Credit Card Default Prediction

Bhargava Gaggainpali  
DSC680 - Applied Data Science  
[bgaggainpali/bgaggainpali\_DSC680 (github.com)](https://github.com/bgaggainpali/bgaggainpali_DSC680)

Business problem & Hypothesis

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 delinquency rate on credit card loans. The increase in delinquency 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.

I have selected the topic, as I was interested in knowing the variables which influence the credit card default key factors. As I explore more about the domain, I understand that it’s not same set of rules which is being used across domain and each different banks and credit unions are based on different credit score calculation structure when approving credit cards, but the factors which influence the default are same.

Solution Method

I see this problem as a classification issue, where we should try to understand and able to predict the customers, who have high Credit Card default chances. Planning to use supervised machine learning algorithm to work on the classification problem to be trained with algorithms like:

1. Logistic Regression

2. Decision Tree

3. Random Forest

Start with loading data into a data frame and then understand the data, then perform Exploratory Data Analysis (EDA) on the data set. EDA involves doing Univariate and Bivariate Analysis, identify missing values and outliers and fill the gaps with appropriate values. In the next step, building model with starting from logistic regression and observe the accuracy of the model. When the accuracy of the of the model is not high, then planning to use Decision Tree and Random Forest to achieve higher accuracy.

Technical approach involves understanding the data by drawing multiple charts to observe the target variable with respect to each of the variable. Build a heat map to understand the relationship between variables. Build model using the different algorithms and observe the accuracy of the model, evaluate the accuracy of the model by building confusion matrix.

Data

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

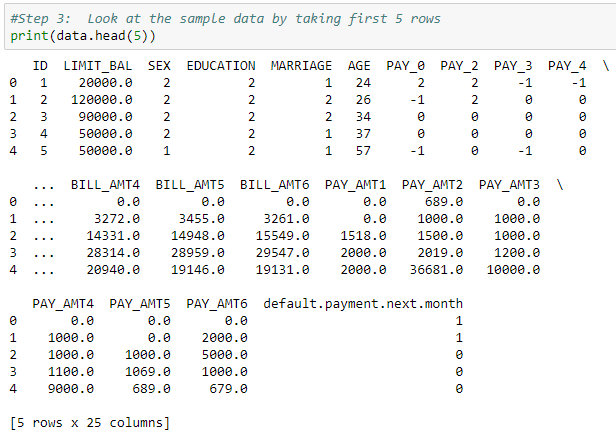
PAY\_AMT4 Payment Amount June 2005

PAY\_AMT5 Payment Amount May 2005

PAY\_AMT6 Payment Amount April 2005

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

Initial Observations



Categorical Features: Based on the data, below are categorical variables.

SEX

EDUCATION

MARRIAGE

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

Ordinal Features: Based on the data with inherent hierarchy, below are ordinal variables.

AGE

PAY\_0, PAY\_2, PAY\_3, PAY\_4, PAY\_5 & PAY\_6

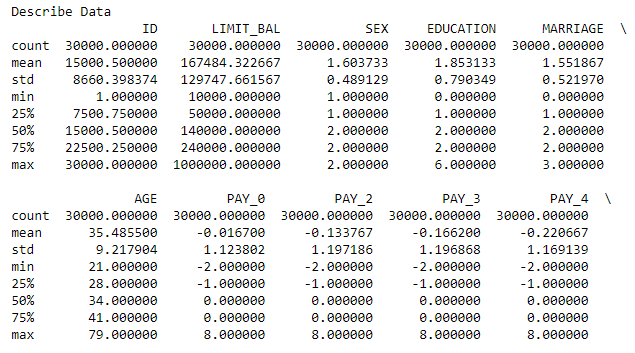
Numerical Features: Based on the numerical data, below are numerical variables.

