Group Fine Mining: Executive Summary

Research Questions and Objectives

This project explores the usage of data mining algorithms to help reduce the risk of default in online lending platforms. Such platforms are much more susceptible to defaulters because of the lack of controls and regulation that their traditional counterparts, such as banks, must abide by. Hence, the project's primary research question is to answer if data mining methods and techniques can be used to help lending platforms reduce the number of defaulting loans. Because of the wide variety of lending platforms, this project focuses its scope on Bondora, a lending platform based in the European Union.

With that in mind, our objective is to train a classification model on existing data that will predict whether or not the user applying will default based on their application information. We explore a variety of classification methods to determine which will provide the most optimal predictive performance. Our hope in doing so is to allow platforms to leverage data mining to build a more trustworthy service in which lenders are exposed to less risk, while also improving the company's profitability.

Importance of Project

Bondora is by no means a small platform. Considering it has issued over the equivalent of about €425M ($500M) in loans since its inception, it is important for a large platform like them to address defaulting effectively. This is further highlighted if we consider that from their dataset, it was found that about 60% of loans issued on the platform have gone into default. However, it is important to mention that Bondora doesn't write off a loan if it goes into default; Bondora just proceeds with legal action to attempt to recover the principal. Despite this, the issue is not helped by the fact that the recovery rate, the extent to which the principal can be recovered following default, has shown a declining trend since 2013. For that reason, it is important to identify defaults to maintain the integrity of the platform.

Data and Methodology

On its website, Bondora offers a publicly-available dataset containing information about every loan approved on the platform. The dataset is rather large with over 160,000 records that each possess up to 112 attributes. Thus, several steps needed to be taken to reduce the dataset's dimensionality. First was that only predictors related to information given at the time of the application's submission were kept; any predictors associated with the operational status of the loan after its approval were removed. Then, attributes like IDs and dates were removed. Next was to consider EU legislation such as the General Data Protection Regulation Act, which limits the sort of demographic information Bondora can ask of customers, and Council Directive 2004/113/EC, which forbids gender discrimination in access to financial services. Finally, after reviewing related literature and performing exploratory analysis, we reduced the dataset to 12 predictors. We then sampled the dataset and retrieved 15% of it, resulting in 16,931 observations in total.

Regarding methodology, we determined the Naive rule was that all records would be classified as 'default' as 60% of records fell under that category. The models applied to the dataset were built from the *k*-Nearest Neighbors, Naive Bayes Classifier, Classification Trees, and Random Forest classification algorithms. Also, the data was partitioned into two subsets (70% training and 30% testing) for fitting and predicting, respectively. Then, to optimize the models' performance, randomized search and grid search were used to tune each model's settings. Lastly, evaluating the performance of the models involved the use of an ROC curve and confusion matrix to calculate accuracy, precision, sensitivity (% of class 1 - default correctly classified), and specificity (% of class 0 - non-default correctly classified).

Analysis and Results

Before implementing KNN, Naive Bayes, Classification Tree (CART), and Random Forest, we saw that the non-default records had 7,104 records and the defaults had 10,245 records in the dataset. To combat this problem, we had to use the SMOTE Technique on our two models. This way, both the non-default and default group would have the same number of records (10,245) and the model can perform equally.

Out of the four models, the Naive Bayes model performed the worse. Even though the default detection for both the train and test data were 100%, the non-default detection accuracy was only 22% and 23%. Though this would be helpful for preventing defaults, it would result in too many false positives and turn down many legitimate applicants, ultimately harming profitability.

The KNN, CART, and Random Forest models were much more stable and performed better. Their results for the metrics were in clustered around the range of 60%-79% for both train and test data. The KNN model had an accuracy of 77% for the training data and then 65% for the test data. The defaults and non-defaults correctly classified by KNN for the training data was 82% and 74%, respectively, while the testing data was 70% and 60%, respectively.

CART scored an accuracy of 68% during training and 67% during validation. It detected defaults correctly 70% of the time during training and 69% during validation. Then, for detecting non-defaulting loans, it was 66% accurate in training and validation.

Finally, Random Forest had an accuracy of 73% during training, which fell to 69% in validation. In terms of default detection, it scored 72% in training and 67% during classification. For non-defaults, it correctly detected them 74% of the time in training and 70% in validation.

An ROC curve was then used to compare each of the four models, and found that Random Forest had the greatest area under the curve (AUC) at 0.76. Classification trees comes second with an AUC of 0.74, KNN with 0.71, and Naïve Bayes in last with 0.61.

Overall, based on these results, we can see that the Random Forest model was the best performing model. This may be attributed to its ensemble nature in that its performance reflects the aggregation of multiple models rather than a single model like the other methods.

Findings and Conclusion

The use of classification models, particularly the Random Forest model, with the ability to detect 67% of defaults, proves promising as a solution to help manage and reduce classification risks. The business implications of reducing defaults not only reduces operating losses from unrecoverable principal, but also prevents valuable company resources from being wasted on initiating legal action as required by the default recovery process. Additionally, it improves the legitimacy of the platform as a lending service, making it more appealing for potential investors.

Nonetheless, the model’s accuracy is still far from perfect and can be improved. In addition to evaluating other classification methods, a variety of future steps should be explored to address defaulting. One is testing a dataset selected on time period to see if it helps account for the state of the economy (e.g., a dataset derived from a recession period v. that derived from a growth period). Another step is to employ parametric models like logistic regression that can help determine which variables most significant in determining default. Additionally, a strategy of building individual models for each country that Bondora services rather than for the platform as a whole should be tested, as doing so would allow the dataset to include each country’s respective credit score system. Ultimately, this project has shown that data mining methodologies can be leveraged by lending platforms to prevent defaults.