Predicting a Startup’s Acquisition Status

Abdul Haseeb Sharif

Intern

Case Study: Predicting a Startup's Acquisition Status

Predicting a Startup’s Acquisition Status

- This project predicts a startup’s acquisition status based on its financial statistics
- An ensemble model approach is being opted to deal with the challenge of sampling of data.
- The model is expected to yield higher precision predictions while without compromising the accuracy and weighted recall.

Predicting a Startup’s Acquisition Status

Objective

- The goal of this project is to predict a former startup's acquisition status based on a company’s financial statistics.
- While the area of using machine learning to predict IPO under pricing has been well-researched, this topic has been surprisingly understudied.

IPO Pricing

- The results of this project may be of particular interest to investors as well as job applicants to pre-IPO companies as it can be extended to look at the likelihood of the prospective company being acquired, closed or reaching and IPO. The results of this project may also give insight to which features have the most influence on the predictions.

Algorithm

- The resulting algorithm takes in a startup’s financial statistics such as total funding dollars, funding dates, number of funding rounds, and headquarter location as inputs.
- The algorithm then predicts whether the startup has been closed, acquired, is operating, or has reached and IPO. The main challenge for this problem is dealing with an imbalanced dataset where one class is overrepresented, but under/oversampling cannot be used as a technique to balance the data.
- In order to address this, an ensemble-based technique that combines the results of a high precision anomaly detection algorithm (QDA) with a random forest classifier.

Predicting a Startup’s Acquisition Status

Initial Understandin g

- There are many ways to address biased data such as using bias-resilient models, over/under sampling the data.
- More recent research suggest that anomaly detection techniques trained for each individual class can also be promising. This paper builds off of these techniques by trying to apply an anomaly detection models in a novel way to modify a training set to be more balanced.

Details of Dataset

- The dataset is taken from Kaggle “Crunchbase 2013 Companies, Investors, etc.”
- The dataset contains 17,727 samples providing information as to the startup’s name, website, sector category, funding received, headquarter location,, funding rounds, founding date, first and last funding dates, and last milestone date.

Status

- Each row is also labeled with the company’s status (‘Acquired’, ‘Closed’, ‘IPO’, ‘Operating’)
- The dataset labels show that the dataset is extremely biased.

Predicting a Startup's Acquisition Status

Dataset Labels

IPO	Closed	Acquired	Operating
1.9%	3.1%	9.4%	85.6%

Other classes are under-represented.

Operating class is extremely over-represented

Predicting a Startup's Acquisition Status

Initial Attempt - Logistic Regression

- This is a multi-class classification problem (with only a small number of classes), so it initially seemed reasonable to apply a basic one-versus-all classification technique such as logistic regression to the problem.
- However, the resulting model performed poorly because logistic regression is susceptible to the biased data. Furthermore, balancing the data either by over or under-sampling creates a model that would not be applicable to real applications.

Ensemble Technique

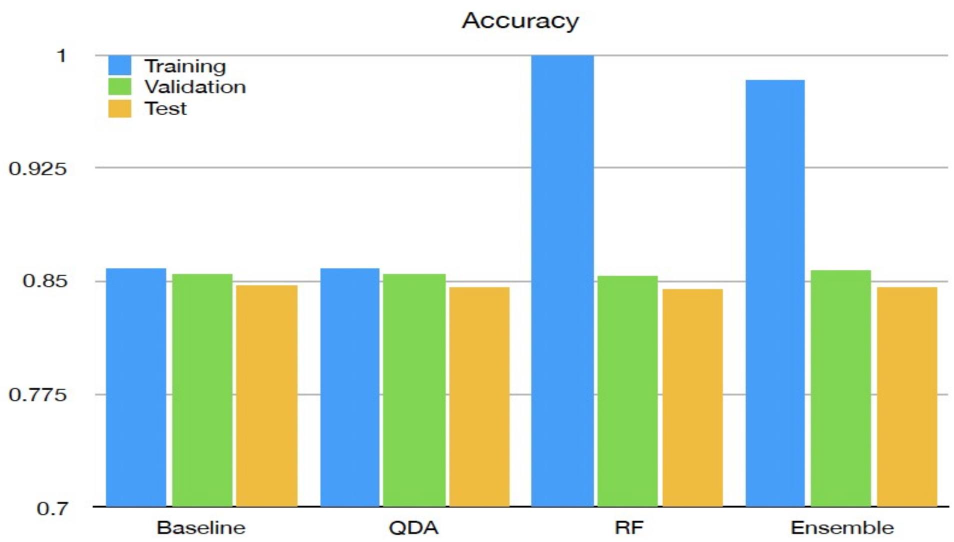
- The underlying challenge with the dataset is the over-representation of 'Operating' classes. Any model can obtain a high accuracy and recall by over-prediction 'Operating', but to the detriment of precision.
- An ensemble technique is used to attempt to address this. The general idea of the technique is to chain together two models that

