Data Analysis 3 – Assignment 3
Balint Thaler
GitHub repo: https://github.com/balint-thaler/DA3-Assingment-3

Technical report

Introduction

The aim of the current report is to present some of the technical choices made during the modelling process. The main results and findings are summarised in the summary report. The current document is designed as a complement of the summary report.

Feature engineering

The original Bisnode dataset contains more than 280,00 observations which required extensive feature engineering so that the modelling process would be made possible. Upon inspection of the dataset, eight columns (“COGS”, 'finished_prod', 'net_dom_sales', 'net_exp_sales', 'wages', 'D', 'exit_year', 'exit_date') were dropped due to high proportion of missing variables. The age of each observation was calculated based on the year of the specific observation and the year of foundation of the firm.

As defined in the assignment, the definition of the default was that a company’s is alive in period t, but its sales is 0 or missing in period t+1.

Regarding financial variables, some obvious errors were corrected, and flags were added. These 223 cases of negative values for some balance sheet figures which were imputed with 0. For better comparisons between the firms, the total assets’ value was calculated from the individual balance sheet items, and all balance sheet-related variables were divided by it. With a similar logic, all profit & loss related variables were divided by the respective sales figures to again get values on a comparable basis. In all cases, the original, “raw” values were also kept.

Two further variables contained relatively high numbers of missing values: labour average and birth year. Missing values were imputed with the average of the respective figures and flags were added for imputed values.

Variable name	Number of imputations
amort	2
material_exp	2
personnel_exp	2
founded_year	31
ceo_count	31
foreign	31
female	31
inoffice_days	31
gender	31
origin	31
region_m	2
age	31
material_exp_pl	2
personnel_exp_pl	2

Table 1: imputations in the holdout set

Data Analysis 3 – Assignment 3
Balint Thaler
GitHub repo: https://github.com/balint-thaler/DA3-Assingment-3

Datasets

To create the predefined set, the original data was filtered for the industry code of "26", the year for 2014 and the sales figures for larger than 1,000 euros but below 10 million euros. The resulting dataset matches that of the description of the assignment in number of observations, number of defaults, average, minimum and maximum sales figures.

As described in the summary report, five different training datasets were created to experiment with not only different models but also different training sets to minimise the expected losses on the holdout dataset. The theory was that the more similar the training dataset is to the holdout data, the more accurate the classifications will be. For this reason, the sales figures, the year and the industry was taken into account when creating the different datasets. The year and the industry are given in the holdout data, and upon inspection of the distribution of the sales figures (see chart 1), we see that the vast majority of firms have a sales figure of less than 1 million Euros.

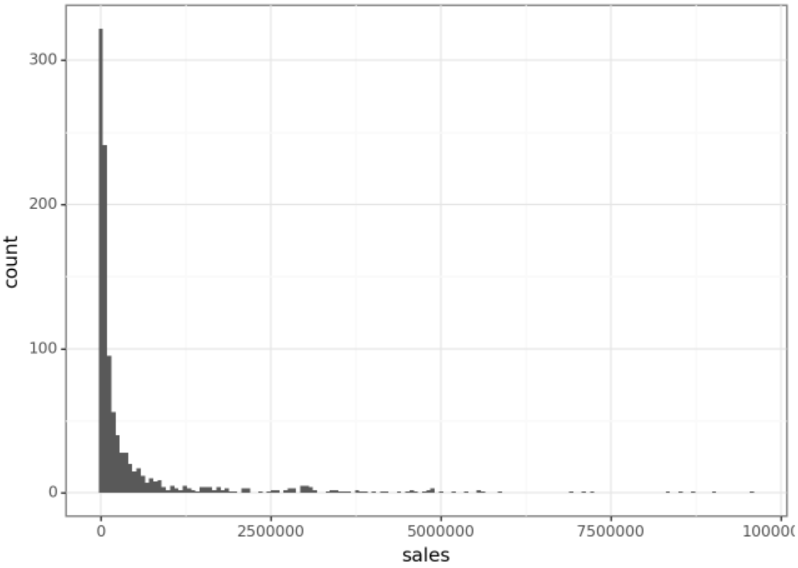


Chart 1: histogram of sales in holdout data

For each training dataset, the missing values were dropped to not dilute the data with imputed values. However, for the holdout set, any missing value was imputed with either the average (for numerical variables) or the mode (for categorical variables) – as before, for all imputations, flags were added. The number of imputations in the holdout set is summarised in table 1.

The dropping of missing values in the training datasets resulted in a loss of 95, 1745, 87 and 2867 observations for datasets 1-4 respectively. Since dataset 5 was created from the already cleaned dataset 2, no further observations were dropped.

Furthermore, for all categorical variables (binary variables were created for all instance) and one was dropped for each variable to serve as a comparison. For industries (ind and ind2 variables) this was industry number 1, for region it was "Central", for gender it was "male", for origin it was "Domestic", and urban it was "1".

