Assignment02

Predicting firm fast growth : Probability and classification

by Skirpichnikova Ksenia

Oscar Armando Leal Marcos

Data: "cs_bisnode_panel.csv" file shared for this assignment.

Dataset contains different types of variables indicating financial, management, employment, corporate, industry classification features of companies.

Label engineering

What is fast growth? There is no general rule what to consider as a fast growth for business, moreover, it depends on the industry. It is accepted that the company is considered as growing if firm's growth is higher than for the overall economy. We could have a few options to determine what is a fast growth:

1) To check the growth for the entire sample and choose a rate that is higher than for the entire sample as a fast growth.
2) To check the growth in each industry in the sample and determine the fast growth for each industry separately.
3) If we would know in which countries the firms operated, we could check the economy growth for considered periods and industries and accept the rates that are higher as a fast growth.

For our assignment we decided to calculate the compound annual growth rate[1] for the firms within the period 2010-2015 and considered as fast growth a threshold 15%. The 15% growth in sales may seem as a moderate one, nevertheless, the firm growing at this pace will double its sales within 5 years, that is a definite fast growth. Dependent variable is a binary variable called *fastgrowth*: 'one' if compound annual growth rate was higher than or equal to 15% and 'zero' when lower.

Sample design

We are going to predict fast growth for SME (small and medium enterprise), thus observations with sales higher than 10 million euro of annual sales and less than 10 000 euro were not considered. We are focused on a cross-section of 2015. Also, as we needed the 5 – years compound annual growth, we included only the firms that operated within the entire period and dropped observations with some missing values. As a result, for analysis we have a dataset of 10 564 observations. 27% of the firms have compound annual growth more than 15%.

We look at the distribution of a few basic variables: sales, profit and loss, total assets:

[1] Compound annual growth rate is a business and investing specific term for the geometric progression ratio that provides a constant rate of return over the time period.

