Business_Report

Ghazal Ayobi and Shah Ali Gardezi

2/17/2022

Introdction

This study focuses on the firms from 2010 to 2015 mainly the firms had high growth rate from 2012 to 2014. In order to classify firms as fast growing and not fast growing we minimized the average expected loss and applied the appropriate threshold to assist investment decisions. Moreover, the main aim of this case study is to build a prediction model which can support individuals in their investment decisions in choosing between fast and non-fast growing firms. To classify firms in the mentioned categories, we need a loss function which quantifies the consequences of the decisions that are driven by the prediction (Gabors, 2021). The loss function has two values, one is a loss due to the false negative and a loss due to the false positive. For this purpose we considered these features of the companies and build 7 different models which are OLS, LASSSO, Random Forest and OLS Logit. The data comes form the Bisnode, a company that offers decision support in forms of digital business, marketing and credit information. The data set contains detailed description of the companies from 2005-2016 for some industries in manufacturing and services. In this study we focus on the cross section of the companies in 2012, we will ask weather they are fast growing or not in the subsequent years.

Label Engineering

Before start of modeling it is vital to define our y variable and start with feature engineering. Based on the Business question we would to build a model to predict fast and non-fast growing firms. Thus, it is important to define what is considered as fast growing firm. For this purpose we consider CAGR, To start with label engineering we define y variable which is whether a company is a fast growing or non-fast growing. Thus, we use compound annual growth rate (CAGR) to be 28% or more. The reason is that on average small or mid-size firms in their initial years have higher annual growth rate than the large companies. Thus, in order to consider a small or mid-size firm as fast growing, we expect it to have CAGR of 28% or more across two years. Thus we define fast growing firms if their CAGR sales value is 28% or more for this purpose we are focusing on the mid and small size firms we only kept the sales between 10 million and 1000 euros.

Sample Design

the Bisnode data set contains 287829 observations and 48 variables. we only kept observation from 2010 to The second task of the project is deciding which firms to keep in the data. As our main focus is on the small and mid-size enterprises captured by 28% of their CAGR sales. To narrow our focus we only kept the companies which had sales between 10 million and 1000 euros in 2012. As a result of the sample design we ended up with 10462 observations and 117 variables. The main goal of the sample design is to reduce impact of extreme values. Moreover, we added anther filter to make sure than companies are still in the market indicating that they are still alive, we filtered the data for the alive status firms.

Feature Engineering

The third task in the case study is feature engineering, which consists of selecting x variables, cleaning them and putting them in appropriate forms for the prediction model. The variables have different characteristics such about the firm which are the firm size, financial factors, human resource and others. The main part of feature engineering is to decide what functional forms of variables should be included. The above figure shows that most of the variables from the data set have skewed distribution. there are a wide range of potential extreme values which also contains errors and unusual values. It is vital to check the distribution of main variables before assigning a functional form. There are various approached to address such values. Such as: grouping the factor variables, transforming the function to its logarithmic form and another method used by the Gabor, 2021 is winsorization. which states that it is a process where for each variable we identify a threshold value and replace values outside that threshold with threshold values itself and adding a flag variables. Thus, we created some new variables based on the result of the distribution of the main financial variables. For example, different types of assets are expected to be positive, thus, we added flag asset to identify assets more than zero. Moreover, we assigned zero for all the assets intangible, current and fixed, for less than negative values. Moreover, we created new columns for all the profit and loss variables and scaled them by dividing the variables by sales and we created another balance sheet ratio variables by dividing the variables on total assets. Moreover, as these ratios contains variation thus, based on the firms nature and we winsorized them and keep the ratios between -1 and 1. Moreover, we also identify counting variables which cannot be less than zero and created a flag variable as *flag_error* to identify such values. In addition, we created a balance sheet variable which sums all of the total assets. To capture non-linearity we also added some variables in the square and quadratic terms. Other flag variables were created to identify the age of CEU of the firm and other variables such as type of industry were transformed to the factors and as result missing variables were addressed in multiple ways such as imputation, dropping. as a result of data cleaning, munging, imputation, and removing null variables, we have 10462 observations and 117 variables.

Modeling

Based on the defined business question the main aim of this project is to build a prediction model to identify fast and non-fast growing firms. For this purpose we use compound annual growth rate with a threshold of 28% over the period of two years 2012 to 2014. Based on the business question we identify the firms as fast growing which performs more than the above growth rate. The table below shows that based on define criteria 16.5% of the firms have fast growth rate. The reason behind considering two years interval from 2012 to 2014. it is difficult to identify firms as fast and not fast growing based on the coverage of one year. Two years gives a more detailed and cumulative growth rate for the considered firms. As the firms grow they maintain their growth from the first year to the second, thus, two year sales CAGR is taken to identify fast growing firms. As a result in the data set the non fast growing firms are 8736 which is 83.5% and the fast growing firms are 1726 which is 16.5% of the data set.

Set up

The best model gives the best prediction in the live data. Before turning to the modeling part of the project, it is worth mentioning that in order to avoid over fitting, the original data is split into two random parts by 20% to 80% ratio. Holdout set contains the 20% and the rest is work data set. For the training set we use 5-fold cross validation, this means splitting the train data into five random samples and calculating and deciding based on the average of 5 CV RMSE result.