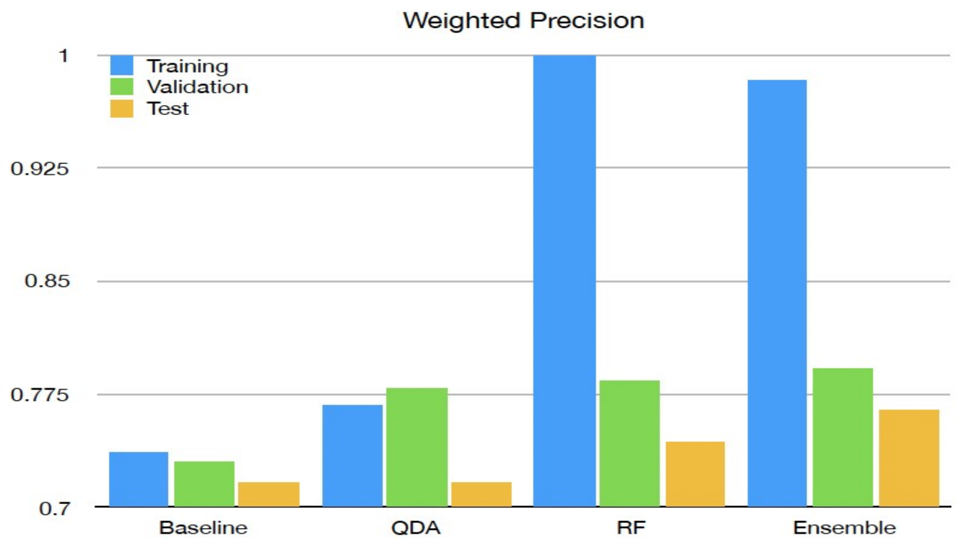
Performance Metrics

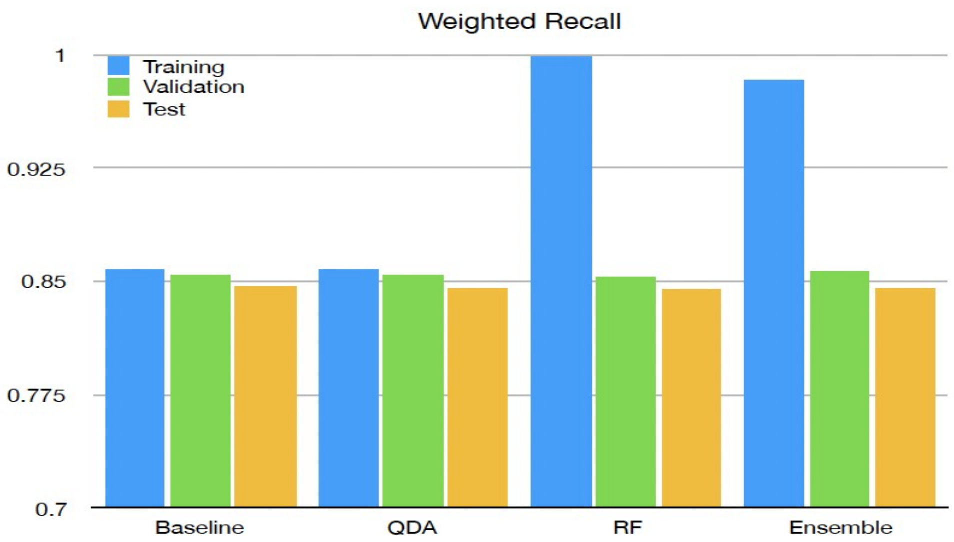
- The performance of the different models are compared by examining the accuracy of their predictions on the validation set. Since this is a classification problem with small number of classes, a reasonable definition of accuracy would be the ratio of correct classifications to total number of corrections.

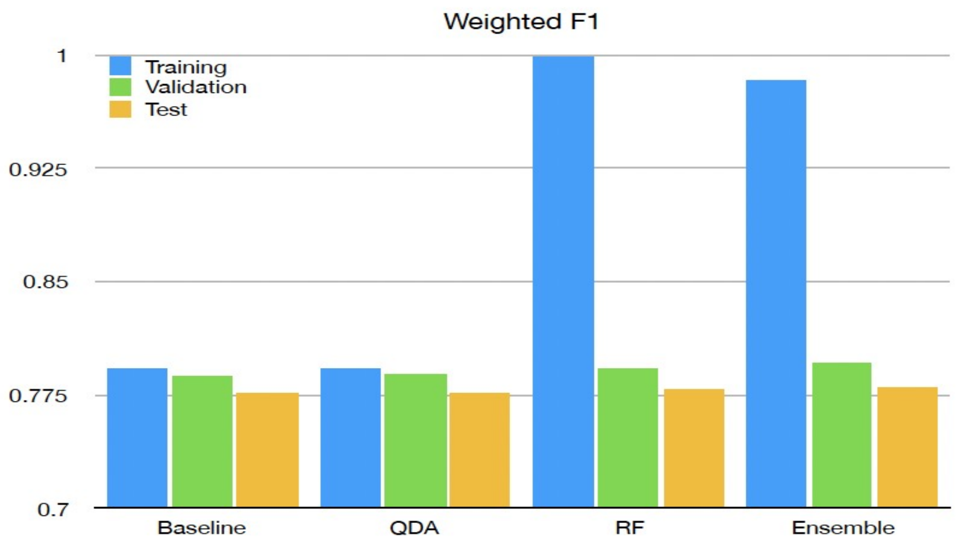
Predicting a Startup’s Acquisition Status

Results









Predicting a Startup’s Acquisition Status

Results

- The above diagram compares the training, validation, and test errors of the different models with respect to the chosen performance metrics. We see that the two-step ensemble technique which combines a high precision model with a high accuracy model gives a higher weighted precision on the test set without sacrificing accuracy or weighted recall when compared to the other models.

- While the increase in performance appears to be somewhat small, they are more significant when compared within each class.

- However, note the disporportionately 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.