Prediction Models For EU Firms

Using EU Firm Data to Predict whether EU Firms will Default

Nicolas Fernandez and Arbash Malik

Introduction

Data from the OSF website [1]contained information about European Union firms from 2005-2016 in three industries: auto manufacturing, equipment manufacturing, and hotels and restaurants. It was consutrcted from publicly available sources by Bisnode for educational purposes. From this data, predictive models were constructed for the purposes of accurately predicting when a firm defaults based on their financial information and reporting as well as other data available including personnel data. For the purposes of creating a model, the selected industry was the manufacturing of computer, electronic, and optical products, modelling based on small and medium sized enterprises, defined as companies that had sales between 1000 and 10,000,000 euros during 2014 within that industry. The code for this analysis can be found on github[2][3].

Data Decisions

As mentioned above, the selected data was that for which we were targeting to create predictive models on as a holdout set. This gave us n=1037 observations from a total of n=287,829 observations in the dataset. Since the industry we were looking at was only one specific industry, a decision was made to take the remaining data and filter for that industry as well. Along with this, the firms we were targeting to predict defaulted firms for was in 2014 meaning any data in or after 2014 in the main dataset would not be as relevant for predicting therefore the data was further filtered for data from 2013 and prior. Lastly, we wanted to target the same size of firm in order to construct our models so we used the same sales restrictions for filtering. This left us with a training set of n=9689 observations from which to construct our models.

From here, we created several engineered variables based on available data such as a total assets balance sheet, profit/loss ratios, and liquid asset balance sheet data points. We also constructed time lag variables within the data to compare firms year-to-year within the data to capture any trends within the data, positive or negative. We also examined the data for any values that seemed to be an error or incorrect value such as negative asset values and corrected them by setting them to 0 along with creating flag variables for whenever such actions were taken.

After examining and collecting the variables which were deemed to be important for analysis, we examined them further to check for null values within the data. Columns for which null values were the minority the mean was imputed in order to maintain the statistical mean and limit bias when estimating other variables. For binary variables with a minority amount of null values the mode was selected for the same reasoning. For columns where the majority of the values were null they were dropped from the data rather than imputed.

A 'default' column was created by characterizing a firm whose value in the 'status_alive' (a binary variable) equaled 1 but then was equal to zero in a subsequent year as 1 for having defaulted, 0 otherwise. This gave us the following descriptive statistics within training set and the holdout set ,respectively:

```
default
0    9091
1     598
Name: count, dtype: int64
```

```
default
0    981
1     56
Name: count, dtype: int64
```

[1] https://osf.io/b2ft9/?view_only=

[2] https://github.com/nxfern/DA3_Assignment_3

[3] https://github.com/arbash-malik/DA3/tree/main/Assignment%203

Models

Five models were created to compare them to each other and provide the best predictions. 5 Fold Cross Validation was used on the training set for training the models. All models used the 'default' variable as the predicted value with 'sales_mil' (a modification of the 'sales' variable divided by 1 million) used as the main explanatory X variable.

M1

A simple OLS model was created by hand which accounted for variables attributed to sales, quality of the firm, and profit/loss.

M2

A more complex OLS model created by hand which comprised of the same variables as M1 along with balance sheet variables, and basic interaction terms with sales.

Logit-LASSO

A Logit-LASSO model was created that used all the variables related to sales, profit loss, quality of the firm, balance sheet, the time-lapse variables created, the engineered variables, and the numeric characteristic variables related to firms such as firm age and allowed for the LASSO algorithm to shortlist which variables were the most important.

Random Forest

A Random Forest model was created that used the variables that were shortlisted from the LASSO model defined above. This was done in order to use only the terms that were deemed to be significant by LASSO.

Gradient Boosting Macihne

A Gradient Boosting Machine (GBM) model was constructed that also used the variables shortlisted from the LASSO model for the same reason as before.

Results and Conclusion

The models were run using Classification regressions given that the predicted variable was not continuous but binary with the target goal of the model to be the model with the lowest expected loss values. Amongst the logit models (M1, M2, and LASSO) the LASSO model performed the best regarding the training set when evaluating their average expected loss. The Random Forest model, however, performed even better than the logit models in this regard, as shown in the table below:

	Number of Coefficients	CV RMSE	CV AUC	CV treshold	CV expected Loss
M1	16.00000	0.23661	0.70373	0.11251	0.85303
M2	37.00000	0.34289	0.66335	0.41384	0.88461
LASSO	38.00000	0.23181	0.75826	0.16532	0.79544
RF	38.00000	0.23010	0.77680	0.21614	0.76045

The expected loss values are, however, still high overall. The GBM model was also run and gave the following values:

	CV RMSE	CV AUC	Avg of optimal thresholds	Threshold for Fold5	Avg expected loss	Expected loss for Fold5
0	0.23863	0.74423	0.15789	0.44301	0.79606	0.84409

The average expected loss was higher than the Random Forest and LASSO models but better than either of the theoretical OLS models.

These models were then run with the holdout set in order to compare their expected losses on the holdout set in order to test whether or not they are good predictive models. The results can be seen below:

	Model	Expected Loss on Holdout
0	LASSO	2.00800
1	RF	0.56410
2	GBM	0.67700

The lower the expected loss, the better the predictive value. In this regard we can see that the Random Forest model vastly outperforms the LASSO model and also does a much better job of predicting defaults correctly compared to the GBM model given our loss function parameters of penalizing a false negative by 15 and a false positive by 3. Below is a table showing the confusion matrix of the Random Forest model on the holdout set:

	Predicted no default	Predicted default
Actual no default	946	35
Actual default	32	24

As we can see, the Random Forest model does a very good job of correctly predicting that a firm will not default while minimizing the most out of any model the number of firms that are predicted to not default that actually do default, which is what we are penalizing the most with our loss function. In conclusion, the Random Forest model does the best job of predicting firms to default based on past data out of the models that were created.