Assignment 3: Technical Report

Alima Dzhanybaeva

Introduction

This project aims to build models that predict the fast-growing firms. A fast-growing firm is typically defined as a company that experiences a high compound annual growth rate (CAGR) over a certain period of time (2012-2014 in our case). The specific CAGR threshold that defines a fast-growing firm can vary depending on the industry, the size of the company, and other factors. However, in general, a CAGR of 20% or higher is considered a strong indicator of a fast-growing company. Therefore, the threshold of 20% is used in this project to divide companies into two groups: *fast_growth*, and *no_fast_growth*.

In total 7 models will be used for the prediction: 5 logit models, LASSO, and Random Forest.

Data preparation

The original dataset *"cs_bisnode_panel.csv"* includes 287,829 observations and 48 variables for the period 2005-2016. The file "ch17-firm-exit-data-prep.R" was used for the data preparation with the small changes and additions.

Filtering

- **year**: The initial time period of 2005-2016 was cut to 2010-2015.
- The firms that did not have data for all six years (2010-2015) were excluded from the dataset.
- **sales**: only the companies with sales between 1000 and 10,000,000 were left.

Creation of the new variables

- **status_alive**: this variable is equal to 1, if *sales* are more than zero or are missing; the dataset was additionally filtered only for firms that are still active, i.e. for which the value for this variable is not equal to 0.
- **cagr_sales**: this variable was calculated for 2012- 2014 with the help of *sales_mil* (sales in millions, EUR) using the following formula:

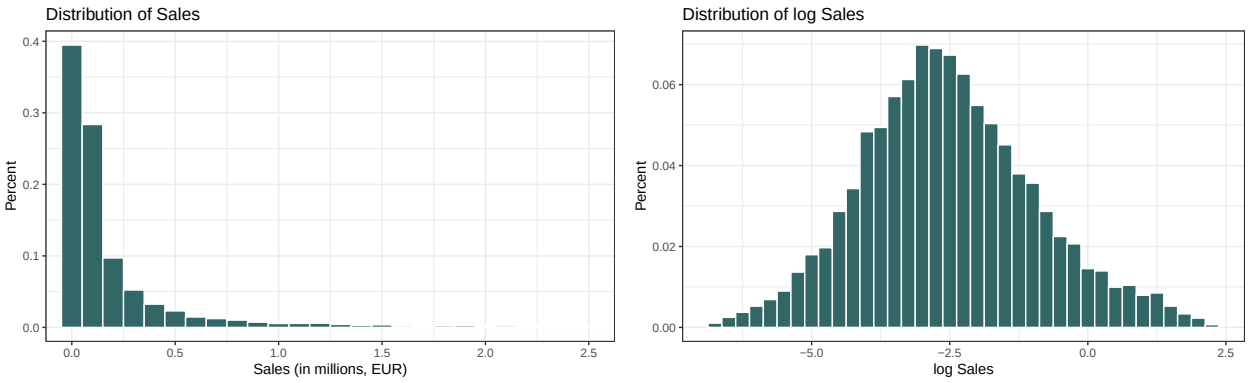
```
data <- data %>%
  group_by(comp_id) %>%
  mutate(cagr_sales = ((lead(sales_mil,2) / sales_mil)^(1/2)-1)*100)
```

All the variables that had missing values for *cagr_sales*, and for which the value exceeded 99% percentile (around 200%) were dropped from the dataset.

- **fast_growth**: firms for which the value of *cagr_sales* was higher than 20% were assigned with 1, and companies with a lower value with 0. The table below shows the number of variables that fell into each category:

fast growth	N	Percent
0	7963	76.11
1	2499	23.89

- **ln_sales** & **sales_mil_log**: the distribution of the sales is highly skewed to the right, therefore, the new variable that takes its logarithm was created.



- **age**: this variable was calculated by subtracting the year the company was established from the current year; further, the additional variable with the squared term *age2* was also added.
- **total_assets_bs**: was created by adding the values of three variables: *intang_assets*, *curr_assets*, and *fixed_assets*.
- **ceo_age**: the variable was generated by subtracting the director's year of birth from the current year; additionaly, three flags were added: (1) if *ceo_age* is less than 25, (2) if *ceo_age* is higher than 75, (3) if *ceo_age* is missing.

Imputation

- **intang_assets**, **curr_assets**, **fixed_assets**: the negative values for were replaced with 0; the flags that indicate these observations were additionally created.
- **labor_avg_mod**: the mean was imputed instead of the missing values for this variable; the corresponding flag was generated.

At the end we are left with 10462 observations and 118 variables.

PART I: Probability prediction

To identify fast-growing firms 5 logit models, LASSO, and Random Forest were run in this project. The dataset was divided into training and test sets, moreover, to identify possible overfitting cross-validation was used for training data.

Logit models

Overall, five logit models were constructed for the prediction, the complexity increased from 1 to 5. Model 1 being the simplest one included only 5 predictors and Model 5 being the most complex contained 84 predictors.

To evaluate the performance of each model the Root mean squared error (RMSE) and Area Under the Curve (AUC) were used. The table below displays the results for 5 logit models with corresponding values of RMSE and AUC.

	Number.of.predictors	CV.RMSE	CV.AUC
X1	5	0.4226723	0.5910082
X2	12	0.4196920	0.6210841
X3	28	0.4170039	0.6413403
X4	71	0.4174206	0.6418947
X5	84	0.4169640	0.6473579

As we can see from the graph, **Model 5** has the lowest RMSE and the highest AUC, indicating that it is performing the best among other logit models in both accuracy and precision measures.

LASSO

The table below shows the results for the most complex **Model 5** which includes 84 predictors and **LASSO** which shrank coefficients for all insignificant predictors to zero and left only 24.

