Abstract

Name: Annas Imtiaz

Student Number: 501203680

Supervisor's Name:

Submission Date: May 15, 2023

The banking / financing industry's life depends on accurate credit decisions. It is practically impossible to achieve 100% accuracy, but the more accurate credit decisions are, the more profitable regular operations will be. Therefore, there is always the urge to improve accuracy of credit decisions. The challenge particularly lies when the bank wants to grow its customer base. In such cases, banks usually have to rely on their subjective expert decision making.

The theme chosen is on similar grounds. I have chosen a dataset which includes demographic characters, as well as selected data points from previous credit experience of more than 1000 borrowers. This dataset is a combination of both regular and defaulted customers. Since all the borrowers included in this dataset have prior credit history; therefore, their credit files can be assumed to be thick and not thin.

The data being used has a combination of numeric and categorical factors for a total of 9579 credit card users (https://www.kaggle.com/datasets/urstrulyvikas/lending-club-loan-data-analysis), with the dependent variable being a dichotomous default outcome. The data has a relatively good representation of both cases (default: 1533, non-default:8045).

The research problem, as described above, is to evaluate creditworthiness of those individuals who are new to the bank / FI and apply for a loan / financing facility. Using the selected dataset for this project, the outcome will be a machine learning model that will learn from characteristics of borrowers to be able to discern between good and bad credit, and then apply this learning to predict whether a new borrower will like default or not.

The research questions I will be answering through this research are:

- Which is the classification algorithm that produces the best results in terms of accuracy and area under the curve?
- Are any of the independent variables correlated, leading to potential over-fitting if one of them is not excluded from analysis?
- Do the chosen algorithms have the same level of accuracy on both the train and test dataset?

Since the dependent variable is dichotomous in nature, classification machine learning algorithms will be used for this case. Specifically, decision trees, random forest as well as logistic regression classifiers will be used to train the model, and then test it on unseen data. The ideal model will be the one with the highest accuracy and area under the curve.

While these three models will be used and their performance compared, the de facto industry standard is to use the logistic regression classifier. The reason is that FIs do not want just a yes / no decision. In fact, they are looking for a ranking of borrowers to determine what level of risk they are willing to take. This is achievable only through logistic regression classifier, which generates coefficients to determine how important each of the independent variables are, and in turn allows conversion of the output into probability of default and credit scores.