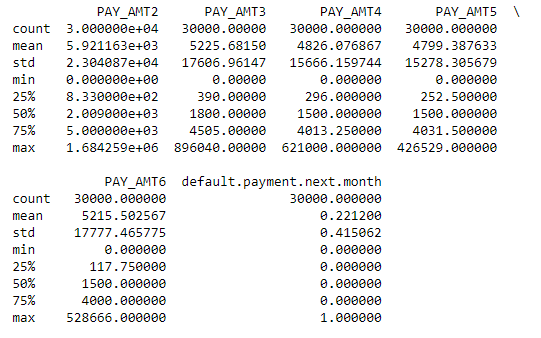
BILL\_AMT1, BILL\_AMT2, BILL\_AMT3, BILL\_AMT4, BILL\_AMT5 & BILL\_AMT6

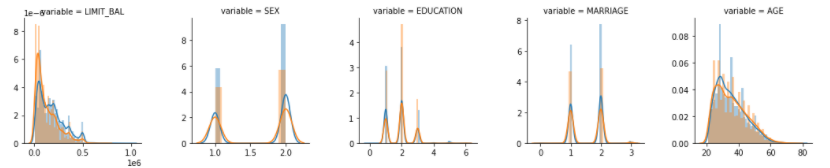
PAY\_AMT1, PAY\_AMT2, PAY\_AMT3, PAY\_AMT4, PAY\_AMT5 & PAY\_AMT6

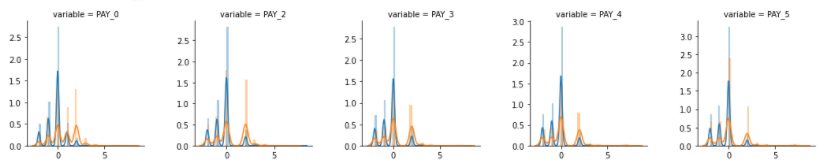
Exploratory Data Analysis

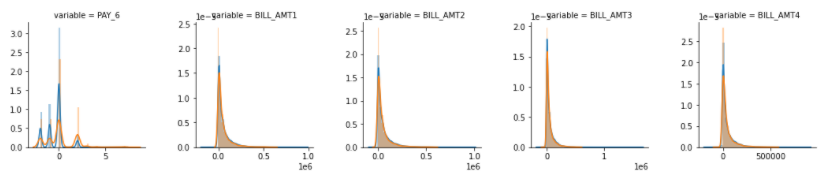
After initial analysis of looking at the dataset values and the basic stats, I had to change my focus on considering many factors. Initially was under the impression that, Credit Card Default depends on Limit\_Balance, Education, Marriage, Pay months and limited factors. I saw surprising stats when I used visualizations to give clear idea on how each factor has its effect on the Credit Card Default. I had to increase my research questions to explore and include more variables, than initially prepared. Its based on the initial analysis using visualization.