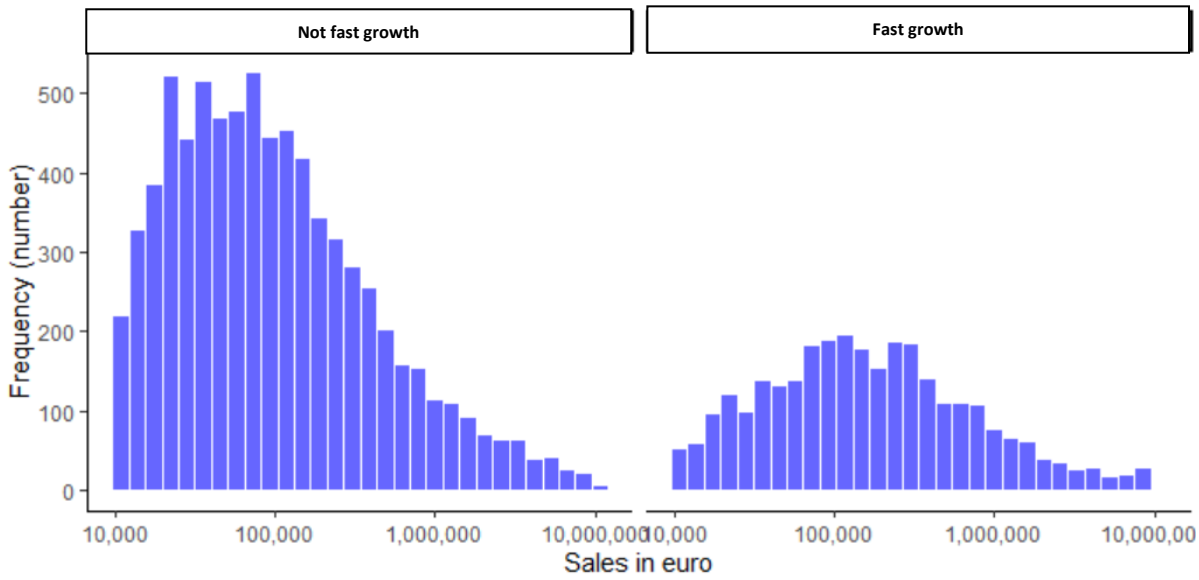
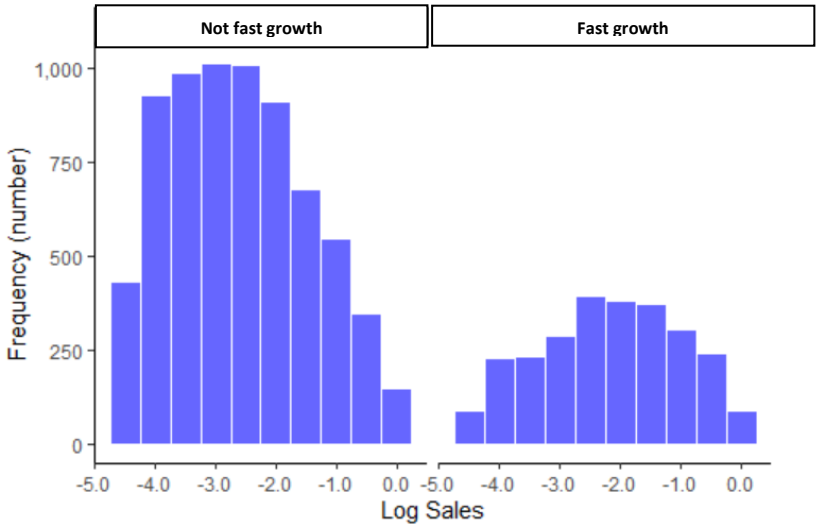
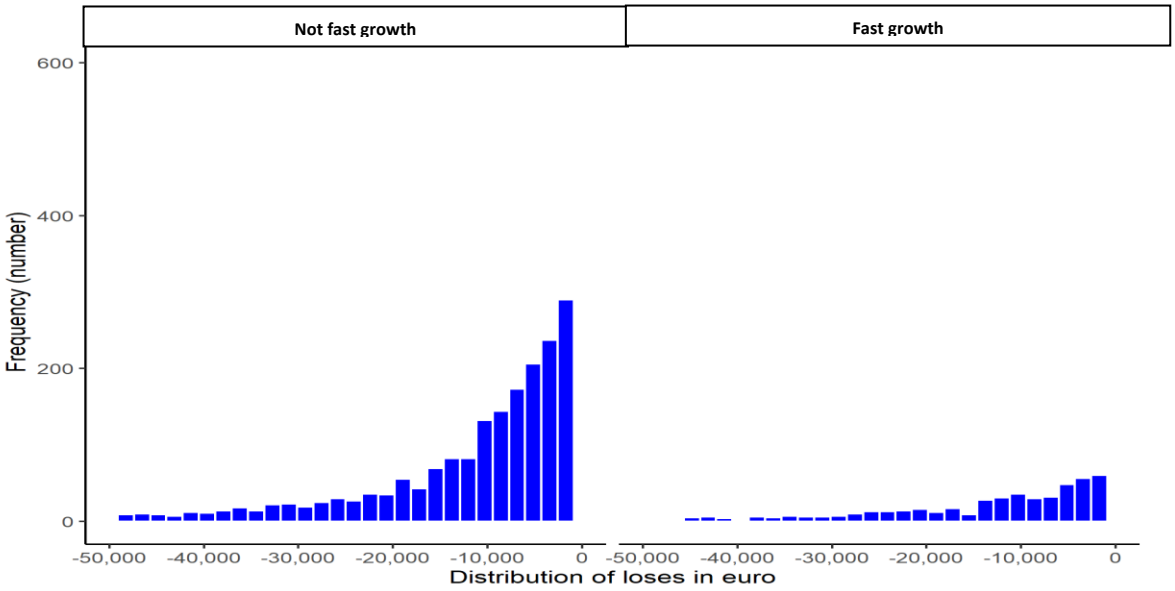


