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

Summary report

Introduction

The goal of this report is to summarise the findings of the classification models about the Bisnode dataset. For more technical details and specifics, please refer to the technical summary.

The dataset(s)

The raw dataset contains information on a total of more than 280,000 observations between 2005 and 2016. The different features in the dataset include a number of financials (balance sheet values, profit & loss statement values, HR-related values etc.). The dataset does not contain information on defaulting, therefore this is defined by us as the company existed in year t with non-zero sales, but had 0 or missing value for its sales in t+1 year.

The holdout set is predefined as a set of 1,037 companies from the "Manufacture of computer, electronic and optical products' industry" in 2015. From these companies 56 defaulted, the remaining 981 have "survived". The companies in this dataset have an average sales of 490,902 euros, with a minimum of 1,070 and a maximum of 9.576,485 euros.

For the detailed description of the training datasets, please refer to the technical summary.

Modelling

After feature engineering[1], 3 different logistics regression models are created with increasing complexity. Furthermore, another random forest model is to be created. The first logistics regression only contains the financial figures, model2 also contains a number of categorical variables on location, industry, gender ratios and origin of the company, while model3 contains some HR-related variables also added to the predictors.

All models will be trained on the different datasets, and their performance will be evaluated on the holdout data. For evaluation cross-validated Brier-score (RMSE) and AUC will be used, however ultimately the best model should produce the lowest expected losses on the holdout set as defined by the predefined parameters[2].

Given the computation heavy nature of these calculations and the initial results of dataset 1 and 2, on dataset 3-4-5 only the random forest model will be used, as it produced consistently better results than the logistics regressions.

[1] For details, please refer to the technical report.
[2] The loss values are 15 for false negative and 3 for false positive classifications.

Results

The expected losses on the holdout set are summarised for each model and dataset in table 1 below.

Dataset	Best logistics regression	Random forest
Dataset1	0.737705	0.668274
Dataset2	0.697203	0.619094
Dataset3	N/A	0.624879
Dataset4	N/A	0.653809
Dataset5	N/A	0.656702

Table 1: expected losses on the holdout set

Based on these results, the best performing model is the random forest model trained on dataset2. It is notable that dataset3's random forest also scores fairly close to the best one. The other datasets and the logistics regressions have produced notably higher expected losses.

In the paragraph below, the details of the best performing model will be shown.

Deep-dive into the best model

Dataset2's model has performed with an average expected loss of 0.619 on the holdout set. As seen in the confusion matrix in table 2, the model has successfully identified the vast majority (92%) of non-defaulting firms and more than half of the defaulting ones. The summary statistics of the model's performance on the holdout set can be seen in table 3[3].

The model was trained on a dataset of 18,508 observations. These are firms from all industries with the same sales restrictions as the holdout set but from one year earlier. This resulted in 1,849 defaulters which is 10% (almost two times higher than the holdout set). Firms in this dataset have an average sales of 254,143 euros with a minimum of 1,000 and a maximum of 9,963,926 euros.

	Predicted no default	Predicted default
Actul no defaul	902	79
Actual default	27	29

Table 2: confusion matrix of best model

Metrics on holdout set	Values
Dataset	Dataset2
Avg. expected loss on holdout set	0.619
Optimal threshold	0.237
AUC	0.852
Brier-score/RMSE	0.21
Accuracy	0.898
Sensitivity	0.971
Specificity	0.269

Table 3: results of the best model on the holdout set

[3] For the ROC curve of the model, please refer to the technical summary.

Feature importances

To understand more of the features used in the models, their feature importances were calculated based on dataset5's Random Forest – given that the difference between the models is relatively small, the assumption is that the feature importances should be more or less stable.

Based on this, the most important features are almost exclusively financials with only one non-financial figure making it to the top 20 – which shows the average length of the CEO's time in office.

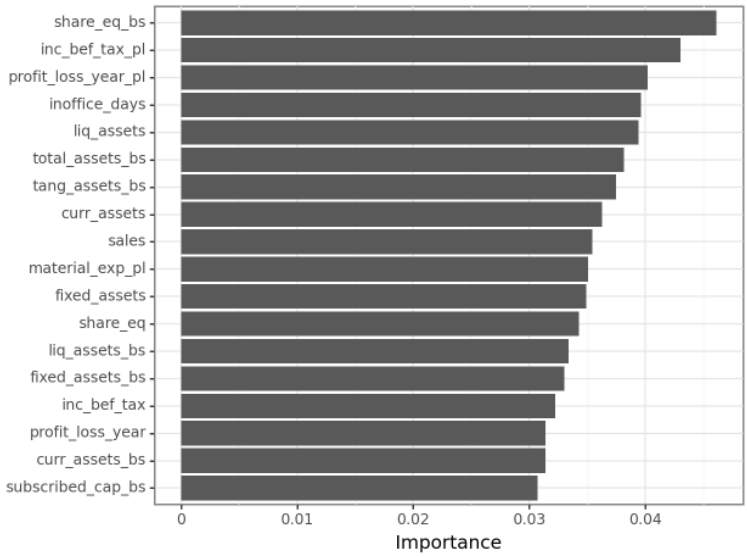

Figure 1: feature importances in the best model

Further to the feature importances, the top variables' pattern of association was also explored to understand more on their relationship with the probability of a default. From these we can observe negative pairwise associations between shareholder equity, income before taxes, annual P&L and the probability of default – these are more or less in line with the intuition. Interestingly, the absolute amount of liquid assets has a positive association (albeit rather weak) with probability of default: having low liquidity (less than 15,000 euros) seems to be associated with lower probability of default.

On top of the financials, the CEO's time in office variable is also observed, which suggests that the probability of default is higher for companies where the average time in position of the CEO is below 1500 days or around 4 years[4].

For the detailed partial dependence plots, please refer to the technical summary.

Conclusions

Overall, the best performing model has an accuracy of 89.8% and is especially successful in identifying non-defaulting firms. Still, only around half of the defaulting companies were classified correctly. The model's most important features are almost exclusively financial figures.

It is noteworthy that the model's external validity is most likely rather limited, as the aim of the exercise was to be precise on a very specific dataset. Other datasets with different industries or time periods are probably behaving somewhat differently and would perform with higher expected losses than the best model.

[4] It is noteworthy that there are very few companies that have CEOs with average time in position of less than 1500 days.