	Number.of.predictors	CV.RMSE	CV.AUC
X5	84	0.4169640	0.6473579
LASSO	27	0.4163375	0.6275310

LASSO has the lower values for both RMSE and AUC, therefore, the choice of the better performing model may be ambiguous. However, I prefer **Model 5**, as the difference in RMSE is much smaller, than the difference in AUC. Additionally, *fast_growth* is quite imbalanced, therefore, AUC may be more important than RMSE, as AUC measures how well the model is able to distinguish between classes.

Random Forest

The results for **Model 5** and **Random Forest** are presented in the table below.

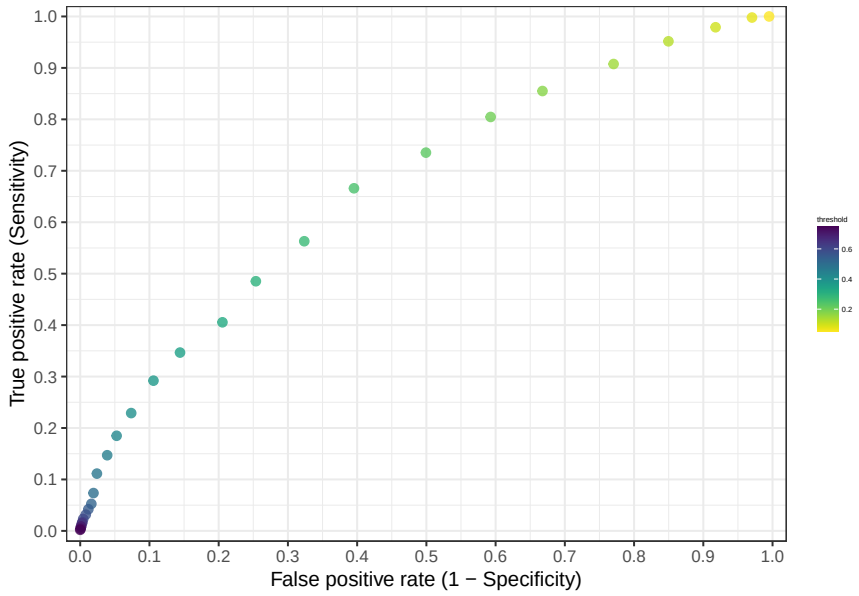
	CV.RMSE	CV.AUC
X5	0.4169640	0.6473579
Random_forest	0.4165901	0.6487081

As we can see from the table, in comparison with **Model 5**, **Random Forest** showed better results for both RMSE and AUC. Consequently, among all 5 logit models and LASSO **Random Forest** has the best predictive performance.

ROC Curve

A Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. It shows the relationship between the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings. The TPR represents the proportion of positive instances that are correctly classified as positive by the model, while the FPR represents the proportion of negative instances that are incorrectly classified as positive by the model.

ROC curve below is constructed for **Random Forest**, as it showed the best performance among all models. The area below points is equal to 0.65, as this value corresponds to the AUC for **Random Forest** in the last table.



PART II: Classification

Loss Function False positive: let's say that we classified firm as a 'fast-growing' one, however, turned out that it is not. If we invest in this company, we can still earn from it, as its cagr may be under the threshold of 20% but still can be positive.

False negative: however, if we classify 'fast-growing' firm into the opposite group we don't invest at all and therefore earn nothing.

Consequently, the assignment assumes that false negative costs us more than false positive, thus we assign FP=1, and FN=2.

The table below displays the the optimal classification thresholds, expected losses with previously defined loss function (FP=1, and FN=2) for the most complex logit model, LASSO, and Random Forest.

	Number.of.predictors	CV.RMSE	CV.AUC	CV.threshold	CV.expected.Loss
Logit 5	84	0.4169640	0.6473579	0.1572260	1.107282
Logit LASSO	27	0.4163375	0.6275310	0.1656744	1.237012
RF probability	36	0.4165901	0.6487081	0.1721368	1.080039

Not only **Random Forest** has the lowest RMSE, the highest AUC among other models, but also the lowest predicted loss (1.08). Second lowest loss is observed for the logit **Model5** (1.11), and the highest value is for **LASSO** (1.24). The optimal threshold is around 0.16 for all models.

PART III: Discussion of results

Confusion Table

The confusion table below was constructed for the model that performed the best, i.e. **Random Forest**, using the best threshold.

	fast_growth	no_fast_growth
fast_growth	411	1102
no_fast_growth	65	514

$$accuracy = \frac{TP + TN}{N} = \frac{411 + 514}{2092} = 44\%$$

$$sensitivity = \frac{TP}{TP + FN} = \frac{411}{411 + 65} = 86\%$$

$$specifcity = \frac{TN}{TN + FP} = \frac{514}{1102 + 514} = 31\%$$

Accuracy is equal to 44%, meaning that the model classified firms correctly only 44% of the time. The remaining 54% of predictions were incorrect.

The sensitivity is quite high (86%) indicating that the model is able to correctly identify a high proportion of positive cases, however, it came at the cost of a lower specificity (31%), which means there will be a lot of false positives.

Conclusion

The best model for predicting probabilities and classifying firms into 'fast_growth' and 'no_fast_growth' was found to be **Random Forest**. Among all 7 models (5 logit models, LASSO, Random Forest), it had the second lowest RMSE (0.417), the highest AUC (0.649), and the lowest predicted loss (1.08). The confusion matrix with the optimal threshold showed quite low values for accuracy and specificity (44% and 31% respectively), nonetheless, the sensitivity is high (86%). This is still better than the situation in which the specificity is significantly higher than the sensitivity, as the higher occurrence of false negatives would be much worse in our case.