED AVERAGE SHARE OUTSTANDING

This is a calculations that could help investor compile a position in a stock over several year even they hold for long term investment.

NET CASH FLOW

Companies with higher net cash flows at lower risks means higher share price valuation.

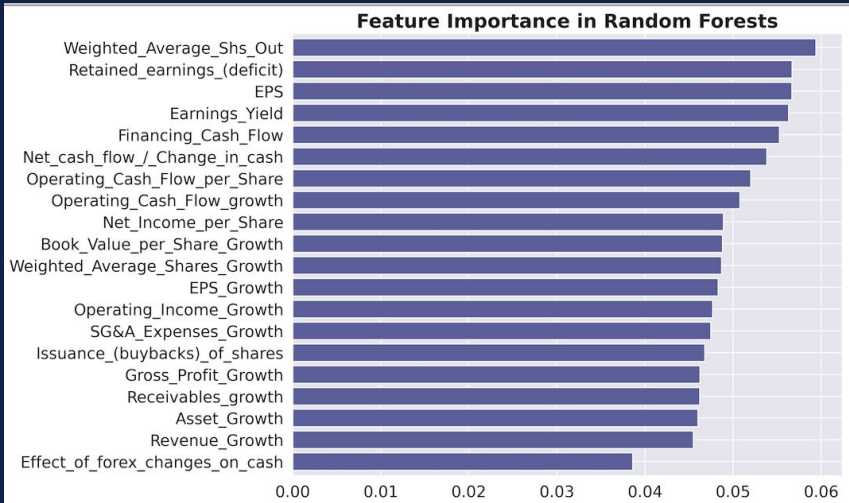
THANKS!

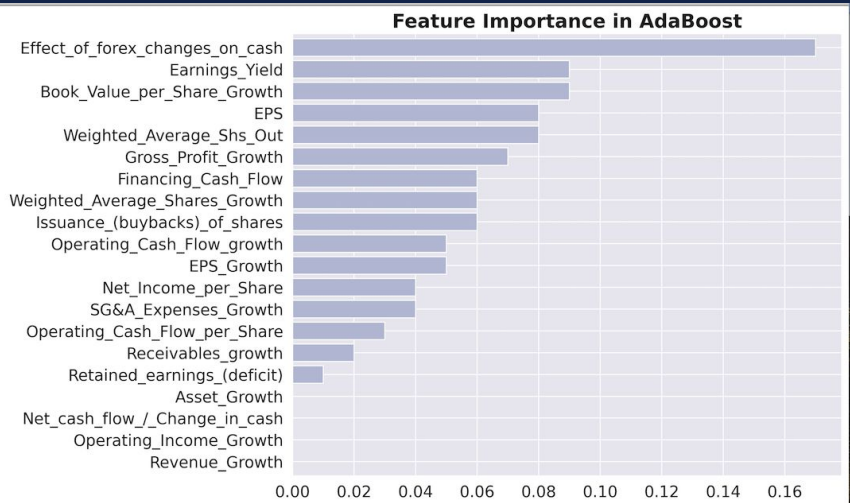
Any questions?

