**Capstone Project**

**CREDIT CARD DEFAULT PREDICTION**

**Shajad**

**Contents:**

* **Abstract**
* **Introduction**
* **Problem Statement**
* **Steps Involved**
* **Data Visualization**
* **Methods and algorithm that we used**
* **Conclusion**
* **References**

**Abstract:** In this project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable then the binary result of classification credible or not credible clints. We can use the K-S chart to evaluate which customers will default on their credit card payments

**Data Description**

Credit card default historical data and information about the customer are following Most of the fields are self-explanatory. these points are:-

* **Amount 🡪** Credit amount
* **Gender 🡪** Gender(1: Male , 2: female , 3:other)
* **Education 🡪** Educational qualifications (1: Graduateschool, 2: university, 3:high school, 4: others )
* **Marital status 🡪** Married or not(1: Married, 2: Unmarried , 3: other
* **Age 🡪** Age
* **History of past payments 🡪** Payment made in past
* **Amount of bill statement 🡪** bill amount
* **Amount of previous payments 🡪** Previous amount paid(1: paid, 2: not paid)

**Introduction:**

we are provided with historical credit card default data of the customers. The task is to forecast the "classifying and predict the default customers" column for the test set.

**Problem Statement:**

This data set contains information of customers that information includes such as Amount, Pay 0,1,2,3 etc. and sex of the customers education Qualification of the customers All personally identifying information has been removed from the data.

Explore and analyze the data to discover important factors that affect the Prediction of credit card defaulter.

We will tackle the problem statement in the following steps:

Step 1: Data Overview.

Step 2: Data Preparation and Cleaning.

Step 3: Visualizing the Univariate and Bivariate features.

Step 4: Correlation Analysis.

Step 5: Concluding Analysis.

**Steps Involved:**

* **Importing Packages:** Importing the various libraries that will help us analyze our dataset properly with visual graphs.
* **Data Overview :** We load the data and go over the basic features, shape and datatypes of various variables.
* **Data Preparation and Cleaning:** We use various features of python to create combined features of date and time and other variables which can be simplified. We also drop columns that we used to create the new columns. Finally we take care of outliers.
* **Univariate and Bivariate Analysis:** We use seaborn and matplotlib to plot graphs starting with one variable graphs and then plotting two or more variable graphs to understand the variables and it spread and range with its frequency
* **Concluding Analysis:** Here after plotting and analyzing all the data we finally make predictions and remarks about our analyzes.

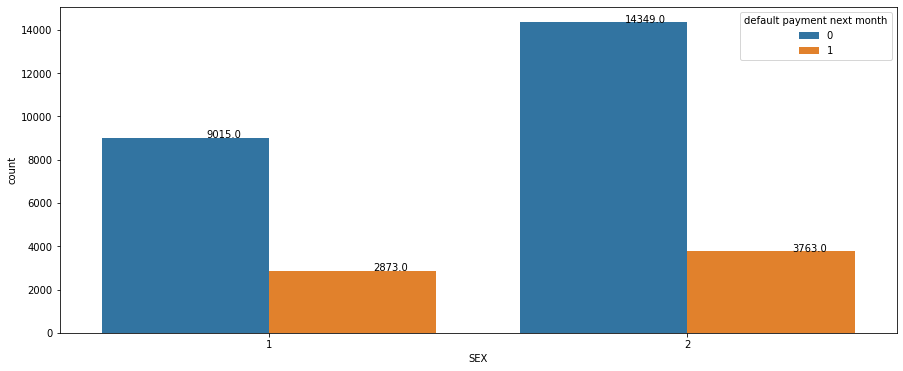
**Data Visualization:**

According to this chart there are 9015 male has no default but 2873 male customers are defaulted.

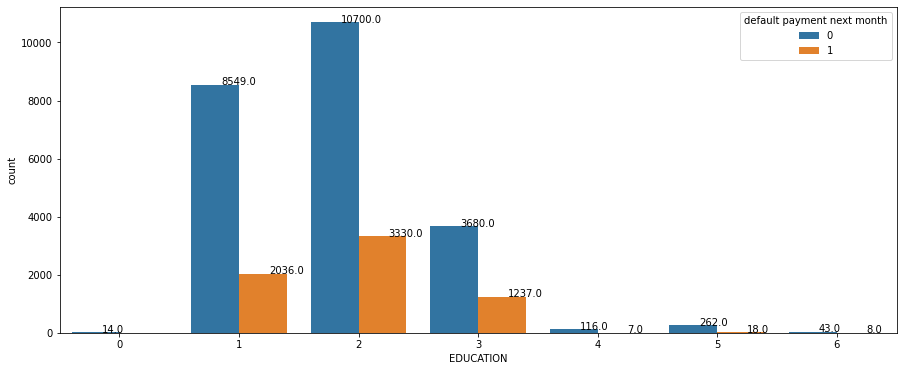
And 14349 female customers are not default but only 3763 female customer are defaulted.