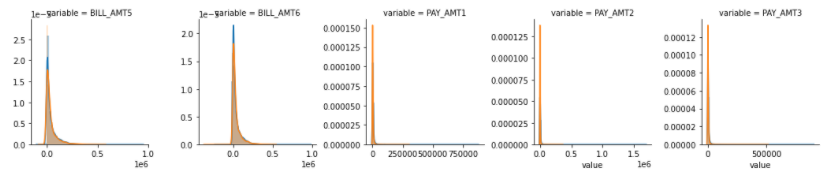


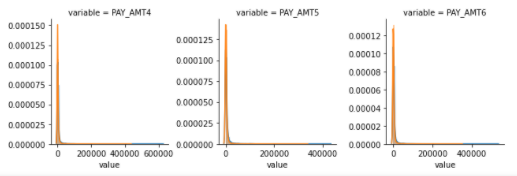


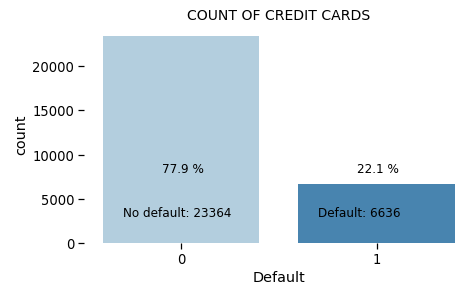
****

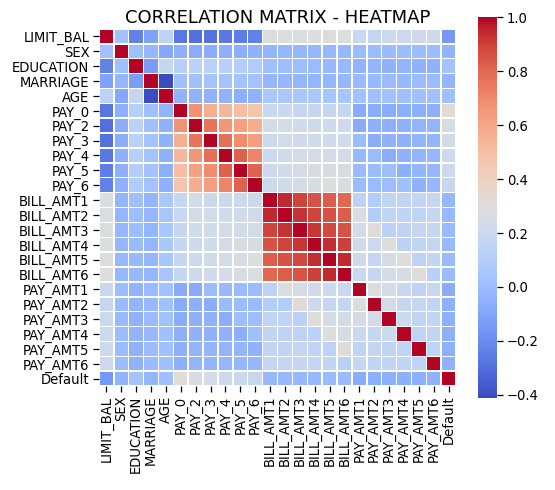
****

****

****

** **

****

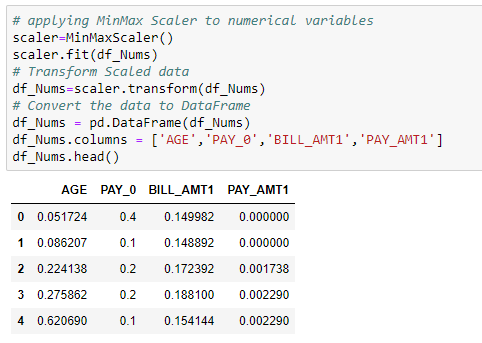
****

Few observations based on the above plots.

1. Customers with low LIMIT\_BAL have higher Default rate.
2. Default rate low among Females(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.

Data Preparation

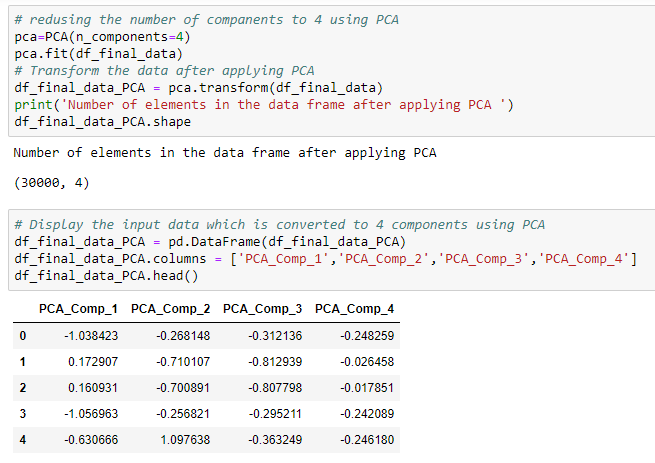
Applied MinMax Scaler to scale the numeric data variables.



Applied One Hot Encoding to convert Categorical data to Numerical data variables.

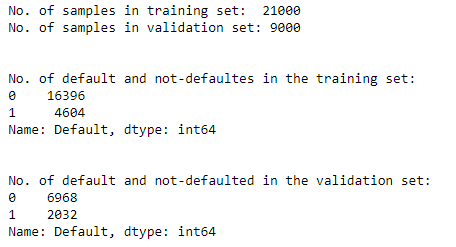


Applied PCA to reduce the number of Dimensions or Variables.



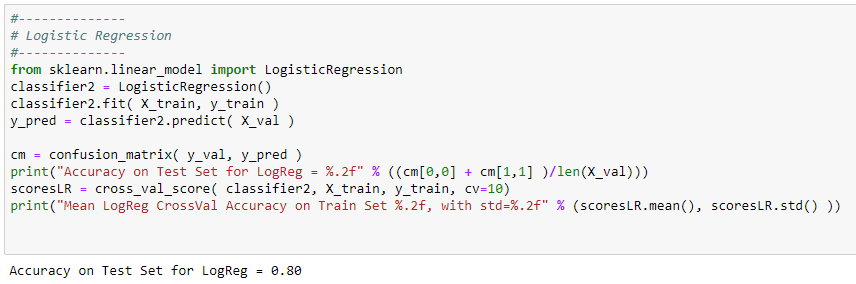
Model Development

As part of the current project, four models were developed after data preparation steps. Data is split in the ratio of 70:30 for train and test, i.e. 70% of the data is fed to the model to understand the patterns and remembering the outcome, later 30% of the data is used to validate the prediction results.

