Project: Predicting a Start-ups Acquisition Status

Made by: Dieksha Priyaa Mishra

Position: Data Specialist, Intern

Company: Technocolabs

INTRODUCTION

- The project is about forecasting the acquisition status of start-ups.
- The goal of this research is to forecast the acquisition status of start-ups based on financial data from the firm.
- A constant predictor on an excessively skewed dataset can provide great accuracy but low precision.
- Problem Statement: How can we enhance precision while maintaining accuracy and avoiding over/under sampling techniques?
- Several models, such Baseline QDA (Quadratic Discriminant Analysis), RF (Random Forest), and Ensemble, have been used to tackle the following issue statement.

DATASETS

- In this project, a Crunchbase dataset called ‘Crunchbase 2013 - Companies, Investors, etc.' from Kaggle was used.
- There are n = 17,727 samples in the collection, and each row provides information about a start-up.
- Each row also includes the company's status ('Acquired,' 'Closed,' 'IPO,' or 'Operating’).
- The data se

IPO	CLOSED	ACQUIRED	OPERATING
1.9%	3.1%	9.4%	85.6%

Figure_1: Showcases that Operating class is over-represented and other classes under-represented.

- The data is divided into 60/20/20 segments for training, validation, and testing.
- The data has been pre-processed, and two transformations have been performed on the dataset, including: converting the date feature from strings to two numbers corresponding to the month and year. The headquarters location feature, when transformed from strings to floats, correlates to the GeoPy API's longitude, latitude, and city/state 'importance.'

FEATURES SELECTION

- The dataset in this project includes features such as company name, permalink, category, financing dates, fundraising rounds, funding amount, city, state, and so on.
- Extraction of features: Converting qualitative input to quantitative data –

Dates Feature – Strings into two integers corresponding to Month and Year

Headquarter Location — Converts strings to floats for Latitudes, Longitudes, and so on.

- The method for selecting feature combinations is based on the **forward selection algorithm**, which greedily adds features that provide the highest accuracy performance when applied to the validation set.
- Using this method, we can observe that characteristics like the number of funding rounds, total funding, and the relevance of a company's headquarters location are consistently picked, implying that they are good predictors of a start-up's acquisition status.
- Each model is trained on multiple combinations of features to reduce overfitting and discover the best feature selection for QDA and RF classifiers.
- Due to time constraints, the features employed in the two-step ensemble approach are those acquired for the individual sub-models.
- This assures that the performance when the features are carefully chosen to optimise for the two-step ensemble approach will be even better, resulting in findings that indicate a lower bound on the ideal performance.

MODELS IMPLEMENTATION

Logistic Regression Model: This model has been applied initially due to bad performance it has been rejected as it is not applicable to this problem statement.

As because logistic regression is derived from a maximum likelihood as it is susceptible to the biased data.

However, the model maximizes the performance of the model over the distribution of the training set, so it would not perform as well on a test set with a different distribution.

Baseline Model :

Input = none ; Output = Operating

Performs well because data is extremely biased.

QDA(Quadratic Discriminant Analysis) Model: An ensemble-based technique that combines the results of a high precision anomaly detection algorithm (QDA) with a random forest classifier.