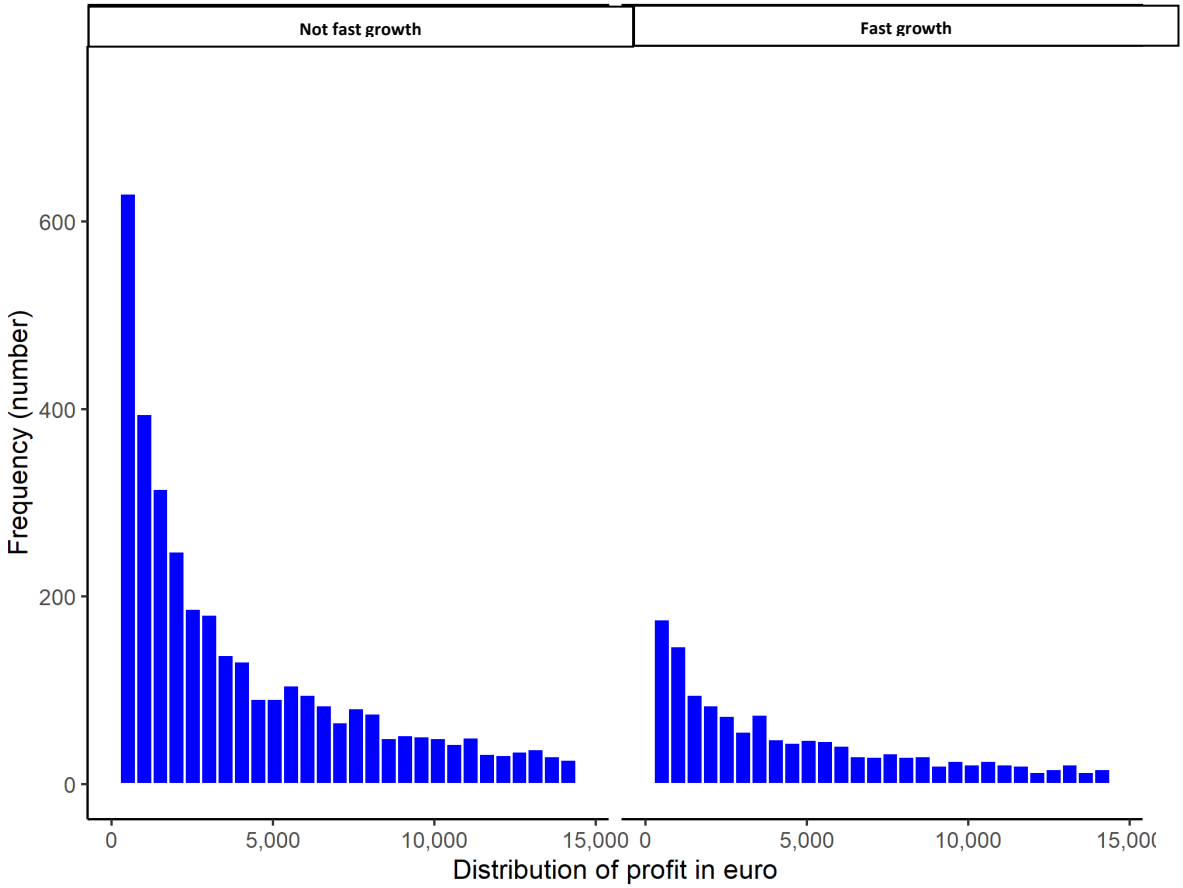
Table 01. Descriptive statistics for sales (euro):