As a result, the following datasets were created:

Dataset 1 contains the same industry as the holdout set, only one year earlier. Dataset 2 is from the same year as the holdout set but with all industries in the data except that of the holdout set. Dataset 3 is the same as dataset 1 but with an even stricter restriction on sales figures. Dataset 4 has no lower limit on sales and includes all industries from two years earlier than the holdout set. And finally, dataset5 is identical to dataset2 however to increase the ratio of defaulting firms (to better train the model to be able to identify them), the non-defaulting firms were downsampled, and 30% of them were randomly dropped to create dataset5.

The different datasets are summarised in the table below.

Dataset	**Filters on sales (in Euros)**	**Filters on industry code**	**Year**	**Number of observations**[1]	**Ratio of defaulting firms**
Holdout set	1,000 – 10 million	26	2014	1 037	5.4%
Training set 1	1,000 – 10 million	26	2013	955	4.8%
Training set 2	1,000 – 10 million	All but 26	2014	18 508	10.0%
Training set 3	1,000 – 1 million	26	2013	849	5.2%
Training set 4	Below 10 million	All	2012	19 753	12.6%
Training set 5	1,000 – 10 million[2]	All but 26	2014	13 510	13.7%

Table 2: summary of datasets

Modelling

For modelling, 3 different logistics regressions were created and then a random forest. The details of the logistics regression models are described in the summary report. For all models, a 5-fold cross-validation was done on the training sets to get RMSE and AUC scores. Once the best performing model was chosen, it was estimated on the holdout set and the best threshold was identified based on the loss values of false negatives (15) and false positives (3).

From the different logistics regression models in dataset1 the most complex model (X3) performed the best in terms of cross-validated RMSE and AUC scores, while on dataset2, they performed extremely close to each other, so the simplest model (X1) would be the preferred choice.

In the case of dataset1 the training data is rather small and the prevalence of defaults is rather small (4.8%), in some folds during the cross-validation, there may be no defaulting firms. As a result, the threshold finding algorithm comes back with infinite values for X1 and X2 models as in at least one of the folds, the "optimal" threshold was identified as infinitely large. Depending on the random seed's setting, one can generate instances were all models optimal threshold is calculated, but this was not added to the code to keep a consistent random seed.

Since X1 and X2 performed inferior to X3, the assumption is that their results on the holdout set would be inferior as well.

[1] Number of observations after cleaning of data

[2] The final sales figures' range is different due to the downsampling process.

As for the random forest models' grid search, a tuning grid is defined for maximum features (5, 6, 7) and minimum sample split (11, 16). The criterion used is gini, and 500 trees are created for all models. Lastly, for scoring accuracy, AUC and negative Brier score is used.

A grid-search is performed to find the pest performing set of parameters, but no further effort is spent on tuning as the potential benefits on accuracy are likely outweighed by the time requirements.

The best models contain 5,5,5,7 and 6 maximum features and 11, 16, 11, 11 and 16 minimum sample split for datasets 1-5 respectively.

Similarly to the small sample problem in the cross-validation of the linear regression models, some random forest folds also result in an infinite value for the best threshold. In these cases the infinite value is dropped from the iterated thresholds, and the average value of the remaining (non-infinite numbers) is used as optimal threshold.

Results

In the case of dataset1-2, the random forest models outperformed the logistic regression models. The random forest models in dataset3-5 were all inferior to that of dataset2's results. The model's ROC curve is well above the 45-degree line.

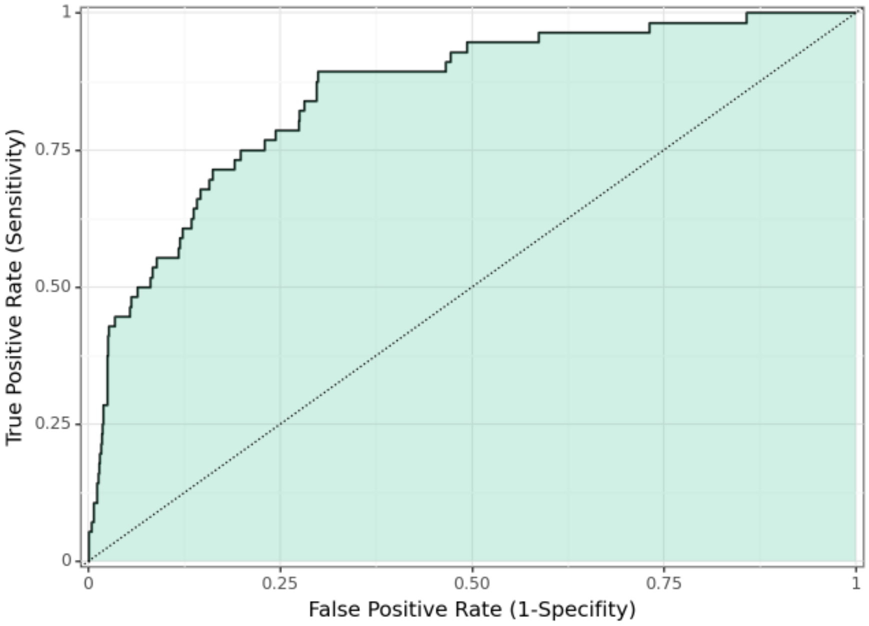
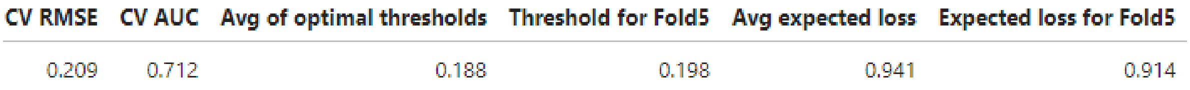


