CEU – DA3 : Growth rate prediction

Introduction and business problem

The purpose of this report is to document the detailed model building steps for the "*Predictive models for high firm growth*" project. The main objective of this analysis is to build a predictive model for fast company growth which enables the client to estimate the probability of a firm being successful on the market. For the model estimation cross sectional data on small and medium sized companies were used for the 2012 year. For the definition of a firm being "fast growing" I used 30% annual growth rate as cutoff value which was calculated as the annualized average rate of sales growth between 2012 and 2014. Three different prediction models were estimated and compared to select the model with the best performance: Logit model with 5 different specifications, Logit model with LASSO and a Random forest model.

Data and sample design

Data on companies was loaded from the *Bisnode* database which contains detailed information about the European companies. For this analysis the 2012 cross section was used with the following sample selection criteria:

- The analysis focuses on small and medium companies therefore companies with sales revenue above 1000 EUR and 10 million EUR were considered only.
- Leading and lagging indicators were derived using the 2011-2014 time interval. In order to ensure data quality of the financial indicators, firms with complete four year data were considered only. Firms with incomplete financial information were also excluded from the sample.
- Firms with extreme high (above 2000%) annual growth rate were removed from the dataset.
- The final sample consists of 11.949 firms and 100 variables.

Label engineering

The target variable is the annual growth rate of the sales revenue which was calculated with the compound annual growth rate (CAGR) formula. The growth was calculated between the 2012 and 2014 sales volume and were annualized. The growth rate over two years was expected to be less volatile than the year-on-year change which latter is more exposed to temporary shocks of the market. The distribution and the descriptive statistics of the target is in **A1.** To define the cutoff value for the "fast growing" firms I selected the 30% annual growth rate based on the distribution of the sales growth. I wanted to have enough firms in "fast growing" group for the prediction therefore set the threshold at the 85th percentile. With this split the proportion of the fast growing firms are about 16% in the sample.

Feature engineering

The predictors of the analysis were separated into three main predictor groups.

- Financial indicators of the firms including balance sheet and profit and loss statement items. Financial indicators with data quality issues were adjusted and flags were created to indicate the correction. Ratios were created from the financial variables by normalizing the asset side items with the total asset and the P&L item with the sales volume. Ratios were winsorized.
- Among the firm related indicators the age of the firm, the industry code, the gender majority, the region and the city type were involved.
- In case of the HR and leadership related predictors the age of the CEO, the indicator for the foreign CEO majority, the number of employees were used. Missing information regarding the age of the CEO and the number of employees of the firm were replace by the sample mean.

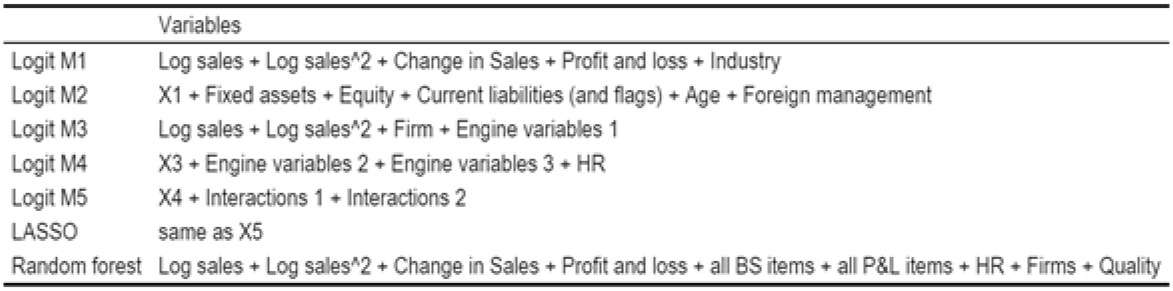
Modelling

For the modelling, train and holdout sets were created by randomly splitting the dataset into two partitions. 80% of the firms (9560 observations) were added to the train set and 20% of the firms (2389 observations) were kept for the holdout set. In case of all the presented models the modelling was done on the train set and the performance of the selected models were evaluated on the holdout set.

