Prediction of Defaulted Companies: A Comparison of ML Models

Apo Duman – Yahya Kocakale

1. Introduction

In this project, it is aimed to create the best possible model to predict defaulted companies in the 'Manufacture of computer, electronic and optical products' sector for 2015. Default is specifically defined as companies that were in operation in 2014 but ceased to exist in 2015. Small or medium-sized enterprises (SMEs) with annual sales ranging from €1000 to €10 million will be used in model selection.

Using panel data spanning 2005 to 2016, we aim to identify key factors contributing to firm default in this industry segment. Leveraging advanced modeling techniques, our goal is to develop a predictive model that allows stakeholders to proactively manage financial risks and increase industry stability.

Through rigorous data analysis and model building processes, we aim to provide valuable insight into firm default dynamics, facilitating informed decision-making for risk mitigation and sustainable business practices in a given industry.

We will undertake rigorous data pre-processing, variable selection, and model building processes to create the best predictive model possible. Using advanced techniques such as logistic regression, ensemble methods and cross-validation, we strive to achieve high accuracy and reliability in predicting firm defaults, thus facilitating proactive risk management practices.

2. Exploratory Data Analysis, Label Engineering and Sample Design

In this project, raw data was taken from datasets of Gabors Data Analysis book. We filter data with industry (Manufacture of computer, electronic and optical products), year (2014), and annual sales (1000 Euros to 10 million Euros). Then, dummy variables are created for firm characteristics and financial variables. Then we focus on two key aspects: normalizing profit and loss (P&L) items by sales and normalizing balance sheet (BS) items by total assets. We define two lists, pl_names and bs_names, which contain the names of profit/loss and balance sheet variables, respectively. These variables will be used for subsequent calculations. For each profit/loss item in the pl_names list, we divide the corresponding column values by the "sales" column. This normalization facilitates comparisons of these financial metrics relative to the company's revenue. Similarly, for each balance sheet item in the bs_names list, we divide the corresponding column values by the "total_assets_bs" column. This normalization enables us to analyze the composition of the balance sheet relative to the total assets of the company. In cases where the "total_assets_bs" column contains NaN values, indicating missing or undefined total assets, we set the normalized balance sheet item values to NaN as well. This ensures consistency in the data treatment. Additionally, we create flags to identify specific characteristics of financial variables. For instance, we classify variables as "flag_high" if they exceed a certain threshold, "flag_low" if they fall below a threshold, and "flag_error" if they are negative. These flags aid in identifying and categorizing anomalies or outliers in the data.

After that, we perform some additional data processing steps, including CEO age calculation, handling of labor-related variables, and conversion of categorical variables.

Then, several new variables are generated based on the growth rate of sales (d1_sales_mil_log). Also, flags are created to identify instances where the growth rate exceeds certain thresholds. Additionally, the growth rate variable is modified to limit extreme values, a process known as "winsorization".

3. Modelling

In this section, after defining the sets of variables like main firm variables, further financial variables, flag variables, growth variable, human capital related variables, firm history related variables and interactions, we run multiple logit models, logit with LASSO and a Random Forest model.

3.1 Simple Logit Models*:* These setups define specific combinations of variables for different model specifications, ranging from simple logistic regression models to more complex setups incorporating various sets of predictors and interactions.

3.2 Logit + Lasso: This set includes variables intended for logistic regression models with LASSO regularization, which helps in feature selection and model regularization.

To predict probabilities with logistic regression and LASSO using 5-fold cross-validation, we follow these steps:

- We specify a 5-fold cross-validation method to evaluate the models' performance.
- We define a list of variable sets for different logistic models (M1 to M5).
- For each model specification, we fit logistic regression models using cross-validation with the specified parameters.
- We calculate the root mean squared error (RMSE) on the test set for each fold of cross-validation.
- And finally, we apply the best logit model to predict defaults in holdout data.

Since we need to select our holdout data from 2014 and predict their default status in 2015, we trained our models with 2014 data (excluding specified holdout data).

3.3 Random Forest: As a final step, we employ a random forest classification to predict defaults. We will compare this model with previous Logit and Lasso models. We created a random forest classifier to predict probabilities. We perform hyperparameter tuning using grid search with cross-validation and evaluate the model's performance using metrics such as ROC AUC and RMSE.

Compared to logit and lass models, random forest gives better results both on training and on holdout data. While M2 model gives an expected loss of 0.686 (least among logit models) on holdout data, Random Forest gives is 0.61.

Table 1: Selected Model (Random Forest) Performance

Metric	Value
Brier Score	0.042922
AUC	0.857288
Accuracy	0.885246
Sensitivity	0.589286
Specificity	0.902141
Expected Loss (Optimal Threshold)	0.610415
Optimal Threshold	0.187648
Number of Firms	1037.000000
Number of Defaulted Firms (Actual)	56.000000
Number of Defaulted Firms (Predicted)	33.000000
Number of Firms Stayed Alive (Actual)	981.000000
Number of Firms Stayed Alive (Predicted)	885.000000
Mean Sales (million EUR)	0.490202
Minimum Sales (million EUR)	0.001070
Maximum Sales (million EUR)	9.576485

The summary table for our selected model shows that we predict 33 defaults out of 56 actual defaults. The expected loss value is 0.61 for our model.

The confusion matrix for our selected model summarizes the true and false predictions (see Table 2).

Table 2: Confusion Matrix for Holdout Data (Random Forest Model)

	Predicted No Default	Predicted Default
Actual No Default	885	96
Actual Default	23	33

4. Conclusion

In this analysis, we explored the use of random forest for predicting defaulted firms in the 'Manufacture of computer, electronic and optical products' industry for the year 2015. Through the development and evaluation of decision trees and random forests, we identified optimal parameters and evaluated model performance using cross-validation. The selected Random Forest model demonstrated promising results in terms of RMSE, AUC, and expected loss. This predictive model can serve as a valuable tool for stakeholders in proactively managing financial risks and facilitating informed decision-making within the specified industry segment.