Credit Card Defaulting Project's Report :

Loading the Data and preparing it:

- After downloading the data from Kaggle I used Pandas to load it into a data frame.. after that i did data where it does have record for Amount of the bill(BILL_AMT) and how much the was the customer payment on that bill (PAY_AMT) was following indexing from 1 but the payment status it start from 0 and jump to 2.. so i rename the column to 1 to match the patten... i think it was just a mistake while collecting the data maybe.
- Changing the column values in gender to be 1 or 0 instead of 1 or 2

Visualizing and Understanding the data:

- It's really important to understand the data we are dealing with..from looking into the data details like the mean, max, min. and standard deviation will give a good understanding for the data.
- Using Math plot library to visual the data and see what feature could have a good effect more than other on our target (Defaulted/ Not Defaulted) also did use a small method to print the percentage of the default based on specific columns..for example in gender the Male tend to get defaulted more than human and specially single male
- Creating a heat map to make the understanding the correlation between data much simpler

Feature Engineering Opportunity :

- After looking into the data there was Opportunities to create some feature if needed..the strong correlation between BILL_AMT column make it easy to fit them under one feature by maybe looking on average or the six column.
- There was a correlation between PAY column but not strong as the previous one..

Choose Algorithm :

- it was clearly a Supervised Machine Learning case. I decide to use Sklearn library because it easy and powerful. I choose at first 4 algorithms and they all did good. the algorithm was (Random Forest, Gradient Boosting. Ada Boost and Decision Tree) all get a good score..

Train the model and get prediction with 2 months data:

- decide to not use the data of last month (the 6th month) so i can get prediction base on the 5 month and accuracy did not get effected much and that due the correlation between the column that i mention before.. so i start Random Forest then chance to Ada Boosting because the result was slightly better after dropping column... i was able to get accuracy up to 82% with only using information from the first two months