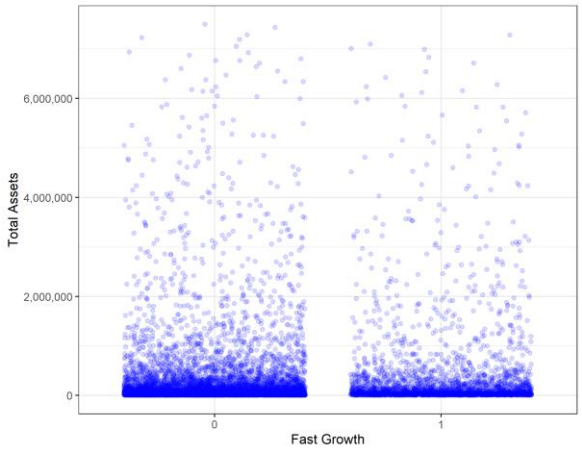
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10,000	35,291	93,715	405,547	284,115	9,964,481

The interquartile range for sales is 35,291 – 284,115 euro, median price 93,715 euro, presence in the sample of firms with high sales shifted the average to 405,547 euro.




Not fast growth
Fast growth
600
400
200
0
Frequency (number)
-50,000
-40,000
-30,000
-20,000
-10,000
0
Distribution of loses in euro


Not fast growth
Fast growth
600
400
200
0
Frequency (number)
0
5,000
10,000
15,000
Distribution of profit in euro



For this assignment we have used the same variables and feature engineering that was used for exit prediction.

PART I Probability prediction

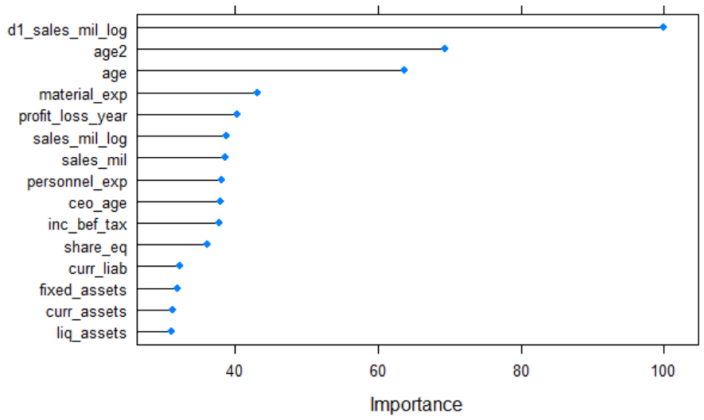
- Logit M1: handpicked few variables (p=7)
- Logit M2: handpicked few variables + Firm (p=10)
- Logit M3: Firm, Financial 1, Growth (p=67)
- Logit LASSO: M3 + LASSO (p=49)
- Random Forest (no interactions, no modified features besides sales and growth (log transformation))

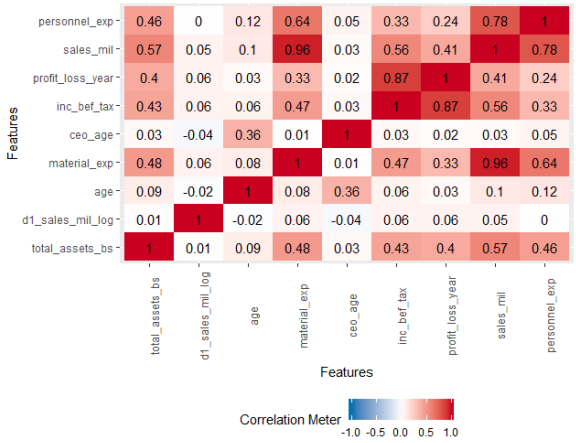
```
M1 <- c("age", "sales_mil_log", "d1_sales_mil_log", "profit_loss_year", "ind2_cat",
"ceo_age", "personnel_exp")
M2 <- c("age", "sales_mil_log", "sales_mil_log_sq", "d1_sales_mil_log",
"profit_loss_year", "ind2_cat", "ceo_age", "personnel_exp","total_assets_bs", "share_eq" )
M3 <- c("sales_mil_log", "sales_mil_log_sq", firm, engvar, engvar2, engvar3, d1, hr,
qualityvars)
# for LASSO
logitvars <- c("sales_mil_log", "sales_mil_log_sq", engvar, engvar2, engvar3, d1, hr, firm,
qualityvars, interactions1, interactions2)
# for Random Forest
rfvars  <-  c("sales_mil_log", "d1_sales_mil_log", rawvars, hr, firm, qualityvars)
```

Training dataset contains 8 447 observations (80%), holdout dataset contains 2 111 observations. We did 5-fold cross-validation. Each model was evaluated by its cross-validated average RMSE.

