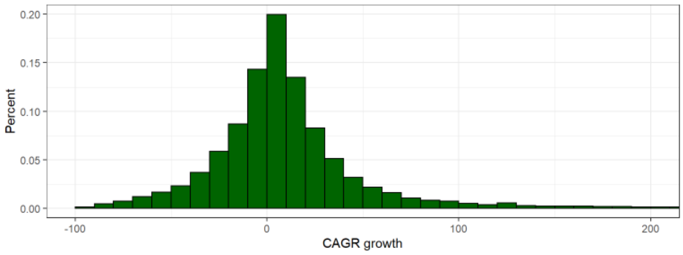
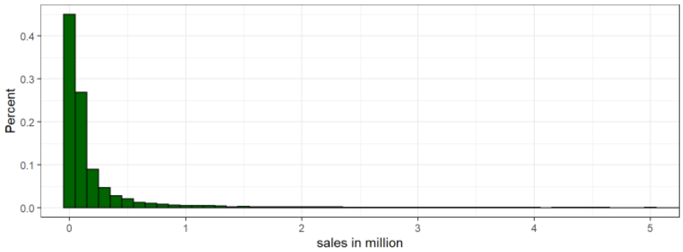
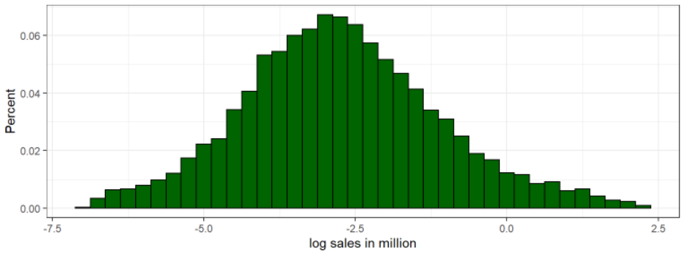
HW3 Salga Benjamin: Predicting fast growing companies

The code of the analysis is on: https://github.com/SalgaBeni/Data_An3/tree/main/HW3
The analysis aims to build a firm success prediction model for helping people in investment decisions, which company will fast grow in the future. The question, which this analysis tries to answer is really strong and common. Many investors or investment company working on calculating, which company will perform better in the future. Where to invest their money? Also, the analysis can be useful for businesses, who want to make long-term contracts with other businesses and they want to check, how the new partner will probably perform in the future.
For answering these questions, the analysis uses the Bisnode dataset, which contains many information about firms in a European country. The raw data was collected, maintained, and cleaned by Bisnode, but further data cleaning and data transformation was during the analysis.
The raw dataset has information about companies between 2005 and 2016 for industries in manufacturing. In the data analysis only focuses on firms between 2012 and 2013.
The data preparation started with 287 829 observations and 48 variables. The raw data has information as balance sheet, properties of the company, profit and loss elements. First, those observations were dropped which had too many NA values. For the analysis, many variables were created, like CAGR, which shows the success of the companies. A firm was considered successful and fast-growing if its CAGR is 30% or more on year-on-year basis. If the analysis would consider only 1 year of CAGR, that would be a too short period to help the prediction. However, the prediction aims to analyse the firms for the 2 years only. Next to CAGR, the logarithmic form of sales is created and checked. At the end of the data cleaning, the data set was 13 846 observations and 114 variables and it mean the analysis worked with 13 846 companies.
The following histograms show the distribution of CAGR, Sales, and the logarithm form of sales:







From the graph the distribution of vales is clearly visible. There is no evidence for any extreme values, which good for analysis. Before the model building and training, the dataset was divided into a train and a holdout set. The holdout dataset contains 20% of the observation, which was randomly selected and the left of the values went into the training dataset. The holdout dataset is used to evaluate the performance of the last model by simulating the use of its unknown, live data. The training dataset is not good for this purpose, since, for each of the models, 5-fold cross-validation was built, sot the dataset was divided 5 times to train and test sets. The training dataset has 11077 companies and the holdout dataset has 2769 companies.
The variables were separated into nine different groups: The first group is Raw variables. It contains key variables for prediction like profit, loss, and balance sheet. The next three groups are Engine variable 1-2-3. These groups contain profit and loss, balance sheet elements along, the squared form of some key variables along with some flags variables. The D1 group has variables measuring the change of sales. The HR group has variables age, CEO gender, and average labour number. The Firm group has the age of the company, region, and others. The last two groups are Interaction 1 and 2. Those contain interactions of variables.

