Capstone Project **(Proposal)** Ray Liu

Machine Learning ND July 10, 2017

Predict Defaults on Credit Card Payments

Domain Background:

Credit cards incur debt when improperly used. Credit cards were invented as a means for customers to spend money that they do not currently have but will pay back in the future. Banks are usually the suppliers of credit cards since they are the ones with large reserves of cash. To make money off of credit cards, banks charge high interest rates for when the consumer fails to pay their debt within a month. To prevent debt from accumulating, consumers must realize where the limit is for their spending's and to stop using their credit card when they have unpaid debt.

Many credit card users overspend and exceed their limit since they do not know how to stop themselves from excessively spending. This happens because credit cards provide users with a false belief that they have money to spend when in reality they do not. To prevent defaults on accumulated credit card debt, it's best to identify the possibility of a default happening early on to prevent the credit card user from continually spending when they cannot pay. This way, the consumer does not accumulate a large pile of debt in which they cannot pay and the bank does not end up having to pursue debt that cannot be paid.

Problem Statement:

Defaults on credit card payments are a problem for both the consumer and the credit card company. Defaults happen as a result of accumulated debt over time with either continually missed payments or meeting only minimal payments every month. To solve this problem, we must find features of the consumer that will help predict whether a consumer, in the future, will default on their credit card payment. Quantifiable features include salary, credit score, interest rate, etc. Both the consumer and the credit card company suffer as a result of a defaults, therefore it's helpful to prevent a default from happening.

Datasets and Inputs:

The dataset I am using is from Kaggle. The dataset is a sample of credit card clients in Taiwan from 2005 with one of the features being a binary feature; either 0 or 1 to indicate a default. Upon evaluation of the dataset, there are 6,635 out of 30,000 users in this dataset that have defaulted on their credit card payments; which means approximately 22.12% of credit card users default on their payments. The dataset's features are as follows: maximum credit line, gender, education level, marriage status, age, and history of payments. By eyeball, it's most likely that the maximum credit line and the education level are features which have impact on determining whether a credit card default will happen.

The maximum credit line feature determines the amount of given credit to a consumer in NT dollars. Gender is an input feature in case there exists a pattern for whether males or females are more likely to default on credit card payments. Education level has 6 inputs, 1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown. In theory, credit card user with a higher education level is more likely to have self-control when it comes to spending in comparison to another user with pre-university education. Marriage status is another feature because there is a possibility that whether the credit card user is single or married has an impact on the outcome of unpaid credit card payments. Age is also of importance since young adults are more likely to spend lavishly despite lacking money. The final important feature is history of payments; attempting to determine whether a credit card user will default early based on the history of payments is a goal. An example of the importance for payment histories happens when there is a pattern for defaults where the credit card user only pays the minimum amount for the bill a few months prior to the default.

I plan to first evaluate the features provided to me by determining which sets of features provide the strongest relationship to defaulting. After determining the strongest features, I will then proceed to use those features to make a model that can predict whether a credit card user will default based on the features provided.

Solution Statement:

With such a large dataset, finding a relationship between the features of the data and whether a consumer defaulted on their credit card is feasible. After finding the relationship between what features provide the most impact on whether a consumer defaults, we then create a model to determine whether a future consumer will default on their credit card payment. The solution of the first problem shall inform us of which features are the most important for determining whether a consumer defaults on their payments while the solution for the second problem shall be a model for predicting whether a consumer will default in the future.

Benchmark Model:

The results obtained from the model we make will be evaluated on the validation set first and then the test set to determine the true accuracy. Since this is a classification problem, the benchmark model that will be used is SVM. I choose SVM to be the benchmark model for this problem because SVM is very flexible for classification problems. SVM can be used with many kernels such as linear, polynomial, rbf, and sigmoid. Given such flexibility, I will obtain a brief idea for how the customer data is distributed.

Evaluation Metrics:

We will be using the accuracy score as the evaluation metric for both the benchmark model and the solution model since approximately 22% of the dataset is default payments. The accuracy score will be calculated as follows: (TP+TN) / (Total) where TP = true positive, TN = true negative, and Total = the total number of guesses. In layman's terms, we take all the correct predictions from the model and divide it by the total number of samples we provide the model.

Project Design:

To tackle this problem, I intend to first evaluate the dataset. Downloading the dataset and briefly scanning through the dataset to ensure there is no missing data. Once the dataset is okay by the eye, I plan to do data preprocessing by removing unnecessary columns such as the customer ID. Since there are no categorical variables I don't need to do one hot encoding. I will then proceed to scale the data so that the range for all the features are evenly distributed. For example gender, education, marriage, and age are all values that don't exceed 100 whereas credit limit and bill statements can reach the thousands. Once the data has been scaled, I shall perform some type of unsupervised learning to determine which features have the strongest relationship to determine whether a credit card user will default. One unsupervised learning method I have in mind is PCA. After determining which features are the most important for this problem, I then proceed to perform supervised learning to create a model that can predict whether a user will default on their credit card payment based on certain features. One supervised learning algorithm isn't optimal for this problem since there exists so many algorithms. I plan to handpick a few algorithms and compare the accuracy score for all of them to better determine which algorithm is the best for this problem. Some supervised learning algorithms I have in mind is random forest, SVM, and Adaboost. The accuracy score will be computed by using the model to predict the validation set. The model with the highest accuracy score in the validation set will be used for computing the accuracy score on the test set. The accuracy score of the test set will be the true accuracy of my model.