GitHub: https://github.com/mateschieszler/DA3-A3

# Assignment III – Technical Report

## Data Analysis 3

Máté Schieszler, Péter Szilvási

2024.03.03

## Introduction

The goal of this report is to provide a detailed description for the prediction model about the defaulted firms in the ‘Manufacture of computer, electronic and optical products’ industry. The project’s objective is to create the best possible model to predict which SMEs defaulted in the year 2015.

Definition of default: *existed in t-1 year (sales > 0), but does not exist in t+1 (sales is 0 or missing)*

## Introduction of the Dataset

The dataset contains detailed company data of over 45 000 unique firms across 68 different industries in 11 years’ period, between 2005 and 2016, totalling 287 829 observations. Furthermore, the raw data contains 48 variables, featuring several financial indicators about the firms, along with data about the location of production, management demographics and many other firms’ specific factors.

## Data Preparation and Label Engineering

The first problem with the dataset is that once a company goes default in real life, it disappears from the dataset, thus it makes it harder for us to detect which companies defaulted and which did not. To fix this issue, we had to first add all missing year and company id combinations, and only after that create a status alive, and default variable indicating if the firm’s sales in the given year is positive and if it defaulted the next year. Initial filtering includes looking at firms that have an alive status, thus dropping NA values that resulted from the creation of the company id and year combinations. We also drop years 2015 and 2016, filter for our industry (industry code 26), and only look at firms with less than 10 million sales.

## Feature Engineering

Initially, we transformed perhaps our most important variable, sales. First, we replaced negative sales values with 1, and we created the natural logarithm of sales, ln sales. We also created a year-over-year change in the logarithm of sales (in millions), d1 sales mil log with its squared equivalent sales mil log sq. We also identified new companies as those who met the following criteria: age of 1 or less, without a full year of balance sheet data, balsheet notfullyear == 1, change in ln sales is NA (we consider those companies new that were not in our dataset the year before).

One important step in the feature engineering process was to create ratios for balance sheet and profit and loss items. Balance sheet items include intangible assets, current liabilities, fixed assets, liquid assets, current assets, share equity, subscribed capital, and tangible assets. Profit and loss items include extra expenses, extra income, extra profit loss, income before tax, inventories, material expenses, profit loss year, and personal expenses. To create the ratios, all BS items were divided by total assets bs which was created by adding up fixed assets and current assets, and all PL items were divided by sales. The outcome of these divisions is all the above-mentioned engineered variables with their respective _bs or _pl distinction. Lastly, year-over-year changes were added to inventories pl, personnel expenses pl, material expenses pl, income before tax pl, extra income, total assets bs, and intangible assets bs.

We set the datatype of urban m, m region loc, default, year to category, as some tree-based models in scikit-learn utilize the ordered nature of categorical variables.

Other feature engineering steps include adding age$^2$, foreign management, ceo age, and ceo young, extra profit loss pl quad, income before tax pl quad, profit loss year pl quad, share equity bs quad.

During the process, we also created 2 kinds of flags. First to identify high or low values for columns d1 sales mil log, extra profit loss pl, income before tax pl, profit loss year pl, share equity bs. Only high flags were also added for all the remaining pl and bs variables. Lastly flags were added to identify those where is cannot be zero, where they cannot be negative. The second reason to include flags was imputation. We imputed with means for numerical variables and mode for categorical and binary variables. After dropping flags with 0 variation, all together 116 flag columns were created.

