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.

*Label engineering*

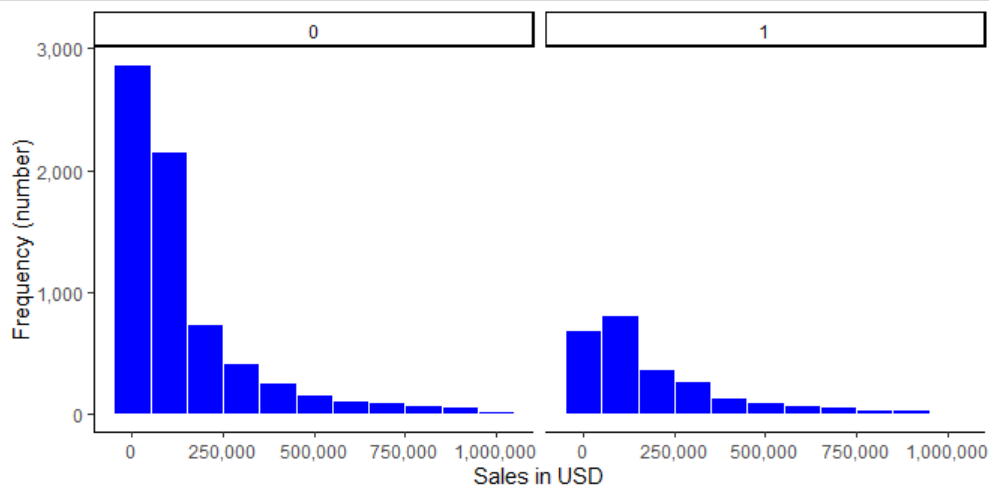
What is fast growth? There is no general rule what to consider as a fast growth for business, moreover, it depends on the industry. It is accepted that the company is considered as growing if firm’s growth is higher than for the overall economy. For example, a 15% growth in sales may seem as a moderate one, nevertheless, the firm growing at this pace will double its sales within 5 years.

For our assignment we calculated the compound annual growth rate[[1]](#footnote-1) for the firms within the period 2010-2015 and considered as fast growth a threshold 15%. 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%.

(1) Comparison of sales (for firms with sales < 1 mn euro)



**Fast growth**

**Not fast growth**

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.

Based on the plot above we can make an assumption

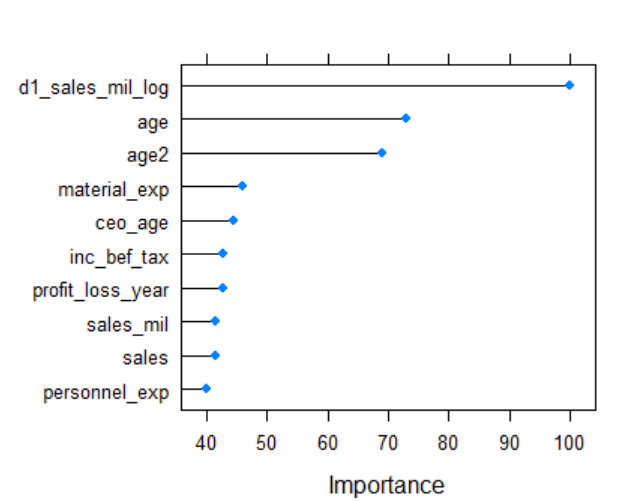
Добавить картинку с log sales

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

Sales : the variable was used in a transformed logarithmic form. Growth was also calculated using logarithmic form of sales.

Industries : out dataset contains 8 categories of industries, variable is a 2 digits NACE industry code.

Financial reports :



|  |  |  |  |
| --- | --- | --- | --- |
| Model Number.of.predictors CV.RMSE Holdout.RMSE | | | |
| X1 5 0.4285132 0.4273307 | | | |
| X2 12 0.4088384 0.4109604 | | | |
| X3 27 0.4002899 0.4014726 | | | |
| X4 69 0.3997036 0.4009388 | | | |
| X5 82 0.3984698 0.4009083 | | | |
| LASSO 75 0.3979136 0.4005383 | | | |
| Model Number.of.predictors CV.RMSE Holdout.RMSE | X1 5 0.4285132 0.4273307 | X2 12 0.4088384 0.4109604 | X3 27 0.4002899 0.4014726 |
| X4 69 0.3997036 0.4009388 | X5 82 0.3984698 0.4009083 | LASSO 75 0.3979136 0.4005383 |  |



*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=69)
* Logit LASSO: M3 + LASSO (p=76)
* 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 8 452 observations (80%), holdout dataset contains 2 112 observations. We did 5-fold cross-validation. Each model was evaluated by its cross-validated average RMSE.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Number.of.predictors | CV.RMSE | AUC |
| M1 | 7 | 0.4101138 | 0.7381311 |
| M2 | 10 | 0.4094188 | 0.7407422 |
| M3 | 68 | 0.3979680 | 0.7642000 |
| Lasso | 49 | 0.3990162 | 0.7660662 |
| Random Forest | 34 | 0.3903306 | 0.7840318 |

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.

*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 the definition of compound annual growth rate for the last 5 years (a false negative), then we lose 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.

|  |  |
| --- | --- |
|  |  |
| M1 loss function | M1 ROC curve |
|  |  |
|  |  |
| M2 loss function | M2 ROC curve |
|  |  |
|  |  |
| M3 loss function | M3 ROC curve |
|  |  |
|  |  |
| LASSO loss function | LASSO ROC curve |
|  |  |
|  |  |
| RF loss function | RF ROC curve |
|  |  |

Best model is Random Forest (RF) the expected loss is 1.07.

*PART III Discussion of results*

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. [↑](#footnote-ref-1)