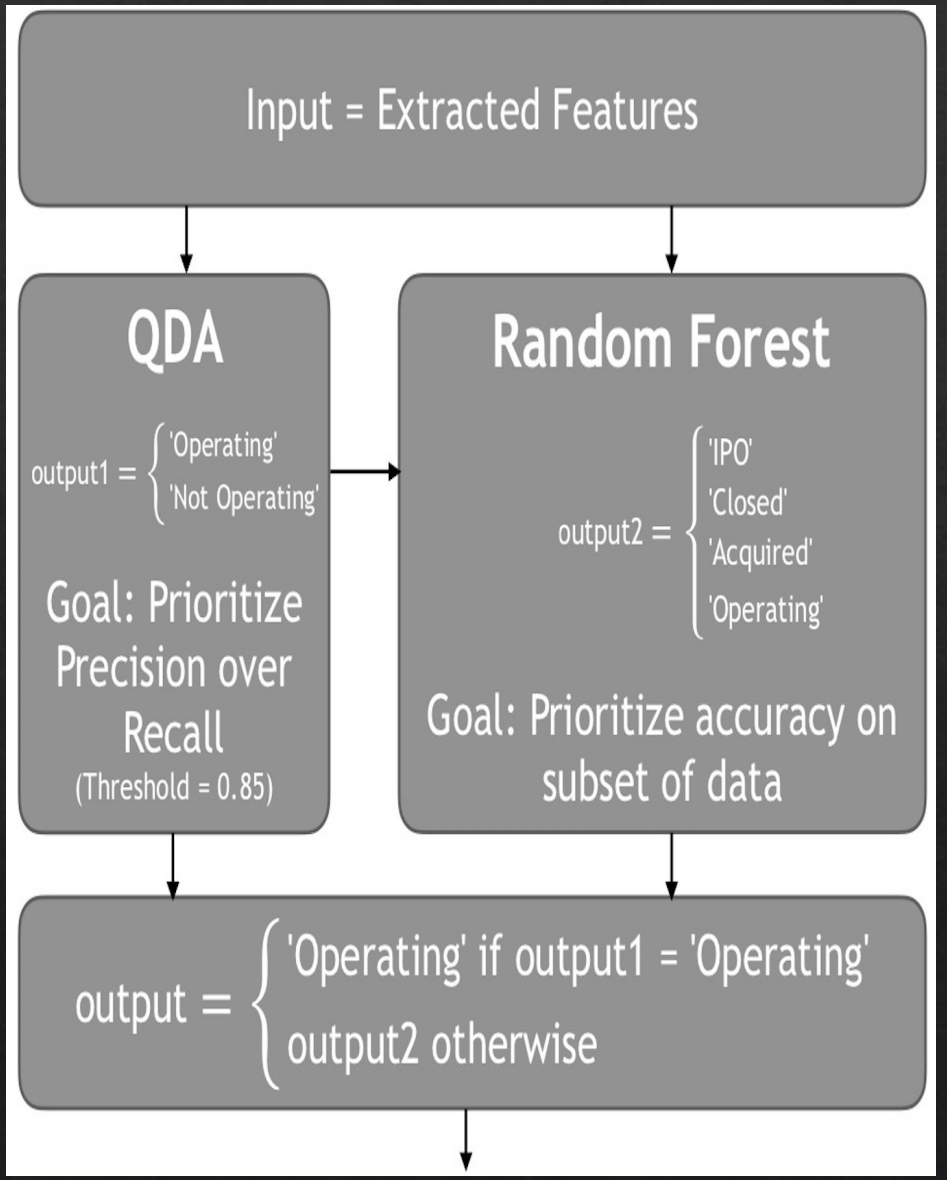
- Random Forest (RF) Classifier
- Ensemble Based Technique

Idea: Use anomaly detection techniques to first identify a subset of the majority class with high precision. The remaining subset now has lower bias.

Step 1: Use QDA, to first identify subset of Operating classes so that data can be balanced. Prioritise precision by increasing threshold.

Step 2: Use RF classifier to classify remaining subset of the data.

Features used at every step are the ones obtained from feature selection for each model individually, but more fine tuning is required for better performance.



