Finding Fast Growing Firms

Data Analysis 3 - Assignment 3 - Prediction and Classification

Peter Kaiser & Xibei Chen

15 February 2022

Executive Summary

The purpose of this project is to build a model predicting the probability of a firm being fast growing to support investment decisions. The criteria for being fast growing is measured by us as the average sales growth rate above 15% for the next two years. The probability prediction paves the way for further classification to differentiate firms from fast growing to not fast growing, by applying the optimal threshold which is calculated by minimizing the average expected loss. Our final model of choice is random forest for probability, as it gives the lowest cross-validated RMSE and the highest cross-validated AUC, despite the drawback that random forest is a black box model and it would be hard for us to explain investment choices to other investors. Then for classification, we first define our loss function that for a bad investment we lose 1000 Euros, whereas to miss a good investment opportunity costs us 3000 Euros. Random forest again performs the best among all the models, as it gives the lowest averaged expected loss. However, according to the confusion matrix, we see that the classification model does not really perform well, as the accuracy is only 51% on holdout sample. And there is not much difference for different industry subsamples. As the performance is no better than random guessing, we would suggest to conduct further research to get a better model by optimizing our feature engineering process and exploring other classification models as well such as GBM classification.

Data Management

The original dataset was collected and already cleaned by Bisnode. The raw data represents all of the registered companies between 2005 and 2016 in a medium-sized European country. Data management includes mainly three parts in our project: Label Engineering, Sample Design and Feature Engineering.

Label engineering

We explored a few alternatives, eventually to take care of the class imbalance issue, and to take longer term and easy interpretability into consideration, our final decision is that firms with average annual sales growth rate above 15% for the next two years are categorized as fast growing.

Sample Design

We only use the observations in year 2012. To mitigate the effects of extreme values, we only focus on the small and medium enterprise sector captured by firms' sales, namely only keeping firms below 10 million euros of annual sales and droping firms with sales below 1000 euros. Furthermore, we also excluded firms that had invalid sales growth rate due to lack of information for certain years. Firms without values for other key variables such as firm age, region and industry were also dropped.

Feature Engineering

We added some features to capture potential non-linearity such as squared age and squared log sales. Some flag variables were also created to indicate foreign management, if there is asset problem, etc. We have also winsorized some variables to mitigate the effect of extreme values, such as CEO age and log level of sales difference. After feature engineering, some observations are dropped due to lack of information of key information such as firm age, share of foreign CEOs, industry category, etc.

At this point, we have 15835 firms in our dataset, among which 6010 or around 38% of firms are categorized as fast growing in the next two years.

Predictor Variables and Model Setup

First, we created following variable groups for later model choices: basic firm variables, firm financial quality variables, extra financial variables group 1, extra financial variables group 2, flag variables, growth variables, human resource variables, firm demographic variables, interaction between industry and some other variables, and interaction between log sales and some other variables. Then we set up the 5 simple logit models with increasing model complexity for further analysis. And we use the same predictor variables of simple logit model 5 for logit lasso. Lastly we set up a random forest model excluding interactions, modified features and flag variables, as random forest can catch those information automatically.

Probability Prediction

First, we use random sampling to create a training set with 80% of the observations, the left 20% will be used as holdout set for evaluation our model choice. Furthermore, to find the best performing threshold-agnostic model, we use 5-fold cross-validated RMSE as well as the average area under the curve (AUC) for each model. In terms of tuning parameters, for lasso we arbitrarily set alpha as 1, and set the lambda parameter to be between 0.1 and 0.0001 letting the machine decide the optimal lambda parameter, which turns out to be 0.0046. For random forest we set the number of randomly chosen variables at each split to be either 5, 6 or 7, as they are around the square root of the total number of predictors used to estimate the model. The number of observations in the terminal nodes of each tree is set to be either 10 or 15. It turns out the optimal random forest model has 15 observations in the terminal nodes and 5 variables randomly chosen at each split.

Model Comparison (threshold-agnostic)

When we do not have a loss function, we compare model in a threshold-agnostic way.

Table 1: Model Prediction Performance Summary (threshold-agnostic)

	Number of Predictors	CV RMSE	CV AUC
M1	11	0.4763	0.6045
M2	18	0.4697	0.6420
M3	35	0.4663	0.6575
M4	78	0.4654	0.6600
M5	152	0.4655	0.6621
LASSO	56	0.4643	0.6474
Random Forest	44	0.4627	0.6674

From the above summary table, we can get following conclusions.

1. As M4 has slightly better CV RMSE than M5, we can say that more complex models do not always generate better predictions, after certain point, including more predictors might result in overfitting training data.

2. Our Logit Lasso model gets better CV RMSE than all the above simple Logit models, however it does not get the best CV AUC. This is due to the fact that here the setting for Lasso is to optimize RMSE not AUC.
3. The model with best prediction performance is Random Forest, as it has the lowest CV RMSE and the highest CV AUC.