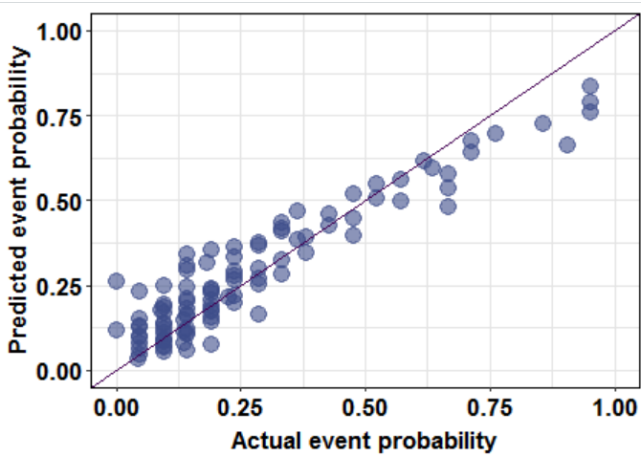
Model	Number.of.predictors	CV.RMSE
M1	7	0.4101138
M2	10	0.4094188
M3	67	0.3979680
Lasso	49	0.3990162
Random Forest	34	0.3903306

After doing a cross-validated performance, we decided to pick Random Forest as our favorite model since it has a considerable amount of predictors, and the RMSE is the lowest.





The variables describing scale and maturity of a firm have higher importance for predicting fast growth. The most important variable for fast growth is a log transformed variable indicating growth for the last year. Next two variables describe age of a firm in absolute value and quadratic form. Then we have variables for expenses for materials, profit or loss amount, sales amount, personnel expenses, age of CEO, income before tax, shareholder equity. Importance of these variables is relatively high (~40%) in comparison with growth for the last year.



Calibration curve for Random Forest

We plotted calibration plot for Random Forest on holdout dataset, observations were grouped (n=100). Based on the plot we can say that for firms where higher probability was predicted more firms actually were fast growing than we predicted, thus, we did not capture some fast growing firms. It can be that their fast growth can be explained by features we did not include in our model/did not have in our dataset.

PART II Classification

For classification we have the loss function: FP = 10, FN = 2.

If the model predicts that a firm has a fast growth, but it is actually NOT a fast growth by the definition of compound annual growth rate for the last 5 years (a false positive), then we lose 10 thousand euros.

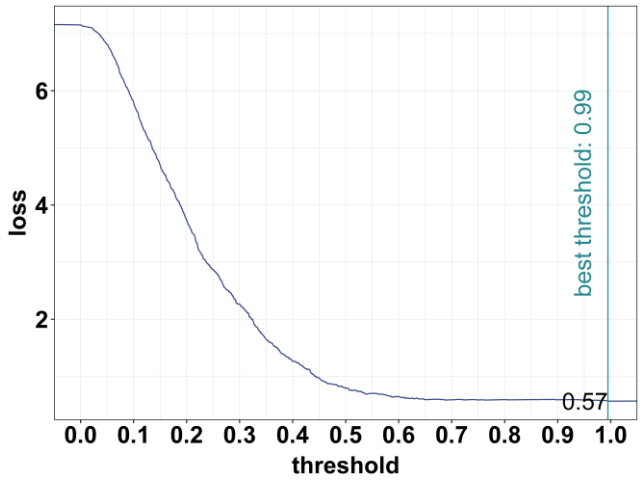
If the model predicts that a firm doesn't have a fast growth, but it is actually growing fast by t he definition of compound annual growth rate for the last 5 years (a false negative), then we l ose 2 thousand euros.