So it means female customers has less default then the male customers.

**There are 34% male are defaulter but 21% women customers are defaulter so women are more safest customer then men**

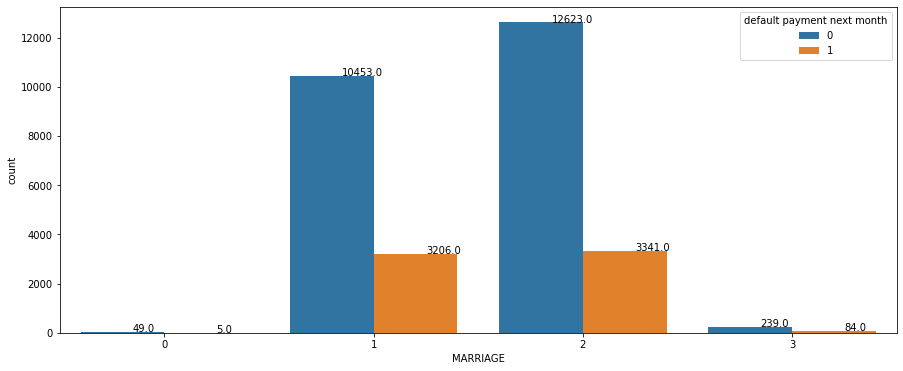


* from the Graduate school There are 8549 customers are not defaulted but 2036 customers are defaulted.  
  From the University there are 10700 customers are not defaulted but 3330 customers are defaulted.  
  Fromm the High school there are 3680 customers are not defaulted and 1237 customers are defaulted.  
  and other we include them in outliers.  
  **from graduate school have least defaulter that is 19% then University comes with 24% defaulters then High school comes with 25% defaulters**

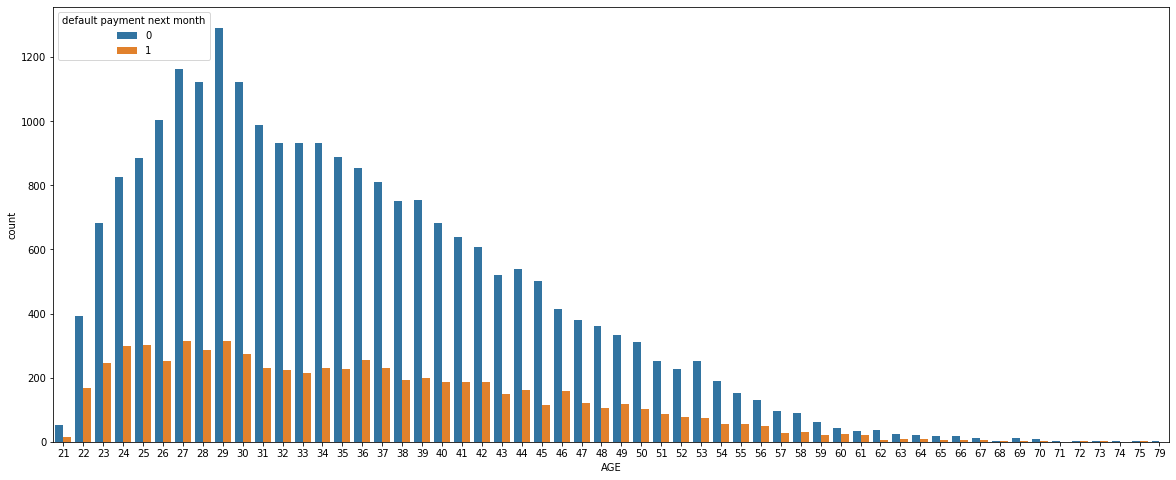


From the 1 there are 10453 married customers are not default and 3206 customers are defaulted.

