CSC240 Proposal

Yiwei Han
Will Yang

March 22, 2023

1 Dataset Choice

The proposal is to analyze the dataset: ”bank-full.csv (Ensemble Techniques)” available on ”Kaggle.com.” The reason that we chose this dataset is it aligns with our background. I am also a student majoring in Business Analytics. It is interesting to utilize abilities learned in computer science classes to solve problems in the field of business and provide more insights. It is true that not all loans made are perfect investments of the bank. Perhaps that model generated from data mining would give insights about what loans to whom are likely to yield profit and not lost to the bank. This model would reduce risk and loss for loan decision-makers.

When bankers considerations about a loan application, they need to be extra careful about the potential risks of the applicants. The loan consideration is based on many reasons: credit score, income, the purpose of the loan, etc. However, purely manual consideration can somethings influenced by emotions. If a customer failed to pay back loans, the bank is unable to recover the full amount of the loan, resulting in losses for the bank. These losses can impact the bank’s financial stability and ability to expand and compete with other banks in the future. Our model will contribute to providing a decision-making process in a bank’s lending consideration.

2 Team Composition

Yiwei Han and Will Yang will be working together on the final project. Yiwei Han is a full-time undergraduate student pursuing a B.S. in Computer Science and B.S. in Business (Business Analytics). Will Yang is a full-time undergraduate student pursuing a B.S. in Biology and B.A. in Computer Science.

3 Goal of the Analysis

We plan to use supervised learning to develop models to make predictions on whether the bank should give a loan to a specific customer in the dataset. We would like to discover various factors that affect the bank’s decisions to give customers a loan. Metrics like MSE and accuracy score will measure the performance of each model. The results can have insights into the business analytics and risk management team of a bank, as they need to value whether to give a customer loan based on the attributes described in the dataset and their potential ability to pay back. Our model will be very useful for them to reduce the risk of lending out loans.

4 Planned Technical Approach

4.1 Data Visualization

This dataset contains 17 attributes, with one target attribute that determines whether a customer can get a loan, and other attributes are numerical, ordinal, binary, and categorical data types. The optimal way to get to know data with many data types is to use visualization like different distribution charts and confusion matrices. We could see the distribution as well as the importance of each attribute. Proper data visualization gives us insights into the further steps of the project.

4.2 Data Preprocessing

Data Pre-processing includes several parts. First, we need to see whether there are missing values for the attributes. Looking through the dataset, there are no missing values in the dataset. Second, we could use a confusion matrix to see which attributes are more important and select useful attributes. Thirdly, for numerical attributes, we could normalize the data; for binary attributes, we could give it 0 or 1; for categorical variables, we could use the dummy variables to process. In conclusion, data pre-processing is a step that will clean the dataset to make our model for machine learning algorithms to perform more accurately and efficiently.

4.3 Model Selection

The baseline model we will be using is the Naive Bayes classifier and random forest classifier. Those are the algorithm we learned from class. Naive Bayes can be implemented quickly and efficiently, making it suitable for large-scale datasets. A random forest classifier is a robust algorithm that can handle outliers and noisy data without overfitting. It is also good at handling imbalanced datasets.

Moreover, by reviewing the dataset, we found out that the dataset experiences oversampling and that there are a lot of imbalances in the target (whether a customer gets a loan). We realized that the majority of customers do not get a loan. In order to solve the problem, we would like to use SMOTE to fix the oversampling and compare the result with the original dataset's result.

The advanced model we would be using is the feed-forward neural network model. It consists of an input layer, some hidden layers, and an output layer. The input data is fed through the network in a forward direction, with each neuron in the hidden layers transforming the input data. At last, the output layer produces the final prediction. Feedforward neural network is able to learn nonlinear relationships and are able to generalize well to new, unseen data. This means that they can make accurate predictions even on data that they have not been trained on.

5 Role of Team Members

Yiwei Han and Will Yang will be performing all the steps outlined above, delivering the final presentation in class, and writing the final report.