Chart 2: ROC curve of best model



CV RMSE	CV AUC	Avg of optimal thresholds	Threshold for Fold5	Avg expected loss	Expected loss for Fold5
0.209	0.712	0.188	0.198	0.941	0.914

Table 3: results of cross-validation of best model

Data Analysis 3 – Assignment 3
Balint Thaler
GitHub repo: https://github.com/balint-thaler/DA3-Assingment-3

To understand more about the features' effect on the probability of default, the partial dependence plots were drawn based on dataset5's random forest model. The assumption is that although this model performed below that of dataset2's random forest model, the feature importances should be mostly stable across the models and their variation should be rather small.

The detailed interpretation of the plots can be found in the summary report.

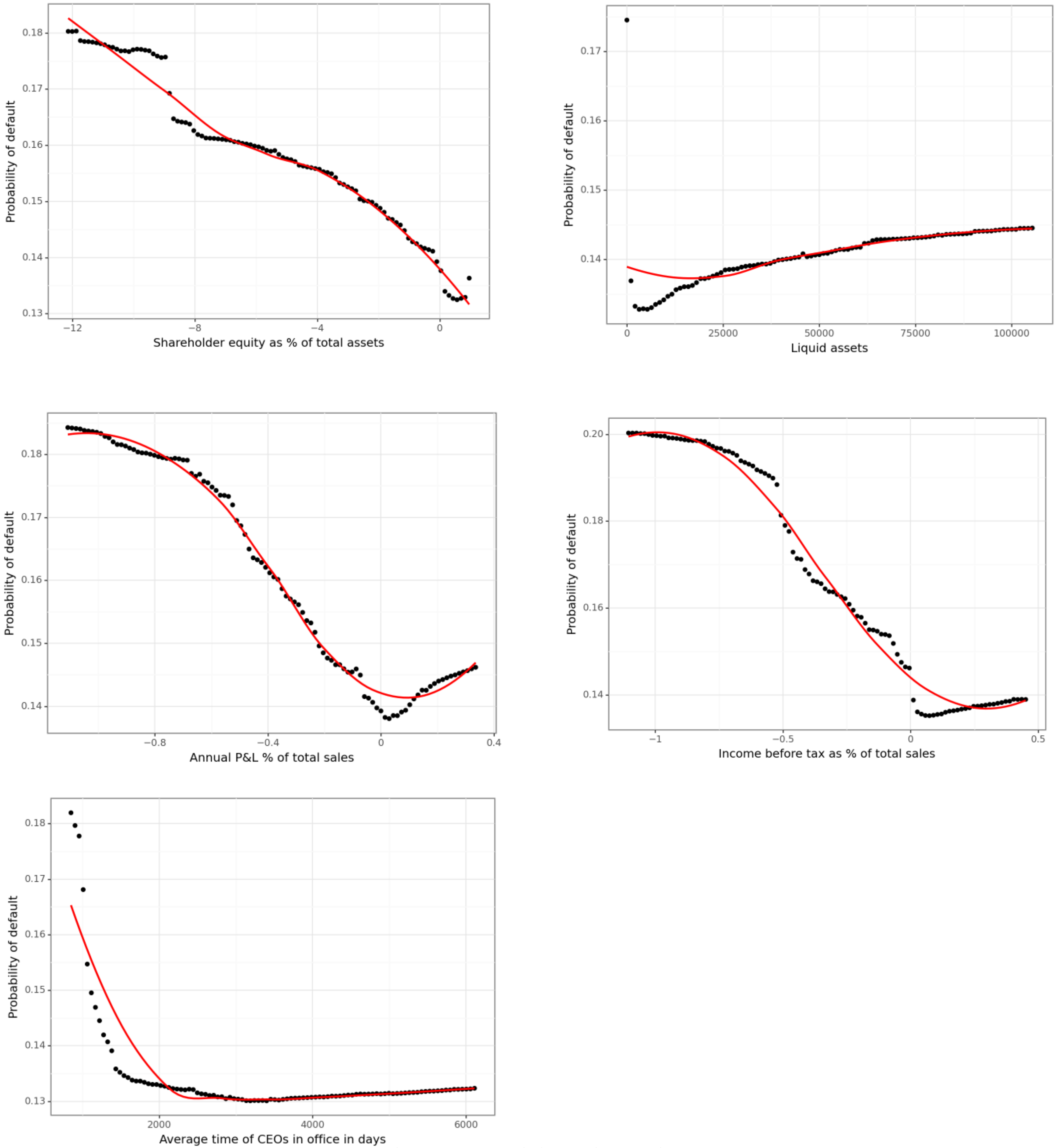


Chart 3: Partial dependence plots