Predictive Models for SMEs

Aftab Alam & Khawaja Hassan

2/17/2022

Introduction

Investment managers need to allocate funds to opportunities where the return can be maximized. As consultants to an investment firm, the aim of our model was to predict probabilities of companies that would have a Compound Annual Growth Rate (CAGR) in sales of 40% or more between 2012 and 2014 and then classify the companies into two classes. This value of CAGR was used to classify fast-growth companies vs non-fast growth companies. Based on these classifications, the investment managers would then decide on whether to invest in the company or not. The value of this CAGR is driven since if a company has sales of 1 million in the year 2012, then what CAGR would result in a sale of around 1.5 million in the year 2014.The prediction analysis utilized the Logit, Logit LASSO, and Random Forest models with 5-fold cross validation. To arrive at the best model, we looked at RMSE, AUC, and the average expected loss as per our defined loss function. These models used several company features, such as balance sheet items, income statement items, and management related variables.

Feature Engineering

The data was prepared by Bisnode; however, it has been sourced from the OSF website. The data followed a time frame of 2005 till 2016, with more than 280,000 observation and 48 explanatory variables. In accordance with the importance of designated variables, we took different methods for dealing with each explanatory variable. For the classification of firms as fast growth we took a CAGR of 40% and calculated it based on 2012-2014 to account for stable growth. Moreover, to account for categorical effect we created dummies for certain variables such as firms' status to check if they are still operational. Further, we created normalized values for all income statement & balance sheet variables to ensure better comparison across medium and small-scale companies. Along with that we created flag variables for financial columns were the value were either missing or shown as negative. Later, columns with flag variable were either imputed with mean value or replaced with zeros. A new level variable was then created to classify fast growth companies and others where the value of 1 was assigned to fast-growth companies. After making all these modifications, we were then left with 10558 observations with 115 variables in total. On this clean data we ran our prediction models and then divided our data based on Manufacturing and Service industries to evaluate prediction performance in each sub-set.

With the help of featured engineering, we were able to store our variables into group based on their designated importance. These helped us in creating models with different levels of complexity.

Prediction & Modelling:

The datasets were divided into two subsets: training data (80%) and holdout data (20%). The training datasets were then used for 5-fold cross-validations in the models used ahead. In total, 5 probability logit models each with increasing complexity, 1 LASSO model, and 1 Random Forest model was run with provided tunning parameters.

Based on average 5-fold cross-validated RMSE for the *5 logit models* across the three datasets the results are as follows. For combined dataset, the lowest RMSE resulting from average 5-fold cross-validated was 0.30058. Although, the AUC was slightly higher in model 4 but we still went for model 2 considering the insignificant difference in RMSE and higher model complexity in the later model. With the same criteria, for manufacturing and services the model 2 & 3 proved to be a better fix with average 5-fold cross-validated RMSE of 0.3273 & 0.2829 respectively. All the model selection was based on lowest RMSE, AUC and model complexity.

For the second type of model, we ran a *logit LASSO model* with most complexity, we added the greatest number of variables in it. It included interactions, dummy variables, HR related variables, management related variables, and other important variables in the dataset. For combined dataset, average 5-fold cross-validated RMSE for this model was around 0.2996, which is lower than the best logit model mentioned above. However, the AUC for LASSO was much lower compared to the simple logit models. On the other hand, the results of LASSO for manufacturing and service predicted higher average RMSE as comparison to the previous preferred models. However, the LASSO average 5-fold cross validated AUC for both sector we either lower or almost equal to the selected logit models.

To build a stronger prediction model, we decided to run a *Random Forest (RF)/Probability Forest (PF)* on the dataset. Although it is a black box, it is better at identifying non-linear relationships and interactions. Hence, we went ahead and used the predictors used in the model 4, but without any feature engineering.As expected, the probability forest returned the lowest 5-fold cross-validated RMSE of around 0.2960 among all the models used so far and it also returned the highest AUC of around 0.7465, suggesting that this model being superior to all others when considered only under these two parameters. This was the case for the RF models in both manufacturing and services datasets; it was superior in terms of both 5-fold cross-validated RMSE and AUC.

Loss Function

We defined our loss function keeping in mind two things, the risk-free interest rate paid by depositing the money in a bank and the rate of return on investing money in a company. We looked at the current interest rate provided by Hungarian banks on deposits as the risk-free rate, 3.3%. Whereas we assumed the rate of return on investment in a fast-growing company would be 10%. Another assumption while creating the loss function was that if an investment manager invests in a company that turns out to be non-fast-growing company, the investment manager would get 0% return out of the investment.Based on these assumptions, we calculated the opportunity costs to arrive at the relative losses by false negatives and false positives. If the manager invests in a company and the classification was false positive, then the manager would lose out on the 3.3% return that could have been earned from depositing the money in a bank. Hence the cost of a false positive is 3.3% risk free return.