Modelling

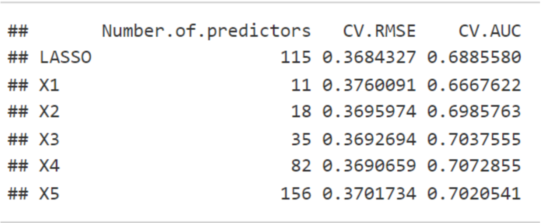
Logit models

The analysis started with five different logit models for predicting. The built models have different variables with increasing complexity. The first two models have arbitrarily chosen variables, which were mainly sales, profit-loss, and industry category of the firms. The following models have more and more features than the first two models.

The model comparison is based on two measures: the Root Mean Squared Error (RMSE) and the Area Under Curve (AUC). The measures averaged on the five different folds used during cross-validation. the results table shows the RMSE and AUC values are not very different for the five models. The best model is the fourth model (X4) since it has the lowest RMSE, and also this model has the highest AUC. So, the fourth model outperformed all the others based on RMSE and AUC. The X4 model is used as a benchmark, however, it is a quite complex model. The complexity comes from having all the financial variables, firm-specific details, some features of the growth of sales, and some variables about the CEO.

LASSO

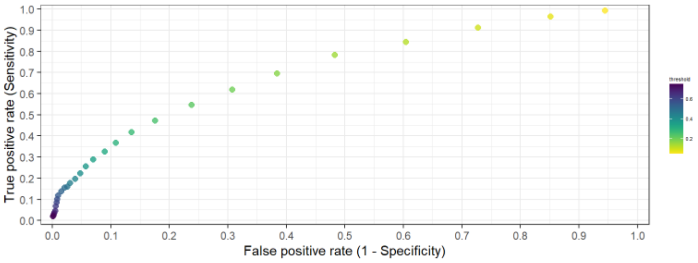
The five Logit models were followed in the analysis LASSO to find the model which contains the most important variables. This model is based on calculation and not manually set up like the logit models. LASSO uses the same variables as the fifth logit model did, but it uses interactions as well. The best LASSO model is compared under the same RMSE and AUC measurements. The best model has only a slightly better RMSE, but the AUC value by the fourth logit model is superior.

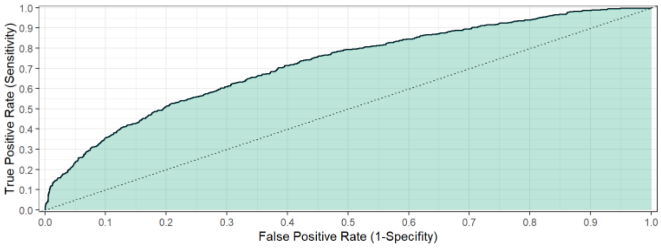


```
##       Number.of.predictors   CV.RMSE    CV.AUC
## LASSO                  115 0.3684327 0.6885580
## X1                      11 0.3760091 0.6667622
## X2                      18 0.3695974 0.6985763
## X3                      35 0.3692694 0.7037555
## X4                      82 0.3690659 0.7072855
## X5                     156 0.3701734 0.7020541
```