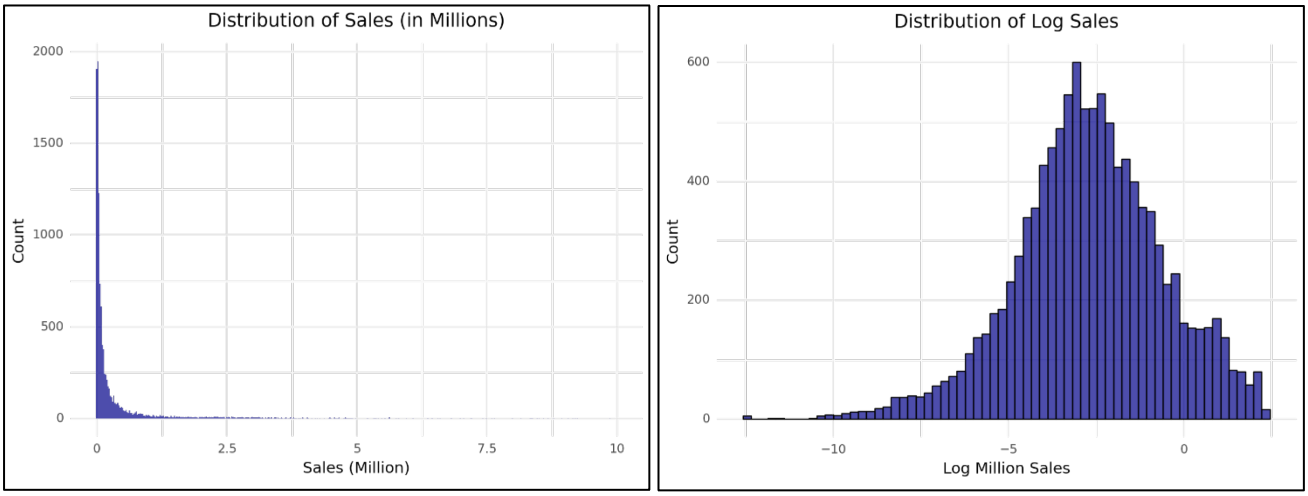
As the last step of the process, we decided to include winsorizing as it limits extreme values in the dataset and mitigates the impact of those outliers. We winsorized current assets, current liabilities, extra expenses, extra income, fixed assets, income before tax, intangible assets, inventories, liquid assets, share equity subscribed capital, balsheet flag, balsheet notfullyear, balsheet length, total assets bs, and labor average to the 90th percentile. Additionally, d1 sales mil log was also winsorized at -1.5 and 1.5 as most values are between these 2 limits.

## Exploratory Data Analysis

In the EDA, we wish to show some of the characteristics of the data, and how we chosen some of our variables later, in the model building process.

The log sales variable was created to fix the distribtuion of the sales variable as it is highly skewed. Figure 1 and 2 clearly show the effects of the transformation. While Table 1 shows the number of firms in the data for each year.

### Figure 1 & 2: Distribution of sales.



### Table 1: Default status on firms for each year

| Year | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | Mean | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Default** | 75 | 65 | 77 | 73 | 96 | 96 | 76 | 88 | 73 | 69 | 79 | 788 |
| **Status Alive** | 1003 | 1035 | 1089 | 1154 | 1191 | 1161 | 1141 | 1131 | 1102 | 1085 | 1109 | 11092 |
| **Ratio** | 7,50% | 6,30% | 7,10% | 6,30% | 8,10% | 8,30% | 6,70% | 7,80% | 6,60% | 6,40% | 7,11% | 7,10% |

GitHub: https://github.com/mateschieszler/DA3-A3

Figure 3.

Effect of Winsorization on Change in Log Sales

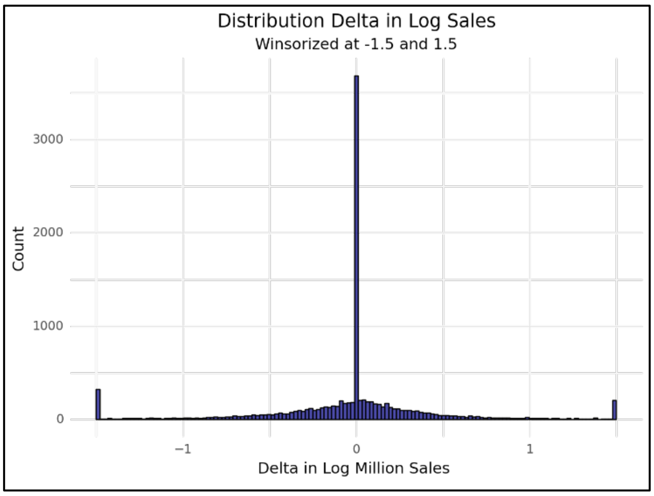
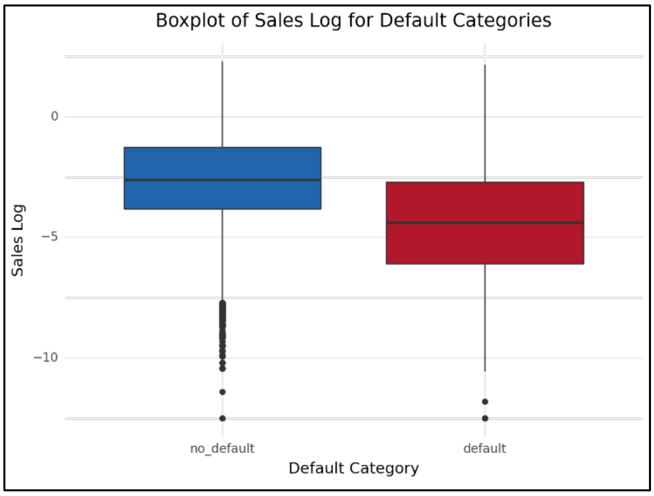


Figure 4.