Model comparison

The goal of our analysis is to build a prediction model based on the observable firm data which is able to distinguish between the fast growing companies and the less prosperous firms in the future. For this, first we will develop predictions using our selected model(s) and then apply a loss function to make a decision about the threshold of the growth of the firms that minimizes the expected loss.

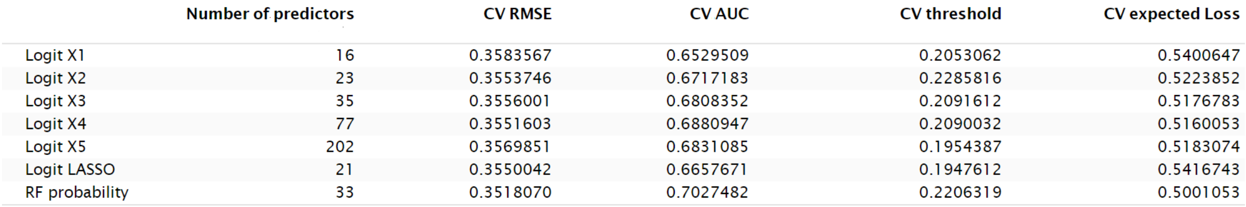
The following table summarizes the model specifications and the included predictors.



	Variables
Logit M1	Log sales + Log sales^2 + Change in Sales + Profit and loss + Industry
Logit M2	X1 + Fixed assets + Equity + Current liabilities (and flags) + Age + Foreign management
Logit M3	Log sales + Log sales^2 + Firm + Engine variables 1
Logit M4	X3 + Engine variables 2 + Engine variables 3 + HR
Logit M5	X4 + Interactions 1 + Interactions 2
LASSO	same as X5
Random forest	Log sales + Log sales^2 + Change in Sales + Profit and loss + all BS items + all P&L items + HR + Firms + Quality

We started the analysis with the simplest Logit specification which only included the sales, P&L and industry information of the firms. Then we enriched the model with more predictors including asset side items, equity and liabilities items, HR and firm related information. For Model 4 and 5 normalized financial indicators were used instead of the raw balance sheet and P&L items and interaction terms were added to the models. Finally, Random forest model was also estimated with the financial variables and its performance was compared to the other 6 models. In each case 5-fold cross-validation was used for the proper estimator performance evaluation.

The next table compares the models based on the calculated cross-validated RMSE and the AUC which were averaged across the 5 folds. The last two columns of the table will be discussed in the classification section.



	Number of predictors	CV RMSE	CV AUC	CV threshold	CV expected Loss
Logit X1	16	0.3583567	0.6529509	0.2053062	0.5400647
Logit X2	23	0.3553746	0.6717183	0.2285816	0.5223852
Logit X3	35	0.3556001	0.6808352	0.2091612	0.5176783
Logit X4	77	0.3551603	0.6880947	0.2090032	0.5160053
Logit X5	202	0.3569851	0.6831085	0.1954387	0.5183074
Logit LASSO	21	0.3550042	0.6657671	0.1947612	0.5416743
RF probability	33	0.3518070	0.7027482	0.2206319	0.5001053

We can see that the Random forest model outperforms all other models looking at both the cross-validated RMSE and the AUC. However, the difference in the performance measure between the models is very small so we can also consider to stick with a simpler Logit model which is much easily interpretable. Among the logit specifications Model 3, Model 4 and LASSO has almost the same RMSE. If we rank the logit models based on the AUC Model 4 has the highest value but includes almost double as much predictors as Model 3 which is just a slightly worse. Therefore if we would like to have a simple model we could choose Model 3 as an alternative to the best performing Random forest model.