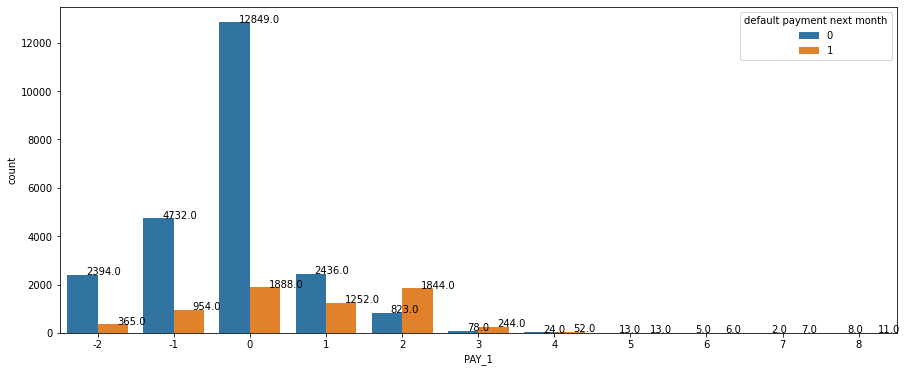
* From the single there are 12623 single customers are not default but 334 single customers are defaulted.
* **in this plot single customers are least defaulter with 21% then married customers are comes with 23.5% defaulters**



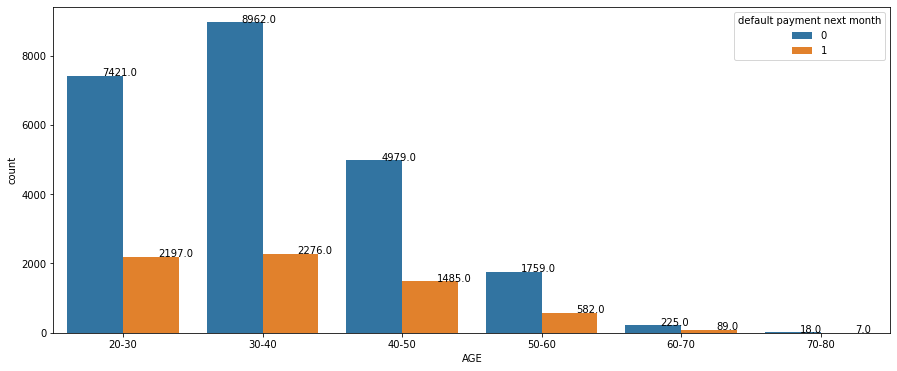
* There are negative relationship between age credit card distribution increase in age that means the credit card holders is decrease.



* from -2(pay duly) there are 2394 customers are not default but 365 customer are default from -1(pay duly) there are 4732 customers are not default but 954 customers are defaulted. from 0(not duly) there are 12849 customers are not defaulted but 1888 customers are defaulted. from1(payment delay 1 month) there are 2436 customers are not defaulted but 1252 customers are defaulted. from 2 (payments delay 2 month)there are 823 customer are not defaulted but 1844 customer are defaulted. from 3(payment delay 3 month ) there are 78 customers are not defaulted but 244 customers are default. and so on.....



* according to this chart
* from the age group 20-30 There are 7421 customers are not defaulters but 2197 customers are defaulters.
* from the age group 30-40 there are 9862 customers are not defaulters but 2276 customer are defaulters.
* from age group 40-50 there are 4979 customers are non defaulter but 1485 customer are defaulters.
* from age group50-60 there are 1759 customer are non defaulters but 582 customers are defaulter.
* from the age group of 60-70 there are 225 customer are non default but 89 customer arre also defaulters.
* **according to this chart from age 60-70 the default percent 29% present in this group**



**Conclusion:**

This project is about the Credit card default prediction analysis. When we start this project it is very difficult to understand and look very complex.

In first step we upload the data and then we start to find the duplicate values and nulls values and missing values etc. but in this data there is no missing values and duplicate values. After completion of 1st step in 2nd step we plot the data in form of graph, pie plot and bar plot and line plot and also box plot, and in box plot we find some outlier but these outliers are not affect the result because thy are very less in number. After plot the data we plot the heatmap of the data and then we find the relationship between the feature and we did some feature engineering in the data. We remove the labels of the data because that labels are not readable and difficult to understand and we make the data’s 1st row as the labels of the data.

After the feature engineering we apply some algorithm, 1)Logistic Regression , 2)Random forest ,3)XG Boost and we also use the dummy model to describe the our model and make them easy to under stand. And we also do the some cross validation in every algorithm.

After applying these algorithm we select the Random Forest as a best algorithm that gives the best score. It gives the precision 0.668 and recall 0.364 and F1 score 0.471 percent. After selecting the algorithm we do some feature engineering and do some cross validation but there is not much change in the result. Logistic Regression model has the highest recall but the lowest precision, if the business cares recall the most, then this model is the best candidate. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. Since Random Forest has slightly lower recall but much higher precision than Logistic Regression, I would recommend Random Forest.

**References:**

* GeeksforGeeks
* Stackoverflow
* Almabetter
* Youtube
* Github