DEPARTMENT OF ECE (PES UNIVERSITY)

COMPANY BANKRUPTCY PREDICTION

SUBMITTED BY:

Aryan Jain (PES1201800488)

FOR THE COURSE:

Machine Learning (UE18EC338)

UNDER THE GUIDANCE OF:

Prof. Vanamala HR (Department of ECE)

INDEX

INTRODUCTION

In the current age of startups, there is a huge increase in the number of companies coming up. But not all companies manage to stay successful over time and burn out pretty fast. So from all the financial and statistical details available from all the companies we perform our analysis and make do predictive analysis.

The aim of this project is to use these features to understand their impact/role on the selected models and how they can help us recognizing the companies that are close to bankruptcy.

PROBLEM STATEMENT

Building a Machine Learning model to predict if a company will go bankrupt on the basis of various financial parameters

METHODOLOGY

1. Perform EDA on the dataset and clean it
2. Remove outliers from the dataset to increase model efficiency
3. Perform Train-Test split
4. Analyze the most important parameters
5. Build multiple ML models using SMOTE (as data is imbalanced) and get accuracy and F1 scores
6. Select the best model using ROC and confusion matrix
7. Predict the target values for test set and check classification report.

SOFTWARE AND DATASET DETAILS

- **Coding Language:** Python
- **Platform:** Jupyter Notebooks
- **Libraries used:**
  - Numpy
  - Pandas
  - Seaborn
  - Sklearn
  - Imblearn
  - Xgboost
  - LazyPredict
  - DominanceAnalysis

SOFTWARE AND DATASET DETAILS

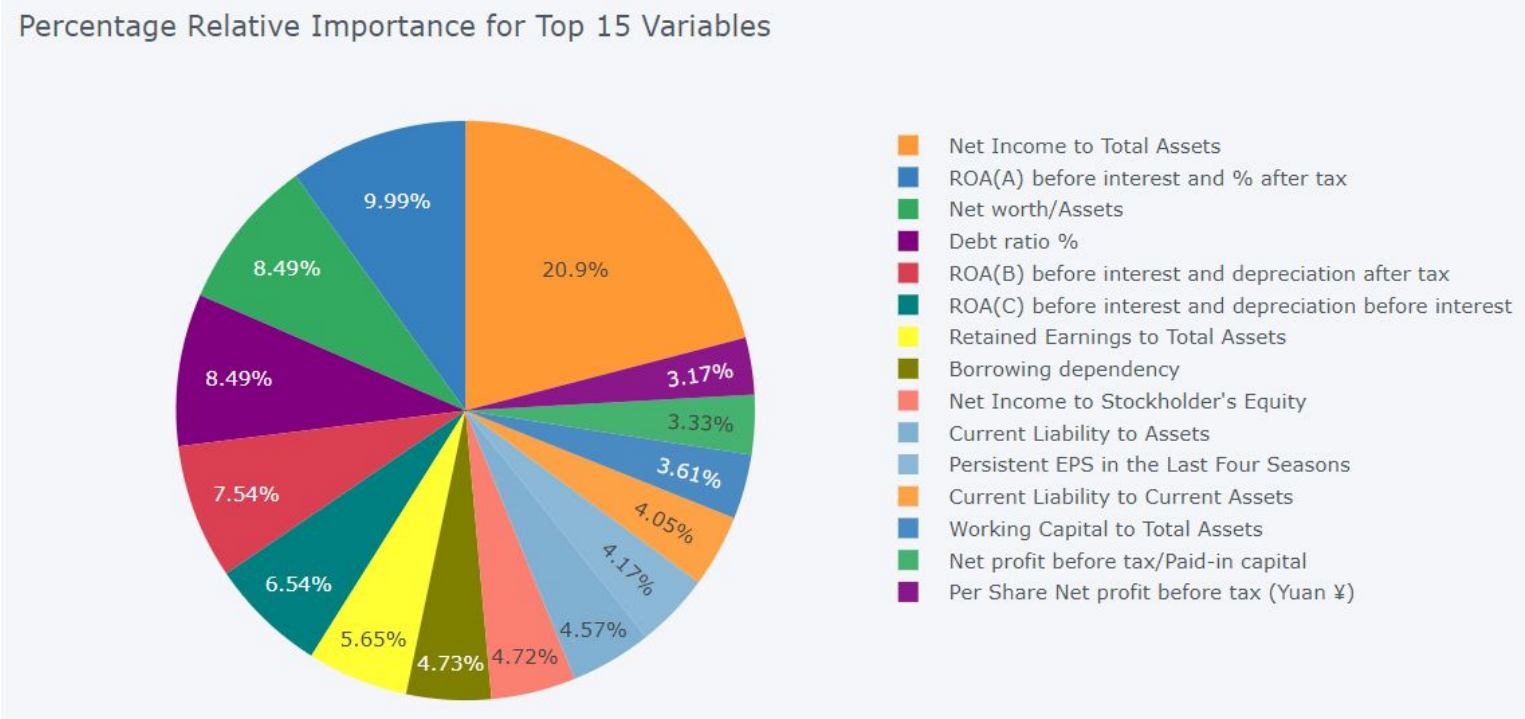
The data was collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.

