Title: Credit Card default Prediction

Bhargava Gaggainpali  
DSC680 - Applied Data Science  
Project 2 - Milestone 2

[bgaggainpali/bgaggainpali\_DSC680 (github.com)](https://github.com/bgaggainpali/bgaggainpali_DSC680)

**Introduction**

In financial industry, banks are playing important role in challenging times like now, with COVID pandemic across the globe. People are losing jobs and financial institutions are facing more defaulters rate on credit card loans. The increase in defaulters’ rate will result in significant financial loss to commercial banks. It is very critical for lending institutions like banks to have a prediction model to be able to predict customers for credit card default.

**Any surprises from your domain from these data?**

Credit Card Default is part of Financial industry. I have selected the topic, as I was interested in knowing the factors of Credit Card Default and the key factors in it. As I explore more about the domain, I understand that its not same set of rules which is being followed across domain. And each category like Credit Card Default, Home Loan Default, Car Loan Default and such classification have different factors which are influencing for people falling default. Bank should not lose business by having more customers as defaulters and should not hurt the good customer, who can pay the credit card payments every month. Also, it should be able to identify the customers who are at risk of falling Credit Card Defaulters and provide them other options like paying the bill amount in installments. We can say that Banks are playing important role in challenging times like now, with COVID pandemic across the globe.

**The dataset is what you thought it was?**

Initially when I looked at the dataset, I had questions about the variables as they are more of general kind and was guessing that if would serve my purpose of analysis. when closely observed the stats, it surprised me as the amount of value we can retrieve from such data. I am satisfied with the dataset which I have selected.

I have identified UCI\_Credit\_Card.csv as source for my work, below is the Kaggle link. There are 30,000 observations in the dataset, each row in the dataset represents a credit card client. Given is the list of variables in the dataset.

Source File: <https://www.kaggle.com/ainslie/credit-card-default-prediction-analysis>

**Variable Description**

ID Credit Card ID - Sequence Number

LIMIT\_BAL Credit Limit

SEX 1 = male, 2 = female

EDUCATION 1 = graduate school, 2 = university, 3 = high school

MARRIAGE 1 = married, 2 = single, 3 = others

AGE Customer Age

PAY\_0 Repayment status September 2005

PAY\_2 Repayment status August 2005

PAY\_3 Repayment status July 2005

PAY\_4 Repayment status June 2005

PAY\_5 Repayment status May 2005

PAY\_6 Repayment status April 2005

BILL\_AMT1 Bill Amount September 2005

BILL\_AMT2 Bill Amount August 2005

BILL\_AMT3 Bill Amount July 2005

BILL\_AMT4 Bill Amount June 2005

BILL\_AMT5 Bill Amount May 2005

BILL\_AMT6 Bill Amount April 2005

PAY\_AMT1 Payment Amount September 2005

PAY\_AMT2 Payment Amount August 2005

PAY\_AMT3 Payment Amount July 2005

PAY\_AMT4 Payment Amount June 2005

PAY\_AMT5 Payment Amount May 2005

PAY\_AMT6 Payment Amount April 2005

default.payment.next.month 1 = default, 0 = On time payment

Below are the initial observations made from stats and plots:

1. Customers with low LIMIT\_BAL have higher Default rate.

2. Default rate low among Female (Sex=2)

3. Customers with highly educated are less like to default (EDUCATION=1 or 2)

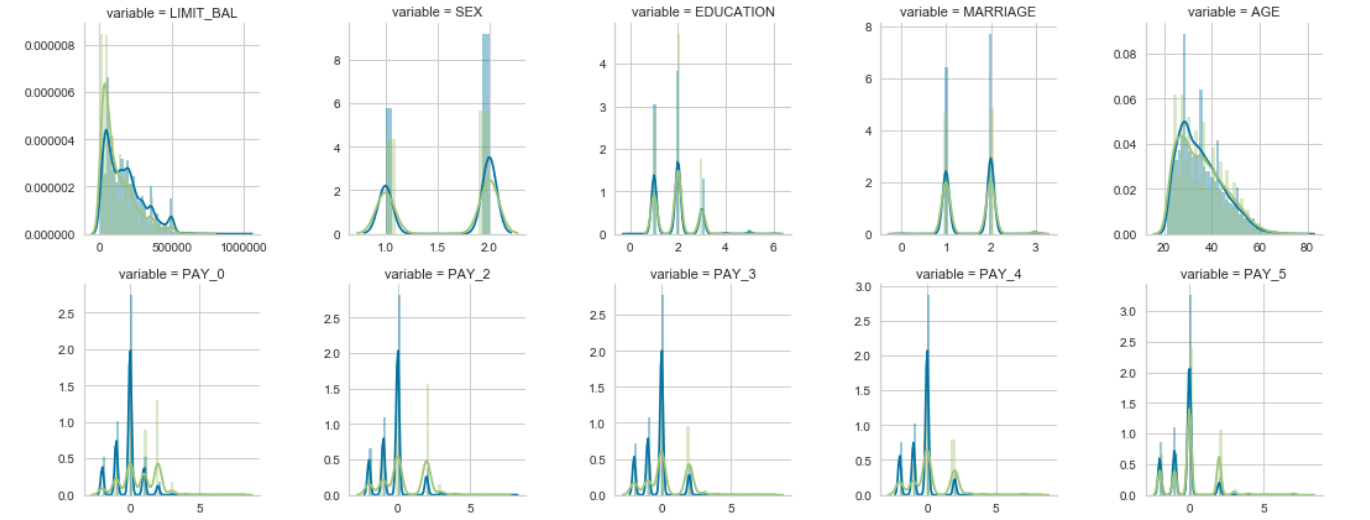
4. Customers with Marital status single are less like to default (MARRIAGE=2)

