Predicting fast-growing firms in the EU using Bisnode data from 2010-2015

Quick overview

The data for the prediction exercise is collected by Bisnode and it contains detailed company data from a middle-sized country in three selected industries (auto manufacturing, equipment manufacturing, hotels and restaurants), between 2005-2016.

The data is restricted to cross-section as the aim is to build several models and select the best in predicting fast growth of these firms. In order to do so, most importantly the target variable has to be defined – what is growth and how fast it should be?

Growth can be incorporated by the increase of:

- Sales,
- Profits,
- Number of employees.

While the speed of this growth can also be growth in 1 year, in 2 years or an average of growths over 3 years. After specifying these targets, thresholds also need to be assigned, above which a firm is considered as *fast-growing*. Some of the potential targets and their thresholds can be seen in Figure 1. Thresholds are moderate (twice or even one and a half times larger amount than before), but the share of potentially fast-growing firms is low across all targets (around 10%).

Finally, **2-year change of sales** was selected as the target variable, as sales data is the most reliable in the sample and 2 years (especially when starting from a low level) can result in great increases, but it also means that this growth is rather permanent than ephemeral.

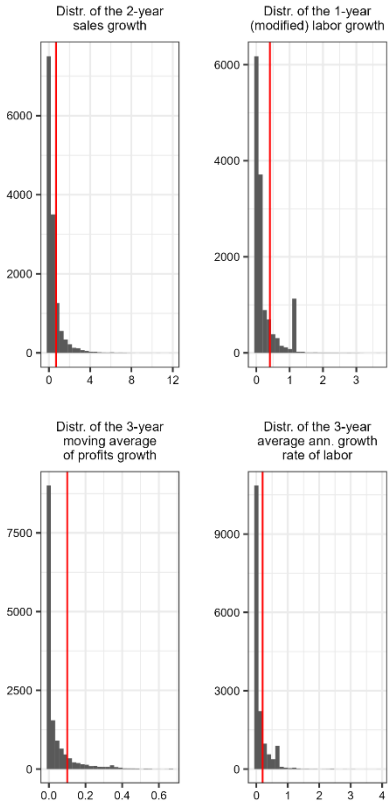

Figure 1

Among the most important features, there appears firstly the unchosen potential targets (profits and number of employees), industry indicators, balance sheet and asset characteristics, age of the firm, HR-related aspects (CEO age and gender, management). The data was very sparse in some cases, which required computations and dismissal of data.

Probability modelling

As the target variable is binary, logistic probability models are constructed to model this relationship. Several models were computed and tested, namely:

1. Multiple (5) Logistic (Logit) regressions with gradually additional number of features
2. Logit LASSO model based on the features of the best-performing Logit model
3. Random Forest (RF) on all available simple features.

Apart from the cross-validated RMSE, in the probability modelling setting one can compare the accuracy of the models along the share of False Positive (FP) and False Negative (FN) predictions. Two metrics incorporating the trade-off between the two are the ROC-curve and the area under the curve (AUC) along the dimensions of sensitivity and 1-specifity, with respect to a probability threshold. In Figure 2, the AUC in green is visible for Logit Model 4.

The performance of all the models is summarized in Table 1, and even though the Random Forest model has the lowest cross-validated error and the highest AUC-values, Logit Model 4 is very close. The Logit LASSO does not perform well despite that it is built on Model 4.

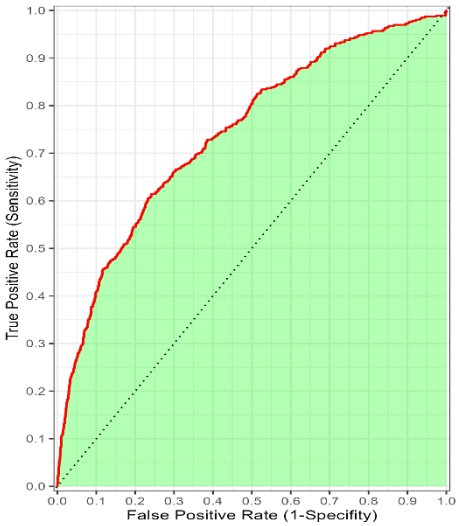

Figure 2

The results are promising as the RF model requires no a priori model or feature selection, it only needs some basic tuning, it is more difficult to formulate, whereas the Logit Model is based on domain knowledge.

Model	CV RMSE	CV AUC
Logit (Model 4)	0.319	0.734
Logit LASSO	0.336	0.711
Random Forest	0.316	0.742

Table 1

Classification

So far probabilities have been predicted but they were not assigned to classes (fast-growing or not). This requires a *Loss-function*, which crucially depends on how the partner weighs the FP and FN prediction errors. It might be reasonable to suppose that missing on a fast-growing firm (larger FN) has a higher cost – lower potential returns –

than labelling a non-increasing (or slowly) firm as fast. The suggestion for this function is FN/FP = 5. Table 2 shows that the smallest expected loss on average occurred at the RF model, but again, Model 4 is very close.

Model	Average expected loss
Logit (Model 4)	0.465
Logit LASSO	0.495
Random Forest	0.463

Table 2

The one last step is to define a threshold so that the classification takes place such a way that it minimizes the loss-function. This can be seen in Figure 3, where the RF model minimizes the function with a threshold = 0.16.

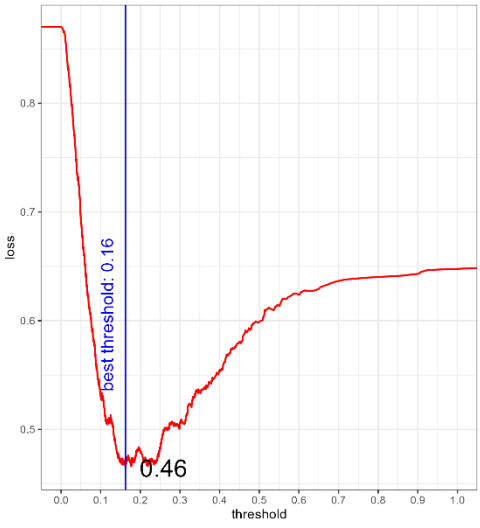

Figure 3

Predictive performance

Table 3 lastly summarizes that the two very different models are indeed similar in every aspect: the errors of the prediction on the holdout set is almost identical.

Model	RMSE on holdout
Logit (Model 4)	0.316
Random Forest	0.314

Table 3

Finally, Table 4 shows the confusion matrix of Model 4's predictions, with the optimal threshold. There are very a few fast-growing firms, but the model aims to predict them well, because FN (228) is more costly.

Reference/ Prediction	Not fastgrowth	Fastgrowth
Not fastgrowth	2564	228
Fast growth	547	236

Table 4

Conclusion

In this analysis, the task was to predict fast-growing firms in the Bisnode dataset. Several possible targets were introduced, from which 2-year increase in sales was selected. The data quality required many imputations and deletions, but the results are reasonable. The logistic model with HR-variables and quadratic terms was on par with the Random Forest model, resulting in the lowest expected loss, which were designed to minimize the misclassification of actual fast-growing firms.

Balint Parragi