Data Analysis 3. Third Assignment

Data exploration and cleaning process

I examined the dataset of bisnode firms to build models to predict fast-growing firms. The original dataset contains 287829 observations representing 46412 different firms, the dataset can be downloaded here. I examined firms that reported sales greater than 0 and filtered out extreme values that would have significantly biased my results. In order to see what the outliers were, I used the data summary function. Based on the results, I filtered out companies with very high (above €11.13 million) turnover.

	P0	P01	P25	P50	P75	P99	P100	Mean	SD
sales_mil	0.00	0.00	0.01	0.04	0.14	11.13	109.95	0.57	4.12

Table 1

To identify fast-growing firms, looked at firms that have experienced a 40% increase in turnover in a relatively short period of time (over two years from 2012 to 2014), which may indicate a successful growth strategy. Companies growing at this rate are likely to have significant expansion and profitability potential. It strikes a great balance between sensitivity and specificity by being able to identify a subset of companies that have experienced significant growth.

The first figure shows us the sales distribution of the firms in million €, we can see clearly that the distribution is right skewed. The second figure captures that how many of the firms were declared fast growing based on our definition, a bit more than 25% of firms were assigned as fast growing.

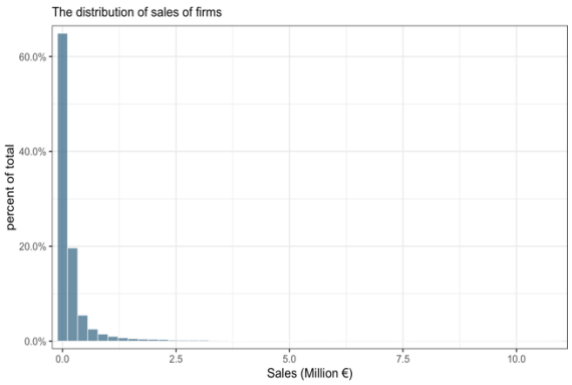

Figure 1

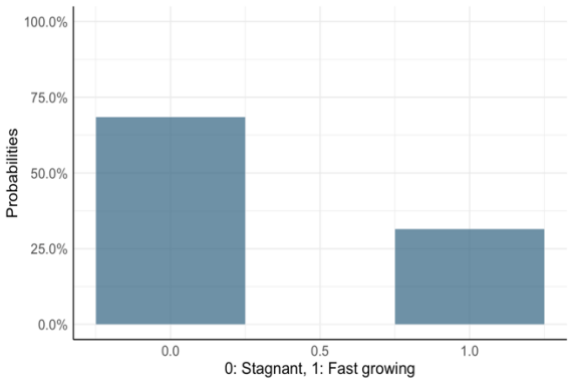

Figure 2

Model definition

I split the dataset into two parts, one of them is the training (80% of the data) on which I built the models, the other is the holdout (20% of the data) on which I tested the performance of them on unseen data.

Four models were built to predict the fast growing companies for firms. Two logit models, a Lasso model, and a Random Forest model. The logit models contained relevant information about the companies such as current assets, extra income, and liquid assets. The lasso model additionally to the logit models included interactions between the industry category codes and the essential firm information such as the age of the firm and the age and sex of the CEO. The Random Forest model included the same variables as the logit models.

Model evaluation

The model evaluation was performed based on Table 1. Based on the table, the Random Forest (RF) model has the lowest CV RMSE and the highest CV AUC, indicating that it has the best predictive performance among the models tested. Additionally, the CV expected loss of the RF model is the lowest among all models, which means that this model is expected to perform the best when it is applied to unseen data.

Therefore, based on these results, I would recommend selecting the RF probability model as the preferred model for predicting fast-growing firms. The RF model performed better than the logit models and the LASSO model in terms of both prediction accuracy and expected loss. The RF model is simpler and does not include the interactions and modified features included in the logit and LASSO models, which makes it easier to interpret and apply in practice.

