**The Busines Problem and Interested Agents**

Credit card companies and peer to peer lenders are keen on lending money to people who will not default. Invariably, some will default. The challenge to lenders is identifying who will default while also not discriminating along race, gender, or other factors. Without discriminating there are some factors that one could consider that will lead to determining whether or not the one receiving a loan will default. Namely, lenders could look at the age, previous defaults, levels of current debt, incomes, assets and liabilities.

**The Data**

In this notebook I utilize a data set of a group of Americans from 2005-2015. I received this data from Credit Bureau Data, Inc ([www.cbdlaxl.com](http://www.cbdlaxl.com)) It represents three attributes: age, income measured in USD, and loan amount also measured in USD. I drop the “Identification” at the beginning as this attribute is not helpful. The data can also be found on my GitHub Repository for this project at <https://github.com/MTM2020/IBM-Data-Science-Capstone-Project/blob/master/credit%20data%20set.csv.zip> .

**Methodology**

First, I describe the data to better understand the variance and, inter-quartile and and loan deficiency. This provided a snapshot into the dataset, by nothing really concrete. In order to account for the “NaN’ values, I replaced then with the mean for that particular attribute.

Then I did some pretty basic data cleaning my dropping the Client ID and changing age to an integer. Next I split the data into a independent variables: income, age, and loan and dependent variable (default) as my “Y” variable. Then I split the train and test set. First I initiated K Nearest Neighbor (N=4) I also printed the mean square error (0.147) and its accuracy 85.20. These were the two inferential statistics that I used.

**Machine Learnings Used**

Next, I employed a confusion matrix to better understand toward which class the model is biased. Next I use the K Nearest Neighbor Classifier to determine the nearest training examples in the feature space. This is a non-parametric methodology for classification and regression.

When I executed the confusion it was clear the model was biased towards the customer not defaulting. So after normalizing the data the accuracy score of the mean square error test increased to 98.00 and the mean square error was 0.020. This is a drastic increase. Further, the confusion matrix significantly reduces it’s Type I and Type II errors.

**Results and Conclusions**

The KNN Model should be used to determine who will default because it minimizes ones Type I and Type II errors. Further research could be done to determine an ordinal degree of default risk (i.e. low, medium, and high) risk, but that is outside of the scope of this paper.