Data Analysis 3: Assignment 3: Business Report

Mohammad Ahmed

Jo Kudo

Git hub repository: https://github.com/Jk33033/DA3_hw3/tree/main

Introduction

In this report, the main task is to build the best possible model to predict defaulted firms in the 'Manufacture of computer, electronic and optical products' industry, 2015. The prediction is only for small or medium enterprise (SME) with certain situation; they existed in 2014 but did not exists in 2015. Holdout dataset with that firms are tested in view of how this prediction is well-modeled.

Data Preparation

The dataset was gained from the OSF website with the following link:

https://osf.io/b2ft9/?view_only=

The dataset nemed "cs_bisnode_panel" is used to make prediction. A company-year long format xt panel data table, 2005-2016 . The number of observations is 287,829, which includes 46,412 firms.

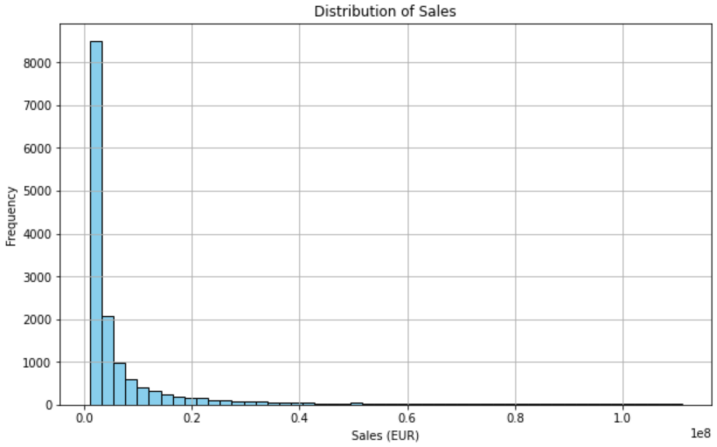
Feature Engineering

The original data has many variables the basic characteristcs such as company ID, year, sales of a certain year, and the other specific columns such as exit year, inventories, started year etc. A few columns were dropped from the dataset after we went through the data transformation and understanding phase because they either had too many NULL values or we believed that the variables won't be useful in our analysis. The best practice while building a good model is to not have any unwanted variables and to make sure there's no memory being wasted on things that aren't being used. COGS column didn't have much in it so that was dropped and even the others. Either they weren't being used so no point of keeping them. For gender, origin, and region, the mode was filled in the missing value bacause they are uncountable value. On the contrary, the other countable variables are filled with the mean value when it is missing. In terms of making a holdout dataset, it is limited to a selected industry, which means ind2 == 26, and the only companies with sales between 1000 and 10,000,000 EUR in 2014. In this report, the definition of defaulted companies is the firms which had positive sales in 2014, but not in 2015. So dummy variable is made for whether the companies are

defaulted or not based on that information. At the end, the total number of the holdout data was 1037 and the number of defaulted firms and live firms are 56 and 981 respectedly, which is the same as the known numbers in instruction.

Exploratory Data Analysis

In order to understand the dataset dealt with in this report, the histgram of the cleaned dataset is made. According to this graph, the distribution is quite left skewed; the number of high sales immediiately drop down. Most of the sales are below 10,000,000 EUR, so it seems to be valid to make a limitaion for the holdoutset to be under that score to make sure there is less outliers.



Modelling

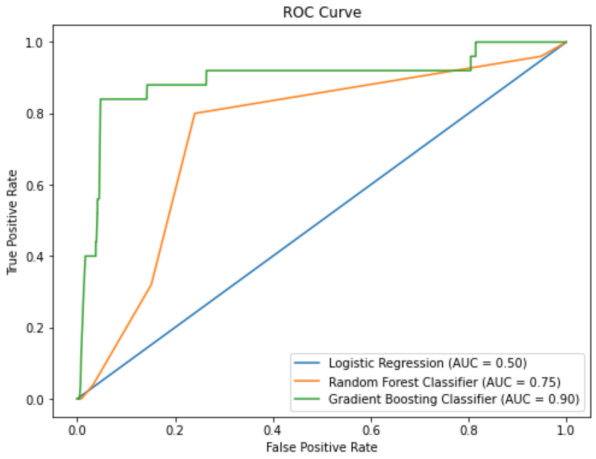
Before making models, the dataset without holdout data is normalised. After that, three models are made: logistics regression, random forest, gradient boosting. In each modelling, scikit learn library was mainly used to make a fair comparison of the indicator related to the quality of prediction.

Evaluating the Models

With the models above, following measures are obtained:

- Brier-score, ROC curve, AUC
- Accuracy, sensitivity, specificity (for optimal threshold)
- Expected loss and optimal threshold

For brier score, Logistic Regression, Random Forest Classifier and Gradient Boosting Classifier got 0.00242, 0.00264, and 0.00252 respectively. The ROC curve and AUC is shown in the graph below.



Besides, Accuracy, sensitivity, specificity for optimal threshold are calculated like this.

```
Optimal Threshold and Metrics:
Logistic Regression: Threshold = 0.05 , Accuracy = 0.9975777540935956 , Sensitivity = 0.0 , Specificity = 1.0
Random Forest Classifier: Threshold = 0.35000000000000003 , Accuracy = 0.9975777540935956 , Sensitivity = 0.0 , Specificity = 1.0
Gradient Boosting Classifier: Threshold = 1.0 , Accuracy = 0.9975777540935956 , Sensitivity = 0.0 , Specificity = 1.0
```

Expected loss of the three models is **375** by chance.

Diagnostics

In terms of brier scores, they are indicative of all three models performing well, with Logistic Regression slightly outperforming the others. However, the differences in scores are minimal, suggesting that each model provides a high level of precision in its predictions. In simpler terms, the Gradient Boosting Classifier was the best at telling the difference between two groups, better than the Random Forest and much better than the Logistic Regression. This is shown by how much its ROC curve bends towards the top-left corner, meaning it has a higher chance of correctly identifying the true cases. Logistic Regression was the least accurate, as its curve was closer to a straight line, showing it struggles more to tell the groups apart. In the other indicator such as accuracy, sennsitivity, specificity, and even expected loss, the large differences are not seen, which means the quality of thes prediction can be compared mainly based on ROC curve or AUC this time.

Conclusion

For the Brier scores, Logistic Regression does a slightly better job than the others. The scores are all very close, though, which means each model is pretty accurate. For telling two groups apart, Gradient Boosting Classifier is the best. It's better than Random Forest and a lot better than Logistic Regression because the curve for Gradient Boosting goes more towards the top-left corner of the graph, which tells us it's good at picking out the right cases. Overall, it is possible to say that Gradient Boosting made the best contribution to prediction in this report.