The data was obtained from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

The dataset contains: 95 feature vectors (X1-X95) and 1 target vector (Bankruptcy). The feature vectors are: Net income, Net worth, Debt ratio% etc.

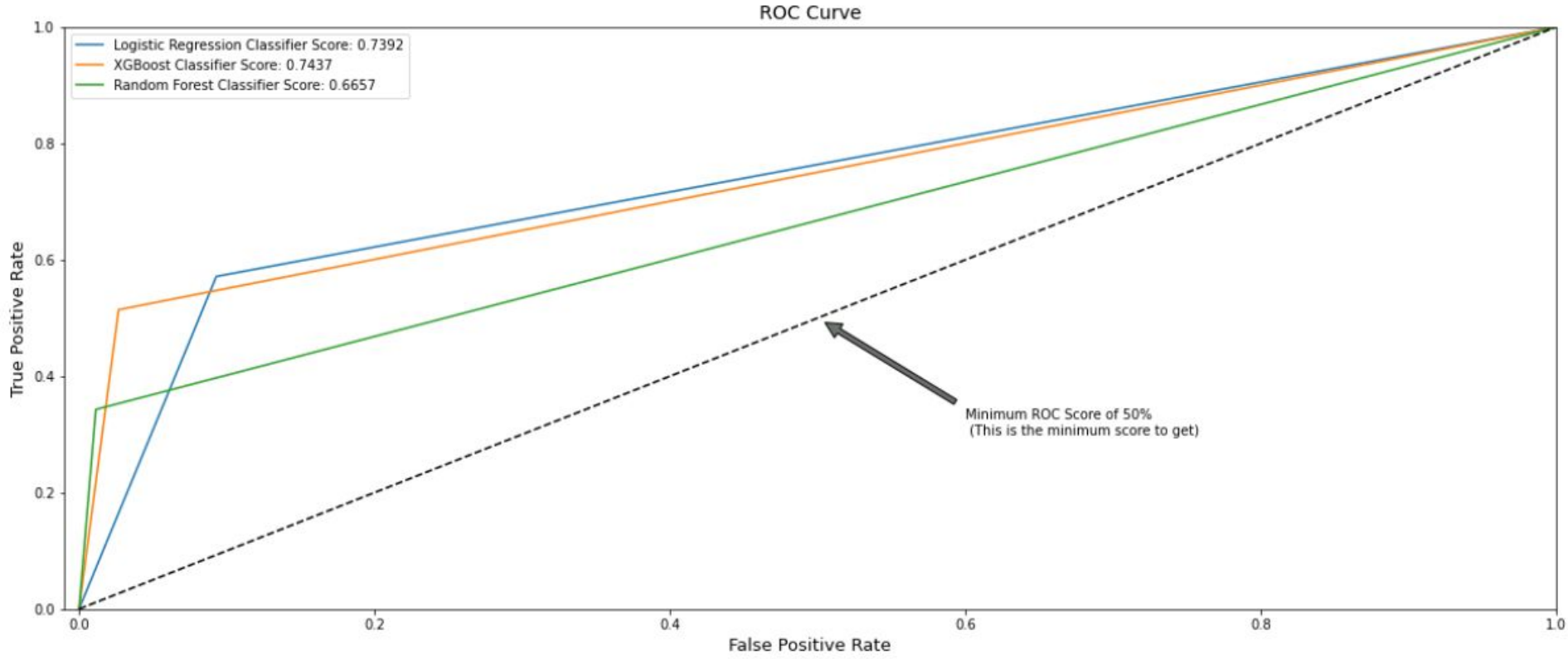
RESULTS AND ANALYSIS

The most important parameters were found out to be using dominant analysis:



RESULTS AND ANALYSIS

The comparison of the 3 ML models chosen:



RESULTS AND ANALYSIS

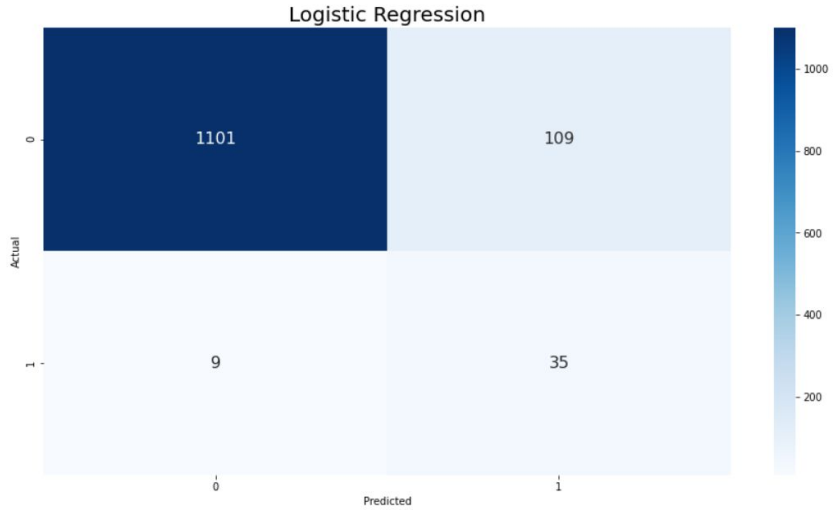
Owning to the ROC curve and Confusion matrix, **Logistic Regression** was chosen (as higher emphasis was there on True Negatives):

Testing data report:

	precision	recall	f1-score	support
Stable	0.98	0.91	0.94	968
Bankrupt	0.18	0.57	0.28	35
accuracy			0.90	1003
macro avg	0.58	0.74	0.61	1003
weighted avg	0.96	0.90	0.92	1003

Training data report:

	precision	recall	f1-score	support
Stable	0.99	0.91	0.95	1210
Bankrupt	0.24	0.80	0.37	44
accuracy			0.91	1254
macro avg	0.62	0.85	0.66	1254
weighted avg	0.97	0.91	0.93	1254



CONCLUSION AND FUTURE SCOPE

Owing to an imbalance dataset skewed in the favour of stable companies, it was observed that the accuracy was very high for those. Logistic Regression was chosen as it had a better accuracy for the true negatives. In a realistic scenario it is better to classify a stable company as bankrupt than vice-versa. The current F1 score is 0.37 with 80-20 split so it can definitely be improved upon.

To improve the F1 score, we can look into combining XGboost and Logistic Regression. It was also shown by LazyPredict that a perceptron model would work with a higher accuracy. The model can also be fine-tuned to fit only specific number of parameters which may affect accuracy slightly.

REFERENCES

1. Azen, Razia, and Nicole Traxel. "Using dominance analysis to determine predictor importance in logistic regression." *Journal of Educational and Behavioral Statistics* 34.3 (2009): 319-347.
2. Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572.