Data Analysis 3 - Assignment 3 - Business Report

Predicting Firm Success

Ali Sial & Rauhan Nazir

2/10/2022

Introduction

The aim of this report is predicting and identifying the fast growth firms was to help Investment managers in identifying and cashing in on the opportunities where the returns can be maximized. We defined our target variable as a binary one, which took the value of 1 if the firm had a fast growth and 0 otherwise. We used the value of Compound Annual Growth Rate (CAGR) to classify fast-growth companies vs non-fast growth companies, where companies with a CAGR of 40% or more in sales in 2 years were put in the fast-growing category. Then we calculated predicted probabilities and with the help of loss function classified them. The models that we used were Logit, Logit LASSO, and Random Forest models with 5-fold cross validation and the criteria to evaluate the model performance was based on RMSE & AUC. The code of this entire analysis has been added to our **GitHub Repository** (please click to access the link).

The Data and Cleaning and Engineering

The data set that we are using for this exercise is related to the set of firms in a European country which was compiled by Bisnode. However, we had to do intensive cleaning and label engineering, taking conscious decisions to make sure that the data we fed into the models was in the desired form, and the quality and accuracy of the models was not compromised in any way (Garbage in Garbage out).

The initial data set had data from 2005 to 2016, but we narrowed it down to 2010 till 2015. One of the most important predictors for predicting the growth of the company is the sales or revenue they are generating and how the sales are changing year on year, so we decided to include the variables that portrays this information. To make sure that the actual growth trend was being reflected we decided to compute CAGR based on 2 years data rather than 1, so that the temporary fluctuations in the growth can be accounted for, and we created a dummy variable for this. Another decision we took was to include firms that had a revenue greater than 1000 euros and less than 10M euros, as values outside of that range are rare and could have been due to some human error, directing our focus primarily on small and mid-sized companies. As far as feature engineering is concerned, we started by looking at the key variables such as sales, age of the firm, total assets, profit and loss, etc. We started by fixing the distribution such as adding logs to normalize it or quadratic terms to capture maximum distribution. We also added factors, classifications or groups wherever it was possible and required. Similarly, we imputed data for the missing values using an appropriate approach such mean values or setting a threshold for categorization of imputation. Flag variables were also added for all the variables that were imputed.

Prediction and Modeling

As mentioned above we used the division of variables based on the groups to develop out prediction models. In total we had 5 logit models, LASSO and a Random Forest with tuning parameter. To begin running the models, we first divided the datasets in two subsets i.e., training data (80%) and holdout data (20%). This division was consistent for all the 3 different dataset we are evaluating. By that I mean the entire data with all the firms, data with only manufacturing firms and data with only services firms. After the division the train data was used for 5-fold cross-validations for each model. The models and results are explained ahead.

Probability Logit models and Logit LASSO

As stated above, we created 5 logit probability models. To measure the performance of models and select our best model, we used Area Under the Curve (AUC) and the average RMSE for the 5-fold cross validation. The tables provided below shows the 5-fold cross-validated RMSE for the logit models, individually for 3 datasets we are working with. Interpreting the results in the tables, it appears that for the main dataset, model 4 outperformed others having lowest avg. 5-fold cross-validated RMSE and largest AUC. For manufacturing and services, the best logit model for both was model 3. For the second type of prediction algorithm, we decided to develop a logit LASSO model with highest number of predictors. The LASSO's RMSE was higher, and AUC was less compered to selected logit models for all the datasets.

Probability Forest

The last prediction algorithm we used in our analysis is Random Forest with tuning parameters. Even though it is known as a black box model, since it builds a stronger model based on the way it selects the predictors, it is better at identifying non-linear relationships and interactions. The predictors we used RF were similar to that in logit model 4, but without any feature engineering. As we expected for all the datasets, the probability forest returned the lowest 5-fold cross-validated RMSE and the highest AUC. This further validates our claim of RF being the best approach for prediction probabilities as far the data we are using is concerned.

ROC Curve

Upon completing the analysis for all the prediction models and the diagnostics, we selected RF model with tuning as best model, and we will be using this to further evaluate our predictions. Using the selected model (RF), we first plotted a Receiver Operating Characteristic (ROC) curves for the models across the datasets using discrete thresholds between 0.05 and 0.75. The ROC curve provided below is for the RF model that ran on the main dataset. Similarly, we also looked at the ROC curves for other two datasets, which also had decreasing slow but remained about the 45-degree line. This indicates that our RF model predictions tend to be better compared to other models with used in this analysis.Below you can find the ROC curve for the main dataset.