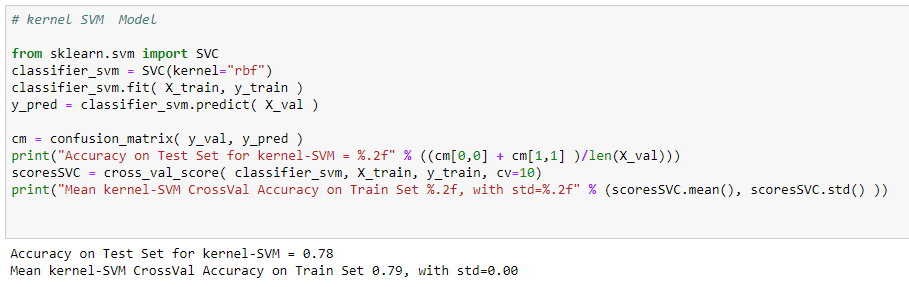


Below are four models

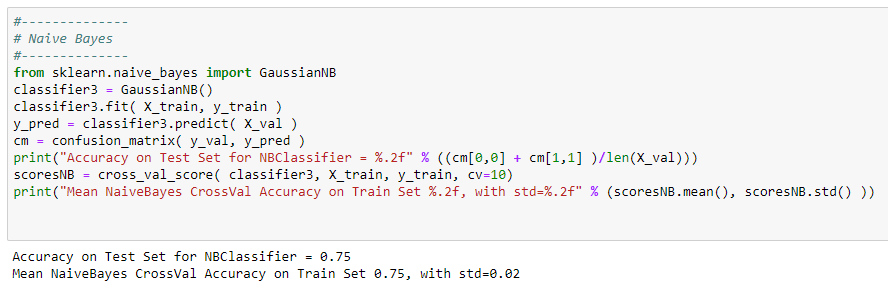
* Logistic Regression



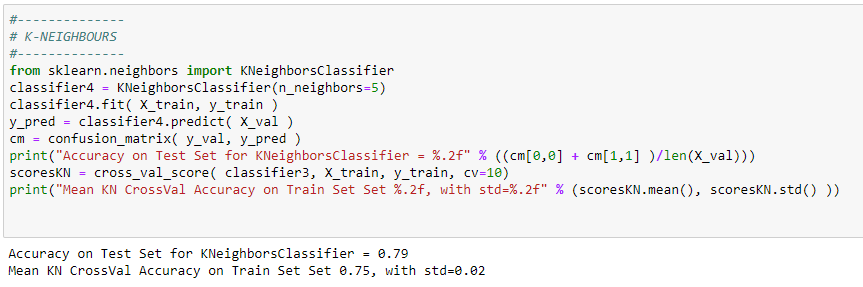
* Kernel SVM Model



* Naïve Bayes



* KNeighborsClassifier

****

Testing and Evaluation

After completing Model building using different algorithms, evaluate the accuracy of the model by building confusion matrix. As part of this project confusion matrix is built for each of the models, below is confusion matrix built on KNeighborsClassifier Model.

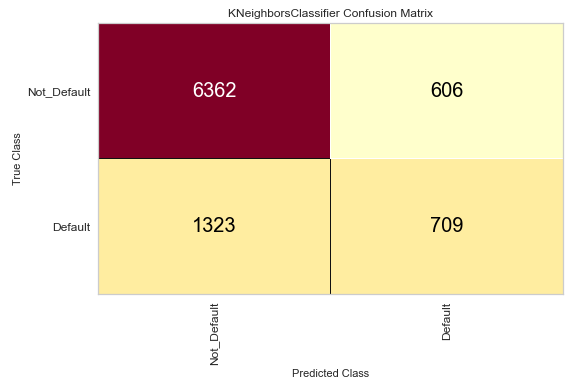


Chart shows the good precision and recall values and f1score 0.868 for not-Default cases indicates that model is performing as expected.