Boxplot on Defaults and Log Sales



## Modelling

Separation of the Training and Holdout set:

We created the holdout set with the pre given metrics from the project description: only industry 26, in the year 2014, SMEs. The results are 1037 firms, with 56 number of defaulted firms, 981 stayed alive, the average sales of the firms is 0.4902 million EUR, with the minimum of 0.00107 million EUR and maximum of 9.57648 million EUR. We created the training set by dropping the indexes of the holdout from the original sampled data. However, the training set is not only looking at 1 year as cross section, but all years before 2014 with some additional firms who have less than 1000 sales, thus making them not SMEs. We decided to include those as we thought it might help the model detect low sales companies better.

Table 2: Separation of Training and Holdout set

| Table | Rows | Alive | Defaulted | Average Million Sales | Minimum Million Sales | Maximum Million Sales |
|---:|---:|---:|---:|---:|---:|---:|
| **Data** | 11092 | 10304 | 788 | 0,44774 | 0,000004 | 9,97430 |
| **Train** | 10055 | 9323 | 732 | 0,44336 | 0,000004 | 9,97430 |
| **Holdout** | 1037 | 981 | 56 | 0,49020 | 0,001070 | 9,57649 |

**Variable Selection**:

In the variable selection process, we pre-define multiple sets of variables such as raw variables, quality variables, engineered variables, delta variables, hr variables, and more firm specific variables. We are including more delta variables than Gabor's book, to better capture the differences between years. We also control for variable year as we think it might help capture trends over time better.

**Model Building:**

OLS, Logit and Lasso models:

We build five increasingly complex OLS models including more and more variables from our pre-defined variable sets. X1 to X4 without interactions and X5 with interactions. We also run a Lasso on our most complex set, X5. We evaluate our models by comparing their Cross-Validated RMSE and AUC values.

Table 3: Performance of OLS Models

| Model | Number of predictors | CV RMSE | CV AUC |
|---|---|---|---|
| **X1** | 4 | 0,2506 | 0,7184 |
| **X2** | 9 | 0,2498 | 0,7484 |
| **X3** | 36 | 0,2483 | 0,7600 |
| **X4** | 167 | 0,2448 | 0,7865 |
| **X5** | 258 | 0,2668 | 0,6580 |
| **LASSO** | 188 | 0,2461 | 0,7754 |

Out of the five models, X4 performs the best in both evaluation categories. it has an average RMSE of 0.244, and an AUC of 0.786. It is also relatively simple compared to X5 and Lasso with 'only' 167 predictors.

Random Forest Classifier:

For the Random Forest, we carefully pick the variable sets, and try not to include multicollinearity, so we do not include the second and third set of engineered variables. After careful fine tuning of the parameters using Grid Search, the final model is using 6 max features, and 9 minimum sample splits. Compared to the OLS Models, it is performing better with an RMSE of 0.242 and a CV AUC of 0.814 on the training data.

Gradient Boosting Method

To try to find an even better solution, we decided to build a GBM model as our last model of choice. After the same tuning of features, the model has 13 max depth 7 maximum features, and 350 number of estimators.

While the GBM creates a better CV Average Loss on the training set as the Random Forest, both the CV RMSE and the CV AUC lags behind.

Predictions on the training set:

The performance of the models on the training set can be summarized in Table 4.

Table 4: Performance on the training set

| Model | Predictors | CV RMSE | CV AUC | CV threshold | CV expected Loss |
|---|---|---|---|---|---|
| **X4** | 167 | 0,244 | 0,787 | 0,158 | 0,833 |
| **LASSO** | 188 | 0,246 | 0,775 | 0,164 | 0,848 |
| **RF P** | 65 | 0,242 | 0,813 | 0,183 | 0,822 |
| **GBM P** | 65 | 0,260 | 0,792 | 0,001 | 0,817 |

**Predicting on the Holdout Set**

For the final prediction of defaults on the Holdout Set, we use the Random Forest model as it tends to have better measures than the GBM and OLS. When used on the holdout set, the model results in the following Confusion Matrix:

Table 5: Confusion Matrix

| | Actual no default | Actual default |
|---|---|---|
| **Predicted no default** | 927 | 28 |
| **Predicted default** | 54 | 28 |