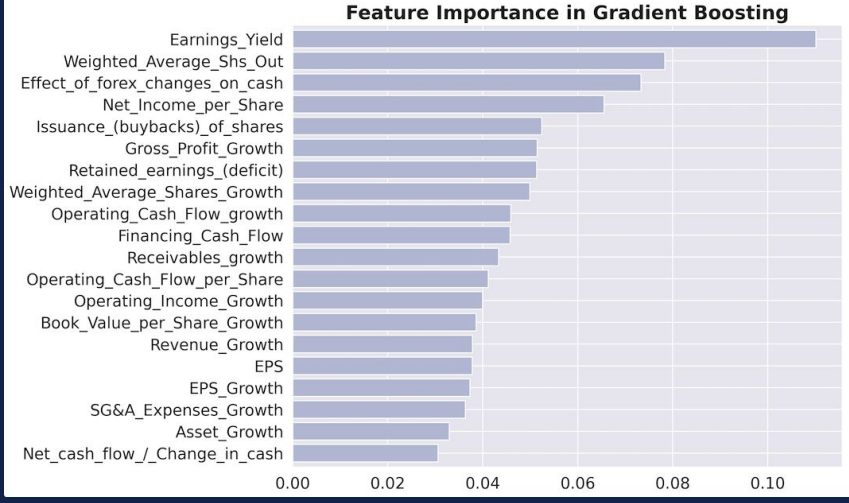
BOSTON UNIVERSITY

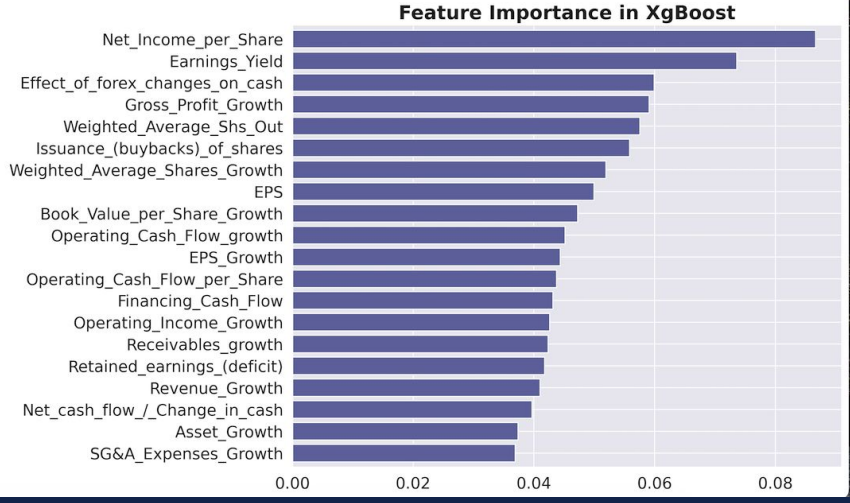
	CV Mean	Std
Radial SVM	0.548584	0.010507
KNN	0.555198	0.003227
Naive Bayes	0.506319	0.006117
Random Forest	0.606523	0.008319
Logistic Regression	0.576444	0.007935
Decision Tree	0.546954	0.003548
Voting Classifier	0.573998	0.004761
Bagged KNN	0.547407	0.003516
AdaBoost(Adaptive Boosting)	0.603715	0.004175
Stochastic Gradient Boosting	0.611234	0.004907
XGBoost	0.606840	0.006958

```
features = pd.DataFrame()
i = 1
while i < 31:
  X_train_fe, X_valid, y_train_fe, y_valid = train_test_split(X_train, y_train, test_size=0.10, random_state=0)
  # calculate mean and standard deviation of train set
  X_train_fe = StandardScaler().fit(X_train_fe).transform(X_train_fe)
  X_valid = StandardScaler().fit(X_valid).transform(X_valid)
  #Create a Gaussian Classifier
  clf=RandomForestClassifier(n_estimators=100)
  #Train the model using the training sets y_pred=clf.predict(X_test)
  clf.fit(X_train_fe,y_train_fe)
  y_pred=clf.predict(X_valid)
  feature_imp = pd.DataFrame( clf.feature_importances_,index=X_train.columns).sort_values( by = 0, ascending=False)
  features =  pd.concat([features, feature_imp], axis=1)
  i = i + 1
```


Feature Importance in Random Forests
Weighted_Average_Shs_Out
Retained_earnings_(deficit)
EPS
Earnings_Yield
Financing_Cash_Flow
Net_cash_flow_/_Change_in_cash
Operating_Cash_Flow_per_Share
Operating_Cash_Flow_growth
Net_Income_per_Share
Book_Value_per_Share_Growth
Weighted_Average_Shares_Growth
EPS_Growth
Operating_Income_Growth
SG&A_Expenses_Growth
Issuance_(buybacks)_of_shares
Gross_Profit_Growth
Receivables_growth
Asset_Growth
Revenue_Growth
Effect_of_forex_changes_on_cash
0.00
0.01
0.02
0.03
0.04
0.05
0.06


Feature Importance in AdaBoost
Effect_of_forex_changes_on_cash
Earnings_Yield
Book_Value_per_Share_Growth
EPS
Weighted_Average_Shs_Out
Gross_Profit_Growth
Financing_Cash_Flow
Weighted_Average_Shares_Growth
Issuance_(buybacks)_of_shares
Operating_Cash_Flow_growth
EPS_Growth
Net_Income_per_Share
SG&A_Expenses_Growth
Operating_Cash_Flow_per_Share
Receivables_growth
Retained_earnings_(deficit)
Asset_Growth
Net_cash_flow_/_Change_in_cash
Operating_Income_Growth
Revenue_Growth
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16