RESEARCH PAPER - RESULTS

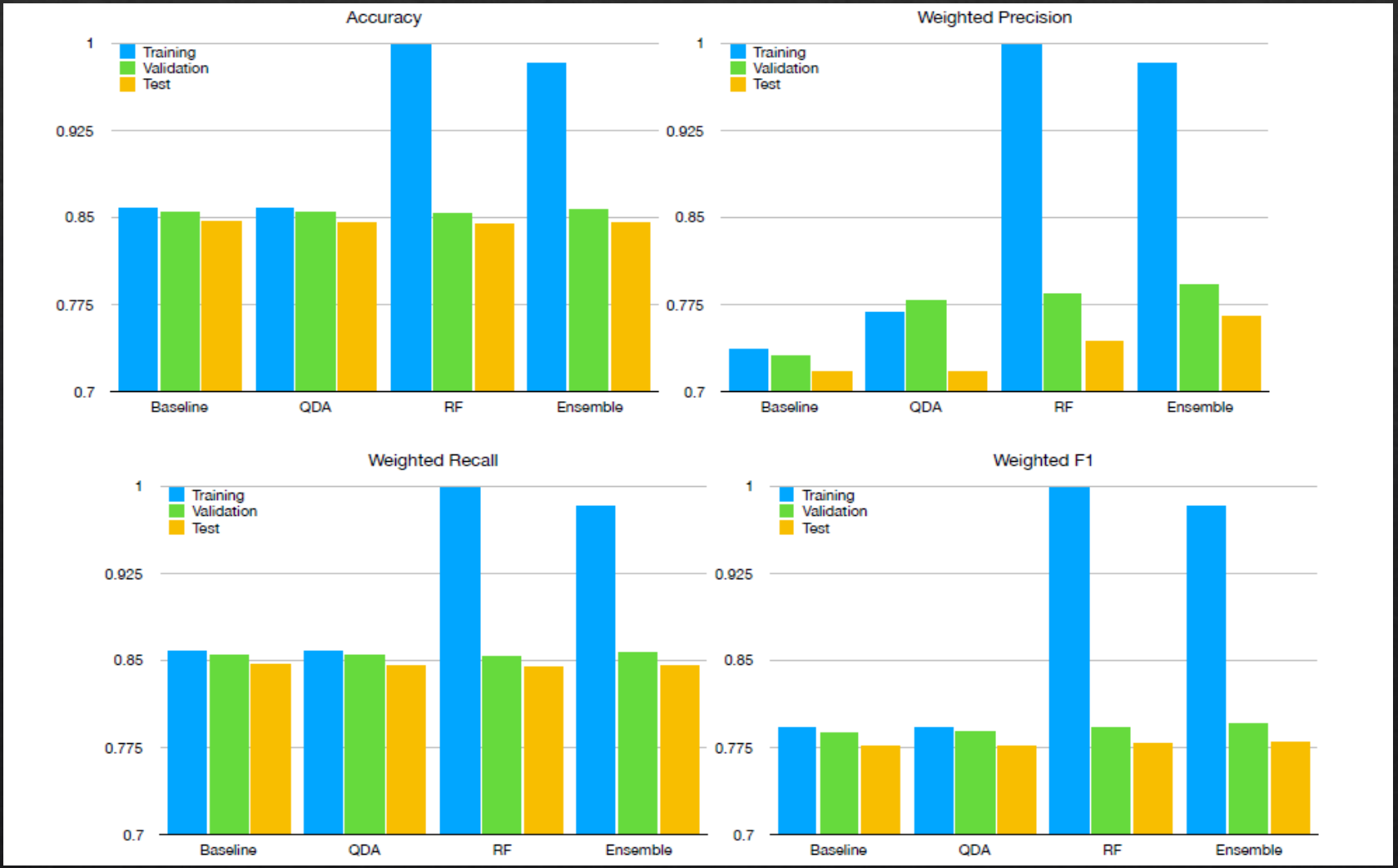


Figure_2: Showcases comparison of performance between the different models on training, validating and testing sets.

- The disproportionately high difference between training, validation and test performance on the RF model.
- This suggests that there is a possibility that the model was not correctly tuned (despite steps taken as outlined the parameter tuning section).
- There is a possibility that the two-step ensemble technique may be outperformed by a properly tuned RF model, which should be addressed in future work.

CONCLUSION & FUTURE WORK

- This two-step ensemble technique is not limited to using QDA and RF classifiers.
- We can explore how other models can be combined. Furthermore, RF models are high variance and dependent on the output of the QDA classifiers.
- While this technique seems somewhat promising when compared to using QDA and RF alone, there are many other techniques to account for biased data such as Boosting.

Thank You