Prediction

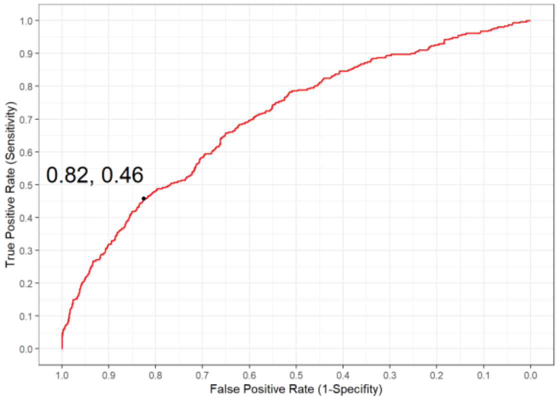
In order to estimate how well the model will perform on the live data we predicted the probabilities of a firm having high growth rate on the holdout set and compared it to the actual probabilities. The calibration curves in **A2.** show how the predicted and actual probabilities of each class compare to each other in case of the Random forest and the Model 3 specification. I used 10 bins for the comparison. We can see that both model predictions are fairly accurate for lower probabilities, but for higher probabilities the models systematically over predict the probability of success in the holdout sample.

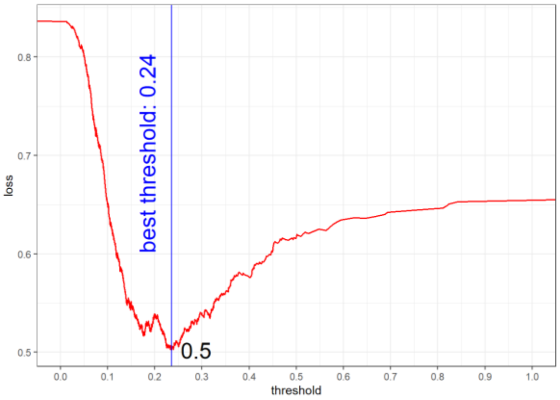
Classification and loss function

In order to support business decisions with our model we need to classify firms into “fast growing” and not “fast growing” groups based on the estimated probability distribution. For this we need to define a cutoff value above which we will label

A. ROC curve

B. Optimal threshold





the firm as "fast growing". Instead of using the proportion of fast-growing firms in the sample, say 10%, we defined a loss function that we can use to map decisions to their associated costs.

Our goal is to invest in companies that we expect to grow the fastest with outstanding anticipated profitability prospects. Therefore if we lose the opportunity to invest in these firms because our model makes too conservative predictions about their future growth than it considerably deteriorates our future profitability (False negative). Of course, there is a risk that we invest in companies that were anticipated to grow above 30% but the sales grew below our expectations or even have negative growth rate (False positive). When we decide about the relative cost of these scenarios we need to consider that the potential gains of a good investment can be higher by a unit than the losses by investing in a less prosperous company. We will consider the relative cost of these two errors 4:1 indicating that the cost of making too conservative decision and losing the opportunity to invest in a prosperous firm is four times higher than making a risky investment. We will express this relationship formally with the loss function.

We used the loss function to decide about the optimal threshold for the 'fast growing' classification. Our goal is to minimize the expected loss which comes from either making false positive or false negative decisions. The ROC curve visualizes the trade trade-off between these two kind of errors for each threshold and finds the optimal cutoff value through optimization. For each model the ROC curve was drawn and the models were ranked again based on the expected loss at the optimal threshold. To increase the precision of the estimation the expected losses were calculated for the folds and were averaged. The last two rows of the summary show the optimal thresholds and the corresponding expected losses.