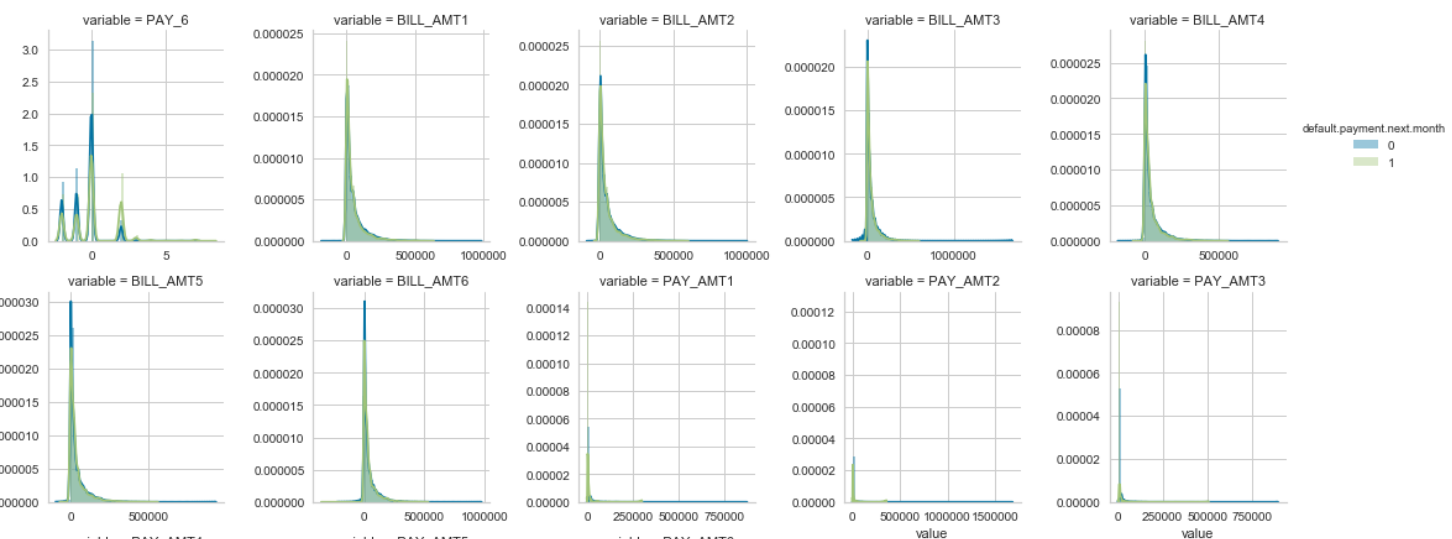
5. People in the age group 30-40 years are less likely to default

**Have you had to adjust your approach or research questions?**

After initial analysis of looking at the dataset values and the basic stats, I had to change my focus on considering many factors. Initially was under the impression that, Credit Card Default depends on Limit\_Balance, Education, Marriage, Pay months and limited factors. I saw surprising stats when I used visualizations to give clear idea on how each factor has its effect on the Credit Card Default.

Yes, I had to increase my research questions to explore and include more variables, than initially prepared. Its based on the initial analysis using visualization.





**Is your method working?**

Till now, I am comfortable and confident that my methods would work as initially planned. The steps which I am following are giving me good results as expected.

After completing the data acquisition and initial analysis, by running the stats and using visualization, I had to increase in the number of variables to explore and analyze and also consider for model building. If the number of variables are significantly more and if need, I am planning to apply dimensionality reduction methods like PCA to reduce the number of variables for model building.

**What challenges are you having?**

So far, I have not faced any major challenges. The process and the steps I am following are giving me good results as expected. Small challenges I have faced is to convert all the data in to common format using Python functions and syntax, which I expect get better on practice. I am now focusing on building the model and might challenge on implementing it. Based on the accuracy of the results, will have to restructure the code if its not as per expected and also planning to build multiple models to choose a better one which suits the dataset values.