With correct decisions, we don't have any loss.

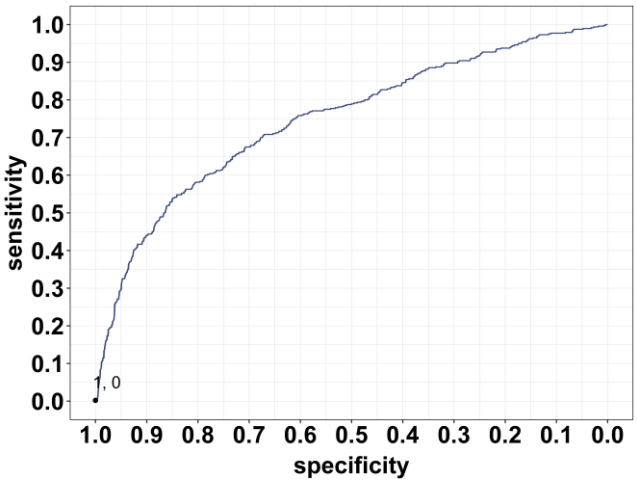
For each model we predicted probabilities, and looked for the optimal classification threshold , calculating expected loss.

We run the threshold selection algorithm on the work set, with 5 - fold cross validation.

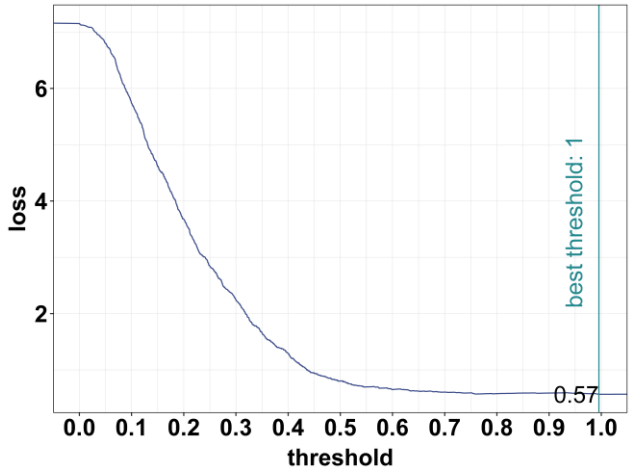
Finding the best threshold using lost function (example, 5th fold):