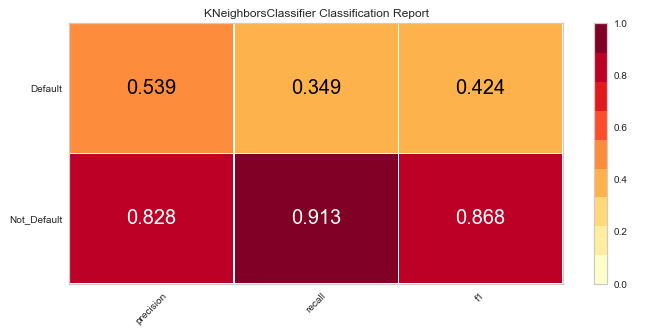
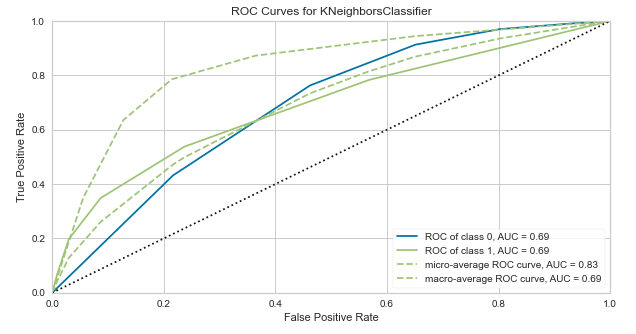
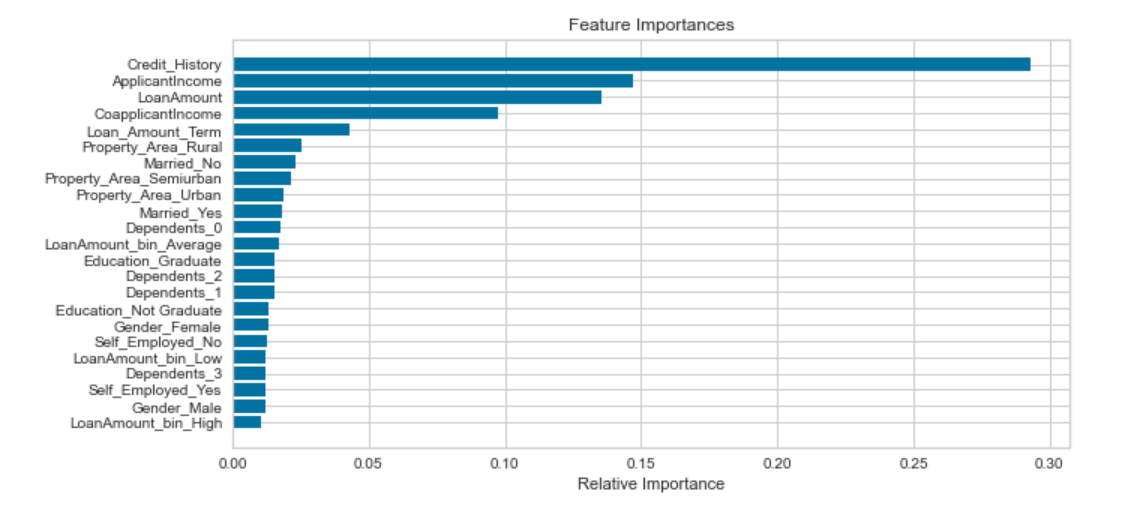
****

Chart shows the AUC (Area Under Curve) is at a value of 0.69 indicates that model is performing as expected.

****

Conclusion

Below chart shows that Bill Amount is most important factor in deciding the Credit Card default followed by Payment Amount and Age.



Future Analysis Questions

I would like to continue my analysis and try to explore further to find answers to the given questions.

1. With the COVID situation in place, are these features still valid to predict the Credit Card Default.

2. As with the different type of defaults like Credit Card, Car Loan, Home Loan, would these features stay in common across industry.

3. Will the feature importance change with respect to geographic location?

Reference:

1. LATOYA IRBY - February 10, 2020 - What You Can Do About Credit Card Default

[What You Can Do About Credit Card Default (thebalance.com)](https://www.thebalance.com/what-is-credit-card-default-960209)

2. Jenny Wang - Jun 24, 2020 - Will You Be Able to Make Your Credit Card Payment?

[Will You Be Able to Make Your Credit Card Payment? | by Jenny Wang | Towards Data Science](https://towardsdatascience.com/catch-me-if-you-can-predicting-credit-card-default-using-classification-analysis-28b2955f7f7d)

3. Marcos Dominguez - Feb 26,2021- Predicting Credit Card Defaults with Machine Learning

[Predicting Credit Card Defaults with Machine Learning | by Marcos Dominguez | The Startup | Feb, 2021 | Medium](https://medium.com/swlh/predicting-credit-card-defaults-with-machine-learning-fcc8da2fdafb)

4. Yashna Sayjadah, Ibrahim Abaker Targio Hashem, Faiz Alotaibi, Khairl Azhar Kasmiran - October 2018 - Credit Card Default Prediction using Machine Learning Techniques

[(PDF) Credit Card Default Prediction using Machine Learning Techniques (researchgate.net)](https://www.researchgate.net/publication/334765725_Credit_Card_Default_Prediction_using_Machine_Learning_Techniques)

5. Bank Rate – 2021 - Credit card default: How it happens, what to do about it

[Credit Card Default: What to Do About It | Bankrate.com](https://www.bankrate.com/finance/credit-cards/credit-card-default/)

 6. Equifax – 2021 - What Happens If I Default on a Loan or Credit Card Debt?

[Process & Potential Effects of Defaulting on a Loan | Equifax](https://www.equifax.com/personal/education/covid-19/default-loan-credit-card-debt/)