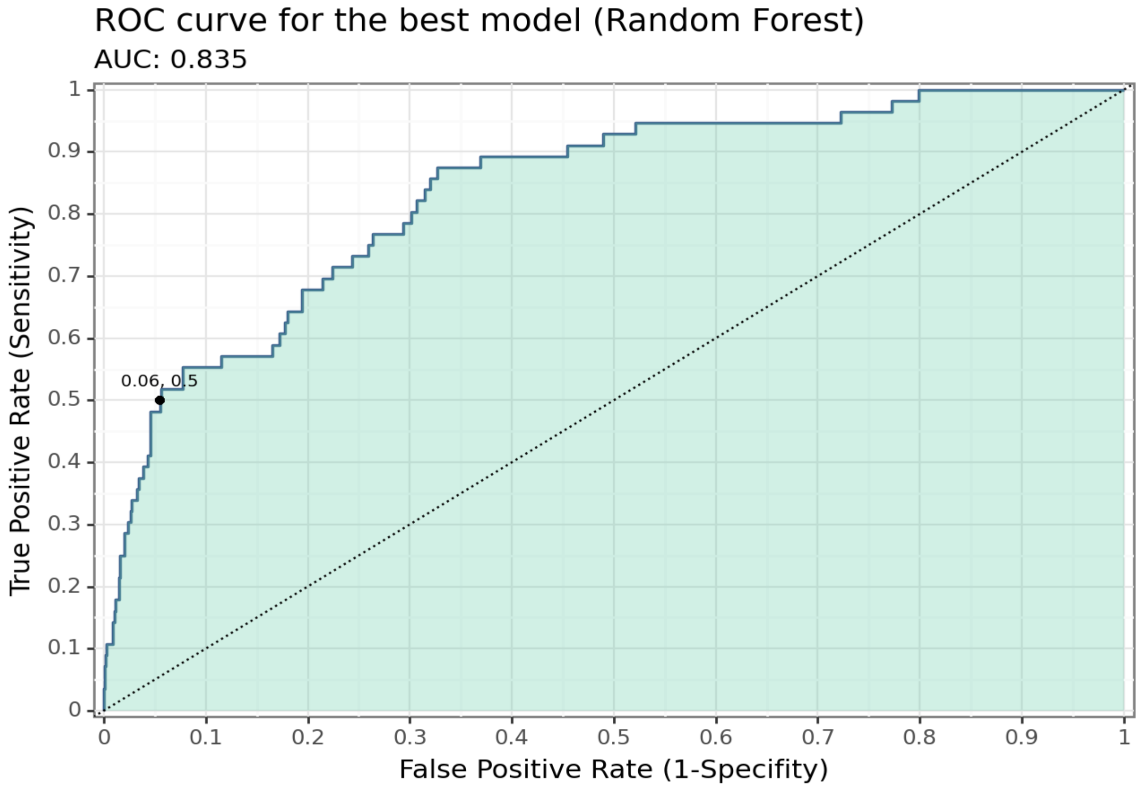
The RF Model predicts 28 firms correctly with default, while it is unable to detect 28 who defaulted as well. However, it is very precise in predicting which firms will not go default. out of the 981 it predicted 927 correctly. This prediction has an expected loss of 0.561 with the parameters defined by the project description, loss (FN) = 15, loss (FP) = 3.

## Summary

The goal of this report was to help better describe how we got to our best model and to its predictions. Four models were illustrated and compared, (OLS, LASSO, RF and GBM), and the Random Forest resulted to be the best model with an outstanding RMSE of 0.209, AUC of 0.835. Feature importances show that sales, log of sales, material expenses, profit loss/ year pl and share equity are the top 5 most influential features of the model.

Finally, Figure 5 shows our best ROC Curve, and Table 6 provides descriptive statistic for the Model.

### Figure 5: ROC Curve



The final loss value on the hold-out is **0.561,** while the hold-out set remains unchanged in terms of the descriptive statistics in the assignment description: The hold-out is 1037 SMEs in 2014 in the chosen industry (‘ind2 = 26’), out of which 56 defaulted and 981 stayed alive, with average sales of 0.4902 million EUR, with the minimum of 0.00107 million EUR and the maximum 9.57648 million EUR.

### Table 6: Final Model’s (Random Forest) Characteristics

| easure | **Brier-Score / RMSE** | **AUC** | **Best Threshold** | **Accuracy** | **Sensitivity** | **Specificity** | **Loss** |
|---|---|---|---|---|---|---|---|
| Value | 0.209 | 0.8353 | 0.1826 | 92.093% | 50% | 94.5% | 0.561 |