Probability Logit Models

For this case study we performed probability prediction by logit and as result selecting the best logit model by cross-validation and evaluate the Model by the holdout, manufacturing and service data sets. Five different logit models are considered ranked from simplest to the most complex one were used to find a better model to use for further analysis. The M1 logit model includes variables based on the domain knowledge, we included variable which we considered to be important. To compare the models we require a standardized measure to select the best model among the created models. Two vital measurement for this purpose is Root mean squared error (RMSE) and area under the Curve (AUC) which allows us to select he best model. The below table shows the result from RMSE and AUCA for the five logit models. The below table is the result of RMSE and AUC for all of the five logit Models. it can be seen from the result that RMSE result have very small differences. Thus, RMSE is the lowest for the Model X3 and X4 same as AUC. For this case study, we consider X3 as it has the lowest RMSE compared to X4 by 0.00028 and it has a simple form compared to X4. For further analysis we will consider model 3.

LASSO MODEL

The set of models in this section includes LASSO for logit. Our Logit Model 5, includes all our variables and interactions. In this section we use this group of variables in the LASSO algorithm to select variables meaning shrink the coefficientsin order to have a better predictive model. As a result LASSO produces a model that includes most of the variables, however it also drop some variables. Based on the below result we can say that LASSO performs well while in RMSE, on the other hand the third simple logit model has better result for AUC.

Random Forest

For this case study the last model is (probability) Random Forest using the set of variables for the random forest from the model section. These variables are raw variables, human resource, and firm variables. Moreover, the variables in the random forest does not have any patterns because Random forest can assign functional forms and interactions. We set the random forest tuning parameter to be either 5, 6, or 7 in each split. The result from the Random Forest indicates that (probability) Random Forest outperforms other models such as LASSO and probability logit by around -0.001 RMSE and the AUC result is higher than the both mentioned model 3 and LASSO logit. As a result the Random Forest based on the RMSE and AUC result is the best performing model. Thus, we use our holdout set to draw ROC curve.

Best Model based on Expected Loss

As a result of we can say that Random Forest has the lowest expected loss and RMSE, compared to Logit LASSO, logit Model 3 and logit Model 1. However, logit Model has the similar performance. The expected loss from Model 3 is same as Random forest which is around 0.30764. The RMSE is different by 0.0006 which is a very small number. The reason behind this decision is logit Model 3 is easily interpretative compared to the Random Forest which is black box model.

Model Evaluation and Confusion Matrix

As we have chosen our best model, we can evaluate the result. As a result we can say that we correctly predicted 83% of the firms. he The accuracy of the model is 83% the model classified correctly 83% of the firms. The specificity of the model is 97% which indicates that from not fast growing firms the model correctly estimated 97%. The sensitivity of the model is 11% which means the model predicted the fast growing firms by 11%.

Industry Analysis

The main data set contained information about two major industries, which we considered them separately. For the both industries we ran the prediction analysis. As a result of the analysis, received the analysis for both industries for logit models one and three, lasso model and random probability forest models. Moreover we used Confusion table on holdout with optimal threshold for the both service and manufacturing industries. The manufacturing firms data set has . From the and The result of Confusion table on holdout with optimal threshold for the manufacturing is as following: The table shows that accuracy of the manufacturing is 78%, sensitivity is 10% and specificity is 95%. The RMSE for the MODE 3 the best selected model in manufacturing industry is 0.377 the area under the curve AUC 0.641 is and the expected loss is 0.333. In the service industry based on the best selected model RMSE is 0.344, AUC is 0.691 and expected loss is 0.273. As a result we can say that based on the accuracy of the model across whole data set and sub sets of manufacturing and services, The complete data set has same accuracy as service industry meaning they classify the right category 83%. However, the manufacturing industry only classified 78% of the firms correctly. The second measurement is sensitivity which is the proportion of true positives among all actual positives. In this case study sensitivity is the actual fast growing firms. Based on the Confusion table on holdout with optimal threshold in main dataset, in the two sub sets of service and manufacture industries, the actual data set predicted 11% of the actual fast growing firms and the other two sub groups performed the same and predicted 10% of the actual fast growing firms. The third classification measurement in the data set is specificity. It means predicting the non fast firms correctly is highest in the main data set, services, and finally manufacturing industry, with specificity of 97%, 96%, and 95% consecutively. As a result the service industry performed better compared to the manufacture industry.

Summary

The final model which was chosen for this case study was Model 3 which has the variable of firm details and engine variables which are firm financial information of balance sheet and profit and loss. The The accuracy of the model is 83% the model classified correctly 83% of the firms. The specificity of the model is 97% which indicates that from not fast growing firms the model correctly estimated 97%. The sensitivity of the model is 11% which means the model predicted the fast growing firms. The result is a helpful tool for the firms to predict which firms are fast and non fast growing. Moreover, we compared across industry result considering the same single loss function and carrying out the exercise for different industries. All result were predicted based on the logit model 3. The performance of whole data and all models were evaluated.