Loss Function, Optimal Threshold & Classification

Loss function is primarily based on two assumptions. First, we looked at the risk-free interest rate an individual receives if they deposit money into an Hungarian Bank. The risk-free rate in the market is about 3.3%. Secondly, we assumed that the if the same amount is invested in a fast-growing company the return would be risk-free rate plus the investment premium, which in our case is around 10%. Additionally, we also assumed that in case the investment is made into a company that was predicted to be a fast-growing and in actuality it wasn't fast growing the return for the investor would be 0%.

Next, we calculated the optimal threshold individually for each dataset using the loss function. With regards to the main dataset without industry filtration the calculated optimal threshold is 0.32. Similarly, for the other datasets which manufacturing and services the optimal threshold are calculated to be 0.32 and 0.24, respectively. Therefore, based on these optimal thresholds the predicted probability of a firm that to be classified as fast growing one, would be 0.32 and above. The same rule would apply to the other two datasets as well. The graphs below are for the associated optimal threshold and AUC for the defined loss function.

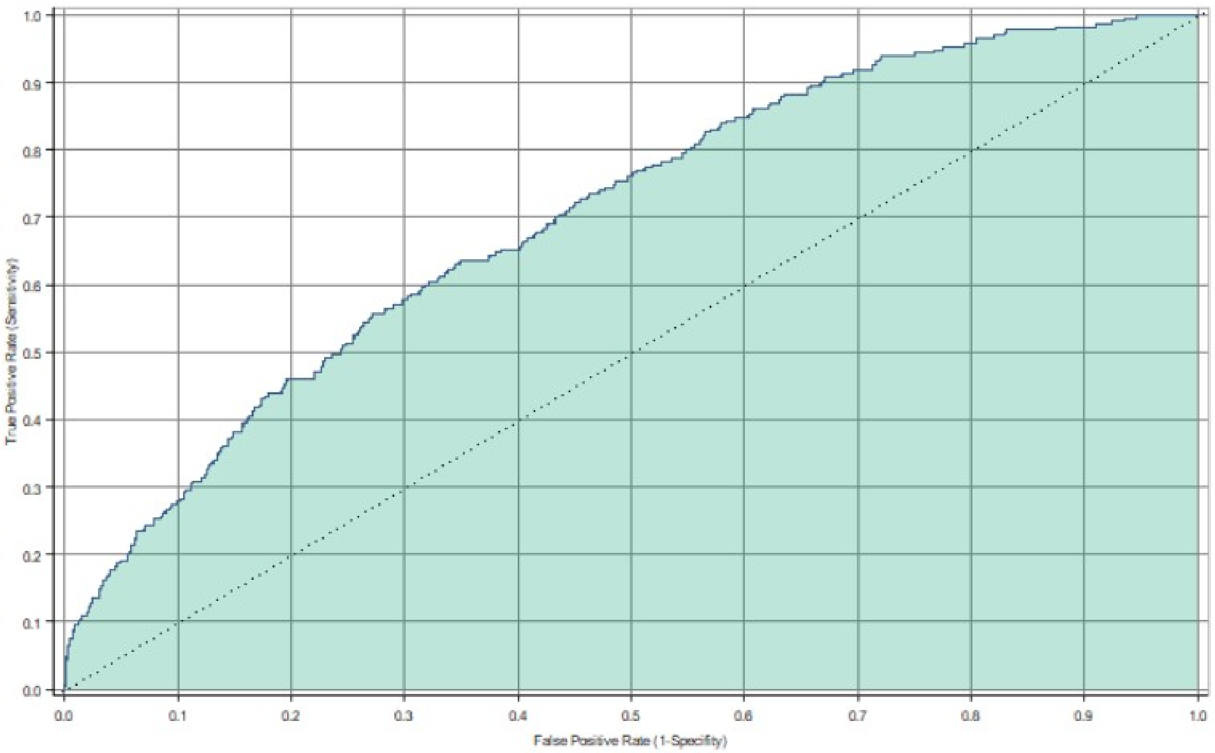
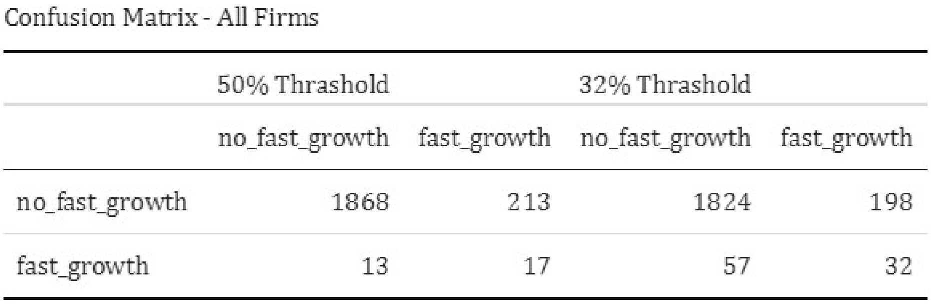


