**Name: Taha Shahid**

**Project: Capstone Project 1: Milestone Report**

**Problem statement: Why it’s a useful question to answer and for whom.**

The data set selected is called “Default of credit card clients Data Set” and as the name suggest we predict which classification method/ data mining technique will give the best accuracy for the probability of default of credit card clients.

Our client will be banks in Taiwan because banks will be utilizing the analysis and machine learning models to better understand their clientele and will be able to make decisions on their customers which will benefit them from customers that will most probably default. The dataset is already provided by the UCI machine learning repository and will be acquired through downloading XLS format of the dataset from the website.

To analyze this problem, different machine learning algorithms such as Decision Trees, SVM, ANN and Naïve Bayes will be used to see which method has the best accuracy for the probability of default of credit card clients. Also, PCA (Principal component analysis) will be used to see if dimensionality can be reduced. Any or all methods learned in the machine learning algorithms will also be applied.

For the deliverables of the project, an IPYTHON notebook code will be provided along with the paper summarizing the findings as well as a PowerPoint presentation.

**Description of the dataset, how you obtained, cleaned, and wrangled it**

**What kind of cleaning steps did you perform?**

The default of credit cards dataset was downloaded in excel(csv) format from the University of California, Irvine. Firstly, the dataset had an extra first row for the variables that was creating problems for calling the columns of the dataset. To eliminate the first row the following code was used to get rid of the first row:

dfs = pd.read\_csv("C:/Users/Taha/Downloads/default.csv",skiprows=1)

Secondly, the decision variable, which is binary in the dataset i.e. if the customer is a defaulter or not had a very long name “default payment next month” and was renamed to just “default” for sake of calling it with ease using the following code:

dfs=dfs.rename(index=str, columns={"default payment next month": "default"})

**How did you deal with missing values, if any?**

No missing values were found in the dataset using simple code shown below to see if each row had integer value or not.

dfs.info()

**Were there outliers, and how did you handle them?**

Yes, there were outliers and they were determined using the criteria that was mentioned with the dataset as follows

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows:

X6 = the repayment status in September 2005;

X7 = the repayment status in August 2005; . . .;

X11 = the repayment status in April 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

The outliers for example in education column were given a 5 or a 6 and there is no information given on what type of education it represented. So, anything other than 1,2,3,4 was given 5. (Not sure if it’s the correct way to clean a dataset but should makes values a bit more consistent and cleaner for the education column as compared to the information provided with the dataset.).

The outliers in repayment status were given -2, which didn't make sense at first but carefully looking at the repayments data of a single row, it was established that it meant that the person paid a month before the time of payment for the bill. So, nothing was changed in the repayment columns because it was established that some people pay a month before the due date.

**Initial findings from exploratory analysis (get this from your data story and inferential statistics reports)**

**a. Summary of findings**

Some of the questions that were investigated from the surface of the above-mentioned dataset are:

**Is the proportion of defaults the same for men and women?**

It was found out that Proportion of male defaulters is 24.16 % and Proportion of female defaulters is 20.77. Also, total number of males in data was 11888 and total number of females in data was 18112.

**Does Education level matter for predicting default or not default?**

It was found out that the highest numbers of defaulters were university students with 23.74%, second highest were High School student with 23.45%, graduate students had a defaulter proportion of 19.24% and all others has the lowest default proportion of 5.7%.

**Does credit card limit have a connection with being a defaulter or not?**

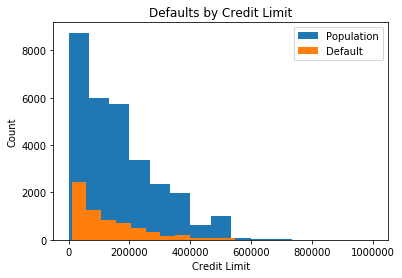
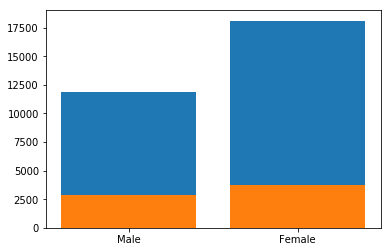
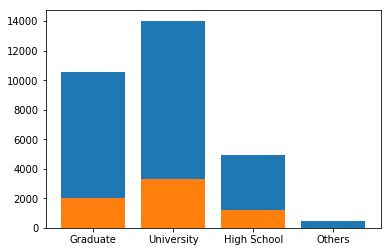
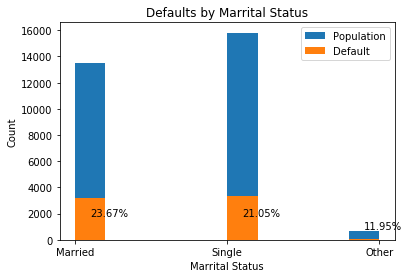
Proportion of defaulters from 0 limit balance to 200000: 25.34%. Proportion of defaulters from 200000 limit balance to 400000: 15.48%. Proportion of defaulters from 400000 limit balance to 600000: 12.28%. Proportion of defaulters from 600000 limit balance to 800000: 7.69%. So, it was found out that with higher credit limit balance there were fewer number of defaulters.

**Does Marital Status have anything to do with being a defaulter or not?**

Married defaulters were the highest at 23.67 %. Single defaulters were 21.05 %. And all others were 11.95 %.

So, from above four questions it was found out that generally men had a higher percentage of default as compared to women. Secondly, students in university and high school had some of the highest percentages when it came to their accounts being defaulted. Thirdly, credit card limit balance almost had a direct correlation with the account being defaulted, as the limit balance increased lower the number of defaulters. Lastly, married and single people had a pretty close default proportion but married people for some reason had a higher proportion of default. The above finding will be investigated further using exploratory data analysis. To get a visual of the above questions investigated below are the four graphs.

**b. Visuals and statistics to support findings**

********