CS412 Machine Learning Project Report

Samuel Lee, Onat Kaya

12/29/2019

Introduction

The main purpose of the project was "Predicting Loan Payment," where we tried to determine whether an individual will repay the loan from the bank given some basic information about the individual. After data analysis and preprocessing, the final decision for the model of application was MLP Neural Network classifier.

Preprocessing

In order to choose the best machine learning model, our group carried out extensive data exploration and wrangling. First of all, we studied all features that were available in the dataset. Some of the data was basic, yet informative, including information like loan amount, annual income and home ownership. However, there were some parts of the dataset that were highly specific and difficult to interpret for application to the model, including "Months since oldest revolving account opened" and "Ratio of total current balance to high credit/credit limit for all bankcard accounts."

To further understand these highly specific datasets, our group found the NaN values within the dataset, to see which information was not available to us. We found the number of total NaN values for each feature, and also found the ratio between the number of datasets available and number of NaN values. When doing this, there was an extensive list of features that had a range of one thousand to over ten thousand NaN values within the given training dataset.

Our group determined that it may be better to give some kind of value to the machine learning model rather than NaN values. Through some research, there were roughly 3-4 methods on how to replace NaN values. We could either replace all NaN values with some constant number(e.g. -1), replace the NaN with the value of the previous value, fill it with either mean or median value of that feature, or fill it with random number that range from the standard deviation

subtracted from the mean. For simplicity and efficiency, we used the third method of replacing NaN values with the mean.

For the main part of the preprocessing, we found all the non-integer/float value features including object and string data types, and gave a numerical value. For most of the features including home ownership, grade, term etc... we replaced all unique values with unique integers. At the same time, we dropped certain features including 'emp_title' and 'earliest_cr_line,' establishing that these features will not have a significant impact on the prediction.

Model Selection

First of all, we were able to eliminate some of the machine learning models during the beginning data analysis phase, including KNN due to the high number of features/dimensionality. With a large number of training data, we decided to avoid decision tree classifiers primarily because it was prone to overfitting. With the presence of high number of features and relatively variable noises, our group finally determined that neural networks will be the best machine learning model for this problem.

After checking the documentation of the MLP neural network classifier, we used "adam" stochastic gradient-based optimizer for the solver, as it works well on relatively large datasets. At the training stage, k-fold cross validation was used with the aim of improving the model's accuracy.

As for the layers, we initially used both default value of 100 and a value lower than default of 20 value for the hidden layers, but this resulted in low accuracy, so we have decided to pick 20 for optimal speed and accuracy. All other parameters were kept to default.

Results

For the final result, our model gave the macro F1 score of 0.79984, which could be considered as a decent score when compared with other competitors' results.