On the other hand, if the investment manager does not invest in the company based on a false negative classification, the loss would be 6.7% as the money would be deposited in the bank and the money will still earn 3.3% (10% - 3.3%). Based on this, the ratio of cost of FP and FN would be 1:2, where false negative being twice as costly as the false positive cost.

Optimal Threshold & Classification

The optimal classification threshold based on these relative costs was calculated as 0.33. This was calculated using the optimal classification threshold formula that assumes that the model being used is the best one for prediction, which may not be true in practicality.

Hence, we carried out the exercise of calculating the optimal threshold using the data itself while incorporating our loss function. We plotted ROC curves to find the optimum threshold. For the main dataset, optimum threshold was 0.35, whereas it 0.27 for the manufacturing dataset and 0.25 for the services dataset.

Based on these classifications, in the main dataset, any predicted probability of 0.35 or above would be classified as fast-growth, 0.27 or above as fast-growth in the manufacturing dataset, and 0.25 or above as fast-growth in the services dataset.

Confusion Matrices

The first confusion matrix was built without the loss function, which runs on the majority vote ideology, where it assigns the value of 1 to any predicted probability of 0.5 or above. This is not the optimum threshold as the losses from false positive and false negative are not always symmetric in the real world.

Given that false negatives are more costly in our case, the goal would be to reduce the occurrence of false negatives in our predictions. However, if we first look at the 0.5 threshold matrix, the percentage of false negatives is around 10.1% and percentage of false positives is 0.56%, whereas, with a 0.35 threshold, the percentage of false negatives is 9.1% and percentage of false positives is 3.12%.

Based on our loss function, the main model suggests that the company loses out around 1,176 Euros per firm and if the company evaluates 1000 firms in a year, the company loses out around 1.176 million Euros. However, when focusing on the confusion matrices for individual manufacturing dataset and services dataset, the results are starkly different.

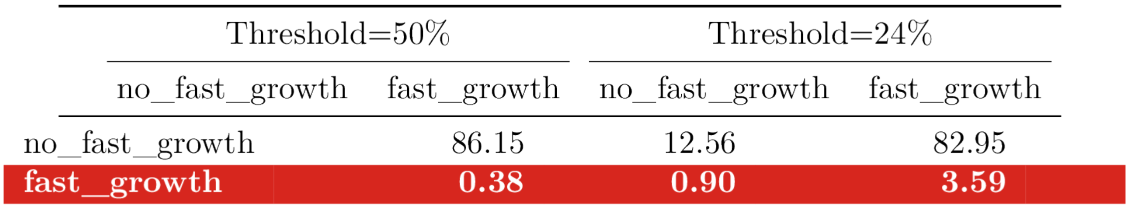
All the values in the confusion matrices below are in percentages

Table 1: Confusion Table-Combined Dataset

	Threshold=50%		Threshold=24%	
	no_fast_growth	fast_growth	no_fast_growth	fast_growth
no_fast_growth	**88.54**	**10.09**	**85.98**	**9.1**
fast_growth	0.57	0.81	3.13	1.8

The manufacturing dataset predictions suggest that using our predictions saves the company around 590 Euros compared to the 0.5 threshold and if the company evaluates 1000 firms in a year, then the company saves 589,743 Euros in total.

Table 1: Confusion Table Manufacturing



	Threshold=50%		Threshold=24%	
	no_fast_growth	fast_growth	no_fast_growth	fast_growth
no_fast_growth		86.15	12.56	82.95
fast_growth		**0.38**	**0.90**	**3.59**

Similarly, services dataset prediction classification suggests that our investment firm will save 1090 Euros per firm and if the company evaluates 1000 firms per year, the company will save around 1.1 million Euros per year.

Table 1: Confusion Table   Service

	Threshold=50%		Threshold=24%	
	no_fast_growth	fast_growth	no_fast_growth	fast_growth
no_fast_growth	88.80	9.70	87.14	8.72
fast_growth	**0.38**	**1.13**	**2.03**	**2.11**

Conclusion

Based on our prediction models, the best model has been the random forest model across all the defined datasets. Although a black box, random forest is result oriented and is optimum for our investment management company, which is also result oriented.

Moreover, our analysis suggests that training the best model on specific industry dataset is better when making classification models compared to training the model on one larger dataset that contains multiple industries. This makes sense in the real world as well where models are trained on same types of companies to make better predictions.However, to check for external validity, we highly recommend running these models on a different time periods than 2012-2014 time, perhaps checking multiple periods would be wise, such 2013-2015, 2014-2016 etc. Additional to this, we would suggest collecting more observations on industry specific small and medium sized firms separately and running this prediction exercise. Perhaps that would give focused results, helping the investment company further.

Link to RMD codes - GITHUB