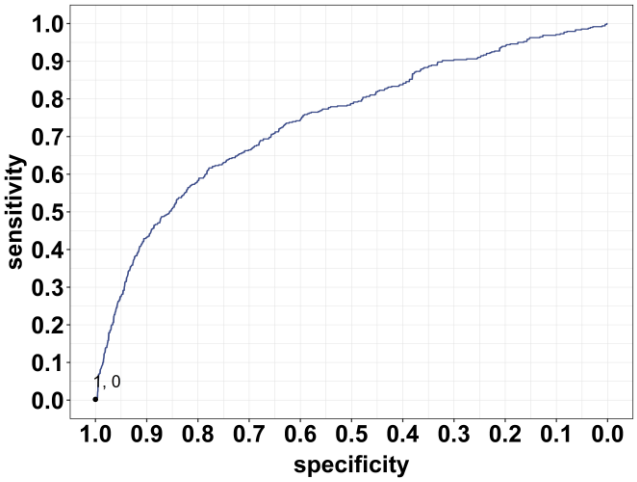
M1 loss function



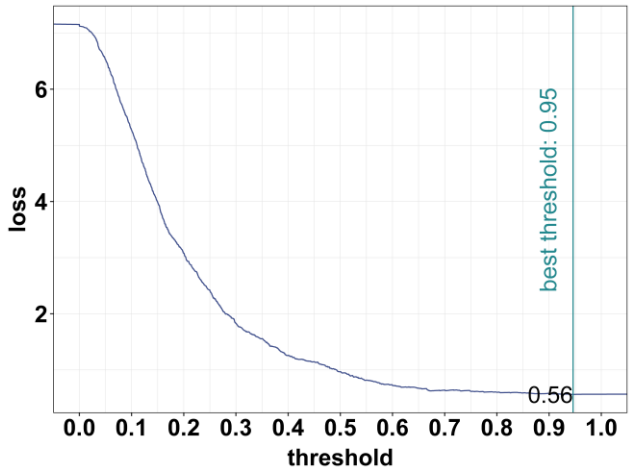
M1 ROC curve



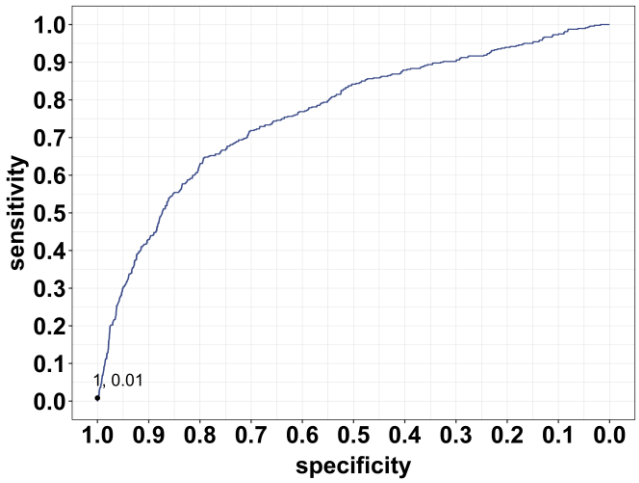
M2 loss function



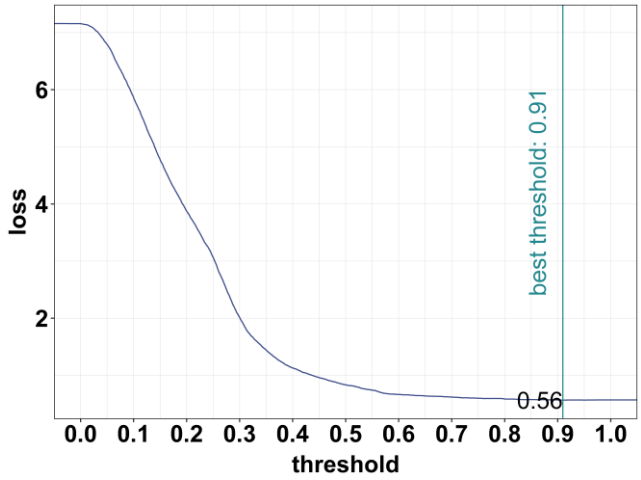
M2 ROC curve



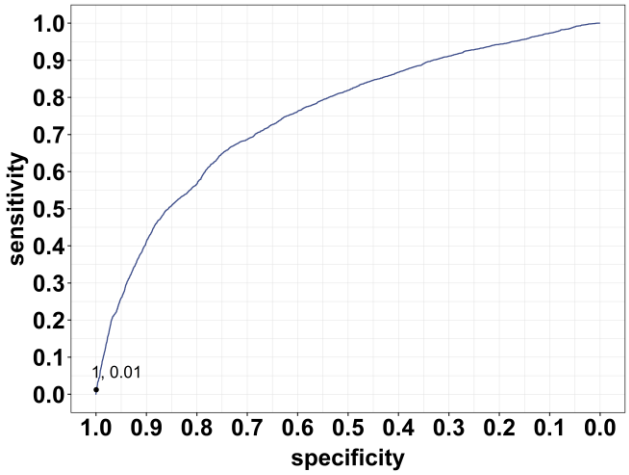
M3 loss function



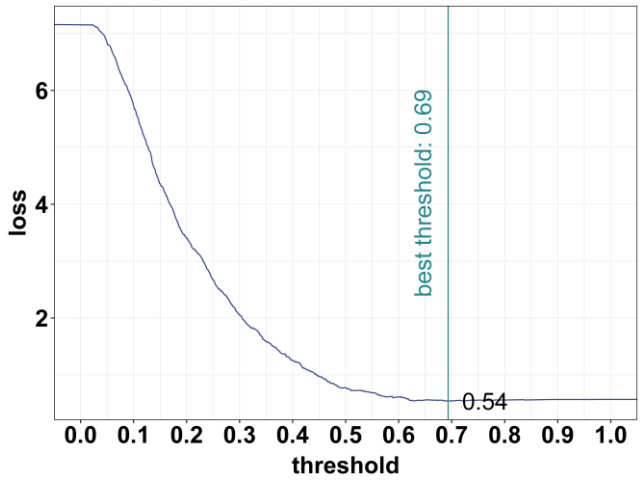
M3 ROC curve



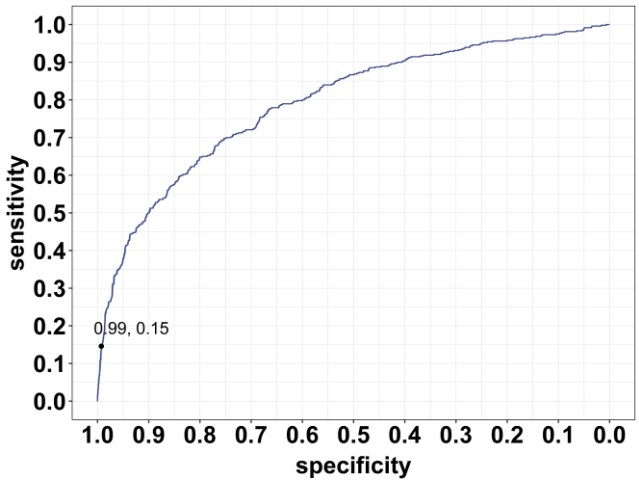
LASSO loss function



LASSO ROC curve



Random Forest loss function



Random Forest ROC curve

Optimal classification threshold is very high for logit models, it is ~100%. It means that to have lowest expected loss we need to classify as “fast growth” only if predicted probability was close to 1. Based on given loss function we do not tolerate true negative at all.

Therefore, with Random Forest model the optimal threshold of it is 69%, we can classify a fir m as a true "fast growth" if the predicted probability was higher than 0.69.
The best final threshold is the average among best thresholds for each fold. Estimation of exp ected loss was done on holdout set.

Best model is Random Forest with the smallest average (over 5 folds) expected loss for 0.506

	Threshold	Expected.loss	Holdout RMSE
Logit M1	0.8200587	0.5476078	0.4104147
Logit M2	0.8601644	0.5466604	0.4097237
Logit M3	0.9066177	0.5400284	0.3979703
Logit LASSO	0.8946155	0.5381336	0.3984113
Random Forest	0.7183525	0.5059214	0.3908759

PART III Discussion of results

Confusion matrix

Threshold =69%				
		Actual		
		no_fast	fast	Total
predicted	no_fast	1,521	509	2,030
	fast	5	76	81
Total		1,526	585	2,111

Threshold =69%				
		Actual		
		no_fast	fast	Total
predicted	no_fast	99.7%	87.0%	96.2%
	fast	0.3%	13.0%	3.8%
Total		100%	100%	100%

The rate for false positive is 0.3% and the rate for false negative is 87%. Thus, applying loss function we can calculate that the smallest average expected loss is 0.506.

Thus, building a better model could mean saving 41.68 euro /one observations:
((0.5476078 - 0.5059214)*1000 euro) for a company who is interested in better classification is a certain firm will have fast growth or not (Random forest vs Logit M1).

If such company estimates 1 000 firms, it means it is saving 41 680 euro if the Random Forest is chosen for this particular scenario instead of Logit M1.