Figure 1: ROC Curve

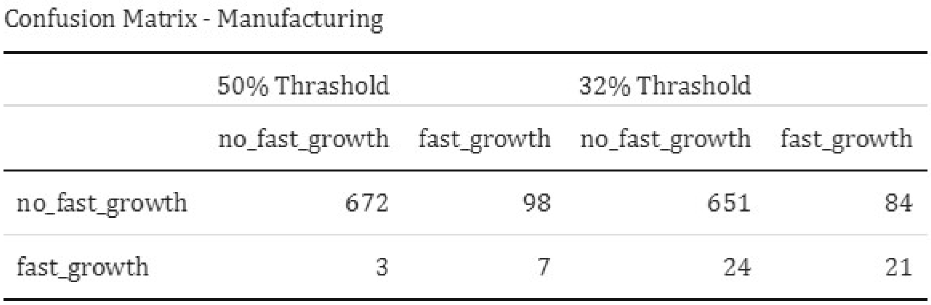
Confusion Matrices

We created 2 types of confusion matrices, one with the no loss function and one with the optimal threshold which was calculated through a function. This was also done on the 2 subsets of the data, Manufacturing and Services. The threshold level allows us to have a degree of flexibility in terms of increasing or decreasing the number of observations assigned to false positives and false negatives. Both of which are components of a loss function. As, in our case False Negative was penalized more heavily, it was intuitive that our optimal threshold was less than that of the default threshold of 0.5. It was predicted that if the company decided to use the optimal threshold for the entire industries, they would incur an additional loss, however when run separately on the Manufacturing and Services industry, the company would be saving themselves from losses, compared to using the one with no loss function.



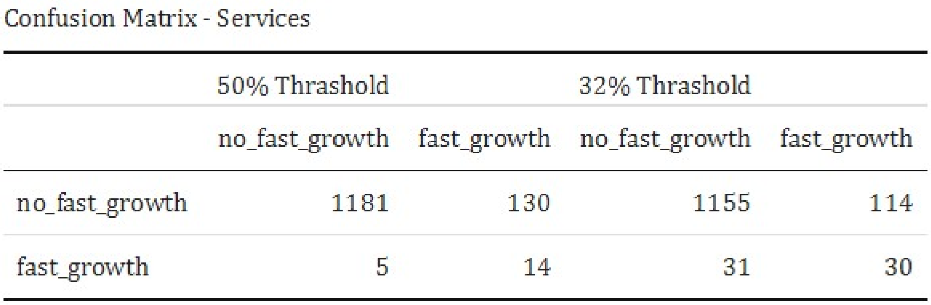
Confusion Matrix - All Firms

	50% Thrashold		32% Thrashold	
	no_fast_growth	fast_growth	no_fast_growth	fast_growth
no_fast_growth	1868	213	1824	198
fast_growth	13	17	57	32



Confusion Matrix - Manufacturing

	50% Thrashold		32% Thrashold	
	no_fast_growth	fast_growth	no_fast_growth	fast_growth
no_fast_growth	672	98	651	84
fast_growth	3	7	24	21



Confusion Matrix - Services

	50% Thrashold		32% Thrashold	
	no_fast_growth	fast_growth	no_fast_growth	fast_growth
no_fast_growth	1181	130	1155	114
fast_growth	5	14	31	30

Calibration Curve

Calibration curve is one way of seeing whether our model is well calibrated in terms of predicting the probabilities, and in turn is biased or not. The closer the curve is to the 45-degree line, the less biased the model is. In our case we could see that after a certain level of predicted probability, the curve starts to move away from the 45-degree line, meaning that our prediction is biased to some extent. However,

for It to work properly and accurately the sample size needs to be large. In our case this is not the case. So, if provided with the opportunity it will be nice to see what the results come out to be when we have a bigger sample size. Below are the Calibrated curves for the entire data set and the two subsets (Manufacturing and Services)

Conclusion

After evaluating all the prediction models on various metrics including the RMSE and AUC, we decided that the best model for our case was Random Forest across all of the defined datasets. Moreover, we uncovered that the model performance improved when the data was trained on specific industries separately rather than including all of them together. This is quite intuitive as well, as it is better that our model is used for predicting the performance of firms belonging to a specific industry, rather than trying to predict performance of all kinds of firms using the same model (Jack of all trades, master of none).

Moreover, there are certain improvements that could have been made to increase the quality of models, for instance getting more data and increasing the sample size, which would allow the models to be trained better and help the investment company make better decisions. We could also run models on different time periods than the one we filtered our data down to, which would allow us to check for external validity.