Technical Report: Finding Fast Growing Firms

Shahana Ayobi

2023-02-26

Introduction

The purpose of this project is to build a model that can predict the probability of fast growing companies vs non-fast ones. To classify those companies in the mentioned categories, a loss function is required that will quantify the the results of the modeling decisions. For this purpose, a threshold of 25% or more of Compound Annual Growth Rate (CAGR) in sales is chosen. Which then classifies the companies into two categories: fast-growth companies with CAGR of 25% or more growth and non-fast-growth companies otherwise. The Logit, Logit LASSO, and Random Forest models with 5-fold cross validation were used in the prediction analysis. To determine the best model, RMSE, AUC, and average expected loss as defined by the loss function were considered. Several company features were used in these models, including balance sheet items, profit and loss statements, and management.

#Label and Featutre Engineering The data set contains detailed company specifications from 2005 to 2016 and includes 287,829 observations and 48 variables. Data was prepared by Bisnode, but was obtained from the OSF website. In order to find a model to predict the probability of fast growing companies, the CAGR treshold was classified for companies with 25% or more sales growth over a two-year period. Companies with a CAGR of 25% or higher are classified as fast-growing, while others are classified as slow-growing. Furthermore, the data is filtered into a panel data of these companies from 2012 to 2014. Then the data was filtered on the companies that had data for these two years and CAGR was calculated. The reason for selecting a two-year time span rather than a single year was to ensure consistent sales growth, as sales numbers can fluctuate year to year. Furthermore, the reason for not selecting a longer time frame than 2012-2014 was a limitation in prediction power, as with longer time frames, prediction exercise becomes increasingly difficult. In addition, a dummy variable was created to distinguish between businesses that were still operating and those whose sales were either NA or greater than 0. Based on this criterion, we only kept the companies that were regarded as being alive. Additionally, the data was further restricted to only include small and medium sized businesses. This assumption was made based on sales of between 1000 and 10 million euros. Lastly, observations with a CAGR of less than 250 were used with data from 2012 to make predictions. Looking at sales distribution, it is skewed with longer right tail, thus log variable was created to make the sales close to normal distribution. The industry categories were combined combined together to decrease the number of categories of of the variable, and several service sand manufacturing industries were grouped together. By deducting the founding year from the present year of observation, the age of the company was then determined. For other variables with skewed distribution winsorization method was used to identify a threshold value for each variable and then replace values of variable that lie outside the threshold with the threshold value and add flag variables. As a result of the distribution of the main financial variables, we created some new variables. Because different types of assets are expected to be positive, we added a flag asset to identify assets that are greater than zero. Furthermore, for less than negative values, we assigned zero to all intangible, current, and fixed assets. Furthermore, we added new columns for all profit and loss variables and scaled them by dividing the variables by sales, as well as new balance sheet ratio variables by dividing the variables by total assets which was created by adding together the intangible, current, and fixed assets. Furthermore, because these ratios vary depending on the nature of the firm, we winorized them and kept the ratios between -1 and 1. Furthermore, we identified counting variables that could not be less than zero and created a flag variable called flag error to identify such values. In addition,

we created a balance sheet variable that totals all assets. We also added some variables in the square and quadratic terms to capture non-linearity.

As the above graph shows that the distribution of sales is skewed to the right and log of sales has a close to normal distribution.

Modeling

The original cleaned data is split into two random parts by 20% to 80% ratio in order to avoid over-fitting. The Holdout set includes 20% and the rest 80% is work data set. Then, the work data set is split into train and test data sets and 5-fold cross validation is run on the train datas et. Then the best model is chosen based on the lowest average of 5 CV RMSE result.

Probablity Logit Model

The logit probability prediction was performed first by selecting the best logit model through cross-validation, and evaluating the model using the holdout set. To find a better model to use for further analysis, five different logit models were considered, ranked from simplest to most complex. The M1 logit model includes variables based on domain knowledge, and we included variables that we thought were important. The results of RMSE and AUC for the five logit models show that the RMSE results have very small differences. Thus, the Model X4 has the lowest RMSE of 0.3965 and the highest AUC of 0.7077. Thus model 4 is chosen for further analysis.

Logit LASSO

LASSO is an algorithm that fits a model by shrinking coefficients, some of which are reduced to zero by the addition of a penalty term. LASSO for logit is run in this set of models and all of the variables and interactions are included in the Logit Model 5. As a result, LASSO generates a model that includes the majority of the variables but excludes some. Based on the results, it is concluded that LASSO performs well in terms of RMSE with 0.3962, whereas the fourth simple logit model performs better in terms of AUC.

Random Forest

Because Random Forest can assign functional forms and interactions, the variables in the random forest do not have any patterns. In each split, the random forest tuning parameter was set to 5, 6, or 7. the result shows that Random Forest outperforms other models like LASSO and probability logit with RMSE of 0.3946, and the AUC value of 0.7108 . As a result, the Random Forest model with the lowest RMSE and highest AUC is the best performer. As a result, an ROC curve is drawn using our holdout set.

Model Evaluation and Confusion Matrix

As a result, when compared to Logit LASSO, Logit Model 4, and Logit Model 1, Random Forest has the lowest expected loss and RMSE. However, the logit Model performs similarly. The expected loss from Model 4 is nearly identical to that of Random Forest, with only a 0.002 difference. The RMSE differs by 0.0019, which is quite negligible. Thus, choosing Model 4 makes more sense since it can be easily interpreted compared to Random Forest. After creating confusion matrix for model 4, it is apparent that the accuracy of the model is 76% meaning it correctly predicted 76% of the firms. Model specificity is 87% which demonstrates

that from not fast growing firms the model correctly estimated 87%. Eventually, model sensitivity is 38% meaning that model predicted the fast growing firms by 38%.

	Number.of.predictors	CV.RMSE	CV.AUC	CV.threshold	CV.expected.Loss
Logit X1	11	0.4100453	0.6450208	0.3343912	0.4071715
Logit X4	78	0.3964892	0.7077788	0.3362984	0.3735718
Logit LASSO	122	0.3962475	0.6887882	0.3433097	0.3899970
RF probability	36	0.3945716	0.7108368	0.3495282	0.3710488

	no_fast_growth	fast_growth
no_fast_growth	1928	408
fast_growth	286	250

$$accuracy = \frac{TP + TN}{N} = \frac{250 + 1928}{2872} = 76\%$$

$$sensitivity = \frac{TP}{TP + FN} = \frac{250}{250 + 408} = 38\%$$

$$specifcity = \frac{TN}{TN + FP} = \frac{1928}{1928 + 286} = 87\%$$

Conclusion

The final model chosen for this case study was Model 4, which includes the variable of firm details as well as all engine variables such as firm financial information such as balance sheet and profit and loss, as well as HR variables. The The model's accuracy is 76%. The model correctly estimated 87% of firms that were not rapidly growing. It eventually predicted 38% of fast-growing firms. However, running these models over several time periods is encouraged, and it might be a good idea to evaluate many time periods, in order to verify external validity. It is also advised to run the model separately on small and medium-sized businesses that are focused on a certain industry. Perhaps this would produce findings that are more focused, enabling the investing companies to make informed decisions.