Feature Importance in Gradient Boosting
Earnings_Yield
Weighted_Average_Shs_Out
Effect_of_forex_changes_on_cash
Net_Income_per_Share
Issuance_(buybacks)_of_shares
Gross_Profit_Growth
Retained_earnings_(deficit)
Weighted_Average_Shares_Growth
Operating_Cash_Flow_growth
Financing_Cash_Flow
Receivables_growth
Operating_Cash_Flow_per_Share
Operating_Income_Growth
Book_Value_per_Share_Growth
Revenue_Growth
EPS
EPS_Growth
SG&A_Expenses_Growth
Asset_Growth
Net_cash_flow_/_Change_in_cash
0.00
0.02
0.04
0.06
0.08
0.10


Feature Importance in XgBoost
Net_Income_per_Share
Earnings_Yield
Effect_of_forex_changes_on_cash
Gross_Profit_Growth
Weighted_Average_Shs_Out
Issuance_(buybacks)_of_shares
Weighted_Average_Shares_Growth
EPS
Book_Value_per_Share_Growth
Operating_Cash_Flow_growth
EPS_Growth
Operating_Cash_Flow_per_Share
Financing_Cash_Flow
Operating_Income_Growth
Receivables_growth
Retained_earnings_(deficit)
Revenue_Growth
Net_cash_flow_/_Change_in_cash
Asset_Growth
SG&A_Expenses_Growth
0.00
0.02
0.04
0.06
0.08

THE FLOW

DATA CLEANING & EDA

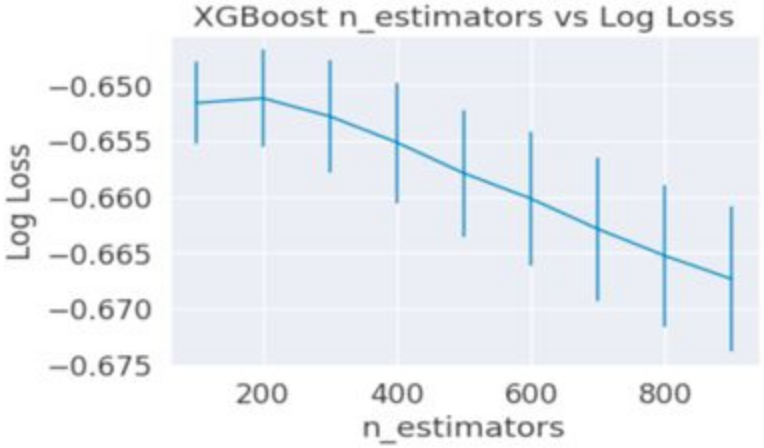
- Remove Duplicates Columns
- Replace NAs with 0
- Remove Outliers
  - Winsorize
  - Standardized
- Plotting

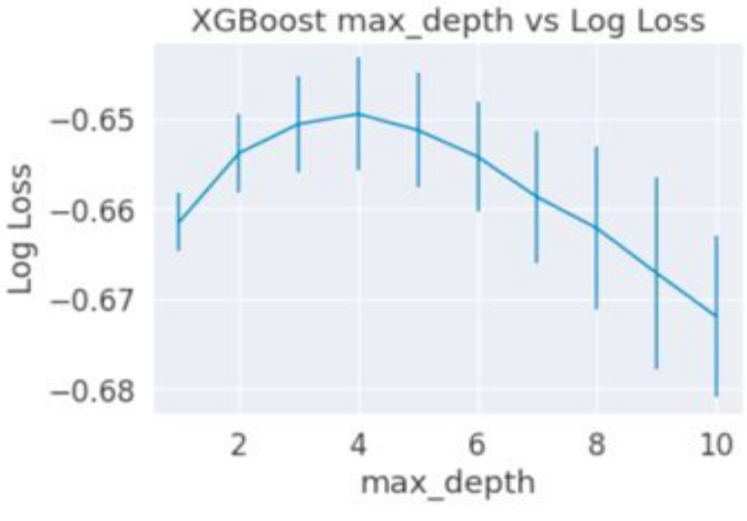
FEATURE SELECTION

- Filter Method
  - Mutual Information
- Embedded Method
  - Decision Tree

MODELING

- SVM
- Logistic Regression
- Random Forest
- Boosting





Hyper-parameters Tuning Process for XGBoost