ROC curve

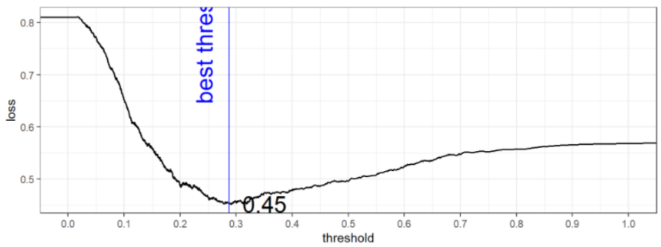
The ROC plot was set up for the best model, the random forest model. The first ROC plot was created with separated dots for different possible threshold values with a range between 0.05 and 0.95. The threshold values are shown in different colors. The second version has a continuous graph making easier understandable the Area Under the Curve. The negative side of this is the shift in threshold values between True and False positive rates. The green part is the AUC. With the lowest threshold values increasing the rate of True positives, there is an increase in the False positive rate as well. To find the optimal threshold values there is a need to set up the loss function.





Optimal threshold

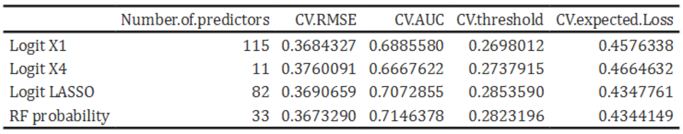
For finding the optimal threshold value, a loss function was set up for the problem of the analysis. The analysis contains some interesting details. So, the analysis aims to predict the fast growth of firms, but in case of a False-negative error, the situation is worse than normal, since the investor loses a good opportunity to invest, because that company will grow significantly in the next years. However, in the case of a False positive error, the analysis advises a firm to invest money in, but the company is not a fast-growing firm. The investor will not lose any money probably, only his/her money will increase slower. In the analysis is key to index the high cost of a False-negative error, so its cost was set 3 times more as a False positive error. With this set up the analysis could come up with the best threshold that minimizes the expected loss. The formula for the optimal threshold would be 1/(3+1)=0.25. This is a good benchmark for the calculation. The optimal threshold selection algorithm was run on the train data using the 5-fold cross-validation. The model with the lowest RMSE and lowest expected loss is the Random Forest model. It has the optimal threshold of 0.282. Both the formula and the algorithm have similar results.



Model choice

To find the best model for the analysis, I ranked them first based on expected loss. The model with the lowest expected loss is the Random Forest, then the LASSO model, after that is the simplest X1 logit model and the last one is the X4 logit model. The order is the same if I ranked them by AUC values. In the case of RMSE, the Random first model is the best. So, for the end of the analysis, the LASSO model will be used.

The numbers in some cases are quite close to each other, but the LASSO model has only slightly off numbers as the Random Forest, and also the Random Forest model is a black-box model.



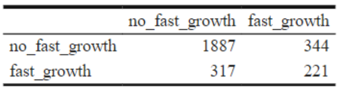
	Number.of.predictors	CV.RMSE	CV.AUC	CV.threshold	CV.expected.Loss
Logit X1	115	0.3684327	0.6885580	0.2698012	0.4576338
Logit X4	11	0.3760091	0.6667622	0.2737915	0.4664632
Logit LASSO	82	0.3690659	0.7072855	0.2853590	0.4347761
RF probability	33	0.3673290	0.7146378	0.2823196	0.4344149

Summary

The final model was the LASSO model, which was chosen by its performance. The LASSO model's accuracy is 76.1%, which means that 86.1% of the firms was ranked as good company. The actual, not fast-growing companies were predicted 80.5% correctly and fast-growing companies with 19.5% correctly, but 41% of those firms are fast-growing. The prediction of the fast-growing companies is really hard since its number is almost 8% compared to the total number of the firms. The prediction of the not fast-growing firms is much easier.

This model can be a tool for investors and investment companies to find which firms they should invest in. They can expect to see an amount of potentially fast-growing firms, from which they can expect 41% of those will be high growth firms.

Investors act and behave differently, so probably they have a different risk tolerance, so they can change it in the loss function. However, those changes will lead to totally different results. A risk-averse attitude will lead to a smaller amount of predicted fast-growing firms and a risk taker attitude to a higher amount.



	no_fast_growth	fast_growth
no_fast_growth	1887	344
fast_growth	317	221