Classification

For turning predicted probabilities into classifications, we need a classification threshold. To determine the threshold, we need to define our loss function. Then we can look for the optimal threshold in each model. Eventually we can decide which model is the best model based on lowest average expected loss.

Define Loss Function

To define our loss function , we need to assign a cost value to False Positive (FP) and False Negative (FN) classifications. First we need to clarify that in our project False Positive means bad investment, that we classify a firm as fast growing and we invest in it, while in fact it is not a fast growing firm, hence we would lose our investment money or not generate as much profit as we expected. False Negative means the opposite, that we classify a firm as not fast growing and do not invest in it, while in fact it is fast growing, hence we would lose the investment opportunity. We decided that FN (losing an investment opportunity for a fast growing firm) would cost us 3000 Euros, more than FP (investing in a firm that is not fast growing) cost us 1000 Euros. Therefore, we set the ratio of the costs of FP to FN decisions as 1/3.

Model Comparison (find optimal thresholds)

For each model, we fit a ROC curve for each fold, saved best threshold for each fold and the expected loss value, then we calculated the average optimal thresholds and the average expected loss. We can compare models based on the average expected loss shown in the summary table below.

Table 2: Model Prediction Performance Summary (find optimal thresholds)

	Avg of optimal thresholds	Threshold for fold #5	Avg expected loss	Expected loss for fold #5
M1	0.2642	0.2433	0.6125	0.6160
M2	0.2299	0.2300	0.6030	0.6097
M3	0.2397	0.2273	0.5916	0.5931
M4	0.2423	0.2064	0.5910	0.5896
M5	0.2371	0.2203	0.5910	0.5935
LASSO	0.2392	0.2321	0.6006	0.5989
Random Forest	0.2780	0.2237	0.5826	0.5848

We can conclude that the random forest is still the best model at classification, as it has the lowest average expected loss. The second best is M4 or M5. The difference between their averaged expected loss is around 8 euros. However, if we review 1000 firms, this would result into 8000 euros difference.

Random Forest Classification with Optimal Threshold on Holdout Sample

We use the random forest model for classification on the holdout sample by applying the optimal threshold of 0.278 and get an average expected loss of 0.57, which is similar to that in the training sample. This means if we use this model for classification, we would expect to lose 570 euros on average.

Confusion Matrix for Best Model Random Forest As the last step of the classification process we create a confusion matrix from the classification on the holdout set using the random forest model. (unit: percentge, rows: predicted, columns: actual)

	not_fast_growing	fast_growing
not_fast_growing	18	4
fast_growing	45	33

According to the matrix, the rate of FP is 45%, while the rate of FN is 4%. This is due to the cost functions we defined. So the ratio of the less costly false decision is higher than the more costly false decision. To sum up, the accuracy of this classification is only around 51%, which is not an ideal classification, as it is just a bit better than random guessing.

Random Forest Classification on Different Industries

The dataset covers two industry categories: Manufacturing (ind2 26-33), and Accommodation and Food Service Activities (ind2 55-56). As an extra task for pair work, we separated the two industries in holdout sample for random forest classification performance. Firstly we explored the difference between two samples in terms of number of firms and share of fast growing firms. There are 2178 firms in Accommodation and Food Service, whereas 980 firms are in Manufacturing. Furthermore, in Accommodation and Food Service, there are around 36% firms that are defined as fast growing. And in Manufacturing, fast growing firms make up about 41%. Therefore, despite there is quite large difference regarding the total number of firms in different industries, the share of fast growing firms are relative close. Afterwards, we applied our random forest model on two industry samples, estimate the expected loss. For Accommodation and Food Service sample, we get expected loss of 0.56. For Manufacturing sample, we get expected loss of 0.61. Below we can also find the confusion matrices on each sample. For Accommodation and Food Service firms, the accuracy is 52%, and 49% for firms in Manufacturing. As we can see the random forest classification performs slightly better on firms in Accommodation and Food Service than on Manufacturing. But it is still not an ideal classification for either industry, as it is no better than random guessing.

Confusion matrix on Accommodation and Food Service Subsample (unit:percentage, rows: predicted, columns: actual)

	not_fast_growing	fast_growing
not_fast_growing	20	4
fast_growing	44	32

Confusion matrix on Manufacturing Subsample (unit:percentage, rows: predicted, columns: actual)

	not_fast_growing	fast_growing
not_fast_growing	13	5
fast_growing	46	36

Further Research

We would not recommend to use this model in live data, as it is not better performing than random guessing. So to get a better performing model, we might need data about more firms, explore other feature engineering choices and ideally get more predictor variables that are potentially linked to growth. In addition, we might look into other classification model choices such as GBM. If we can get a better performing model, to evaluate external validity in terms of time, we should test the model on samples across more years. We should also be careful when we want to apply the model to for example large corporate and other countries, as the external validity might be low.