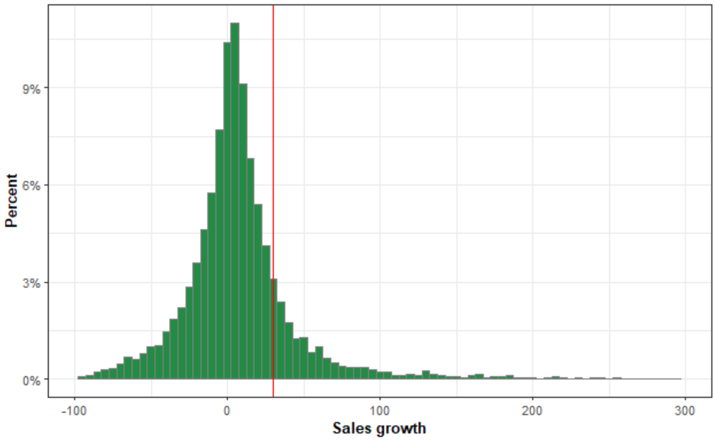
Again, if we compare the models based on the cross-validated expected loss, the Random forest model beats all the simpler models. We can also see that the optimal threshold is around 0.2, close to simply using the ¼ relative cost ratio. The difference in the expected loss across models is not high hence it could have been an option also to turn to a simpler logit model version, Model 3 which has about the same number of predictors as the Random forest model. Also from the calibration curves we saw that both models fail to make appropriate predictions at the higher edge of the distribution but the predictions of the logit model seem to be more accurate. The charts above visually show the ROC curve and the expected loss for different thresholds.

Summary of the results

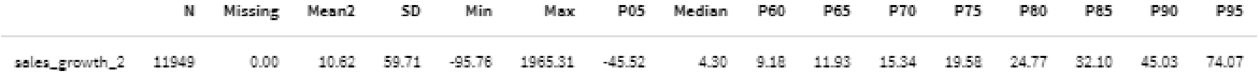
My selected model is the Model 3 logit specification with 35 variables which has 0.358 holdout RMSE. The model includes the logarithm of sales and its square, the normalized balance sheet and P&L items (engine variables 1), the characteristics of the firm. From the confusion matrix in **A3.** we can compute the proportion of correctly classified firms and the different measures of the classification which includes the accuracy ratio which is 80.9%, the sensitivity which is 35.6% and the specificity which is 84.1%. The proportion of correctly classified observations is 76.2% which is seemingly indicates fairly accurate predictions.

Appendix

A1. Sales growth

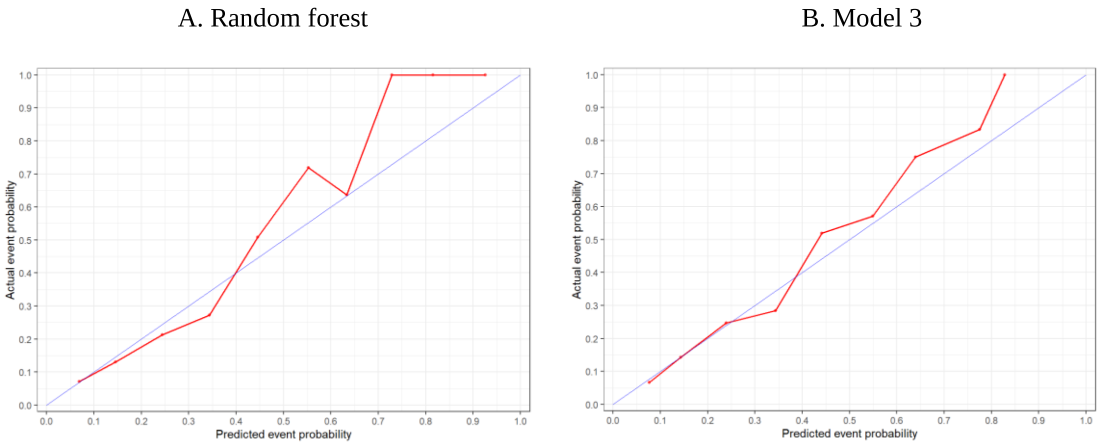


A1. Descriptive statistics of sales growth



	N	Missing	Mean2	SD	Min	Max	P05	Median	P60	P65	P70	P75	P80	P85	P90	P95
sales_growth_2	11949	0.00	10.62	59.71	-95.76	1965.31	-45.52	4.30	9.18	11.93	15.34	19.58	24.77	32.10	45.03	74.07

A2. Calibration curves



A. Random forest

B. Model 3

A. Confusion matrix (Model 3)

	not_fast_growing	fast_growing
not_fast_growing	1682	251
fast_growing	317	139