	Number.of.predictors <int>	CV.RMSE <dbl>	CV.AUC <dbl>	CV.threshold <dbl>	CV.expected.Loss <dbl>
Logit X1	51	0.4495495	0.6511558	0.06812105	0.6844537
Logit X2	95	0.4485251	0.6561120	-Inf	0.6845598
Logit LASSO	156	0.4475206	0.6406234	0.05884198	0.6843634
RF probability	34	0.4203163	0.7675786	0.17237524	0.6313316

Table 2

Random Forest model

Since the best performing model is the Random Forest model, I decided to investigate it further. The ROC plot shows the tradeoff between true positive rate (TPR) and the false positive rate (FPR) at different classification thresholds for a given model, the optimal classification threshold is represented on the plot by the point on the curve, which maximizes the TPR and minimizes the FPR.

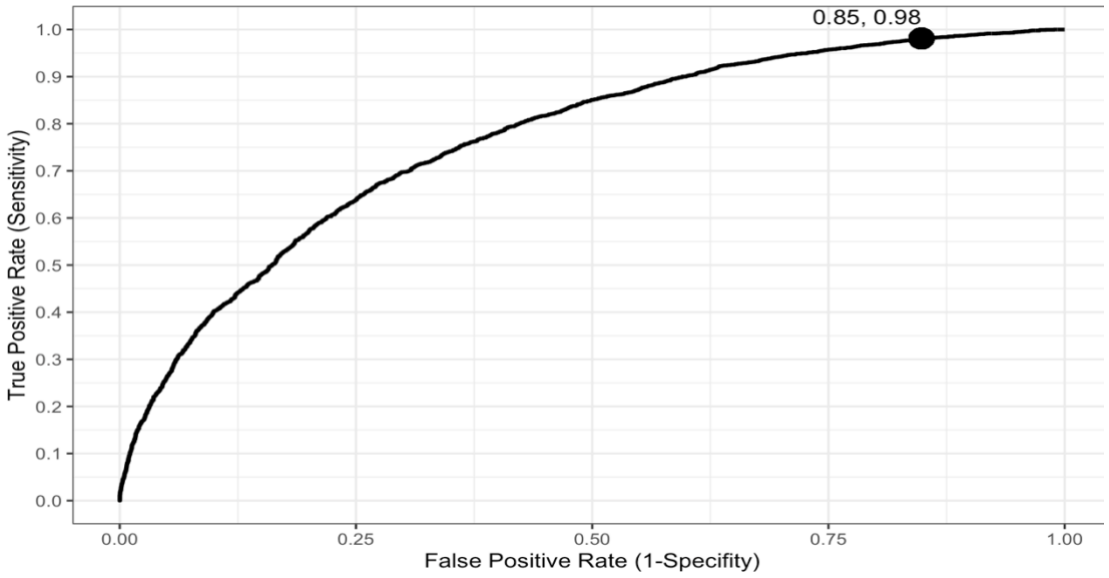

Figure 3

The confusion matrix shows that the random forest model correctly classified 6,082 firms as non-fast-growing and 875 firms as fast-growing, while incorrectly classifying 2104 non-fast-growing firms as fast-growing (false positives) and 356 fast-growing firms as non-fast-growing (false negatives). The model has a relatively high number of false positives, which could lead to missed opportunities for identifying fast-growing firms. However, the number of false negatives is much lower, indicating that the model is generally good at identifying firms that are likely to experience significant growth.

Prediction <fctr>	Reference <fctr>	Freq <int>
no_growth	no_growth	6082
growth	no_growth	356
no_growth	growth	2104
growth	growth	875

4 rows

Table 3

Summary

In summary, the data exploration and cleaning process involved filtering out extreme values and examining firms that reported sales greater than 0. Fast-growing firms were defined as those experiencing a 40% increase in turnover in a relatively short period of time. Four models were built, including two logit models, a Lasso model, and a Random Forest model. The Random Forest model performed best based on the following metrics, with the lowest CV RMSE, the highest CV AUC, and the lowest CV expected loss. The ROC plot and confusion matrix further supported the strong performance of the Random Forest model, I would recommend the RF model as the preferred model for predicting fast-growing firms.

If you are interested in the whole coding process please click here.