CREDIT CARD DEFAULT

ORION INNOVATION
PROJECT

MY EXPERIENCE

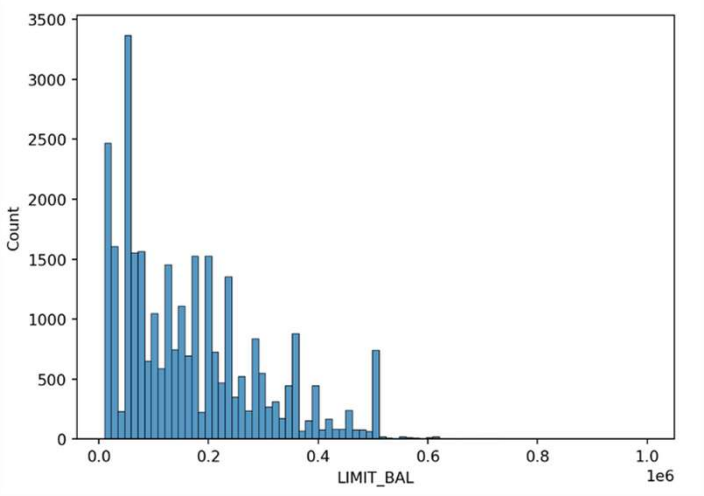
This is my first project with Orion Innovation, during the tenure of this project I learned how to:

- perform Data Analysis in the right manner
- perceive information from the graphs
- transform data (remove skewness)
- perform data modeling
- use quarto for report creation

ABOUT THE PROJECT

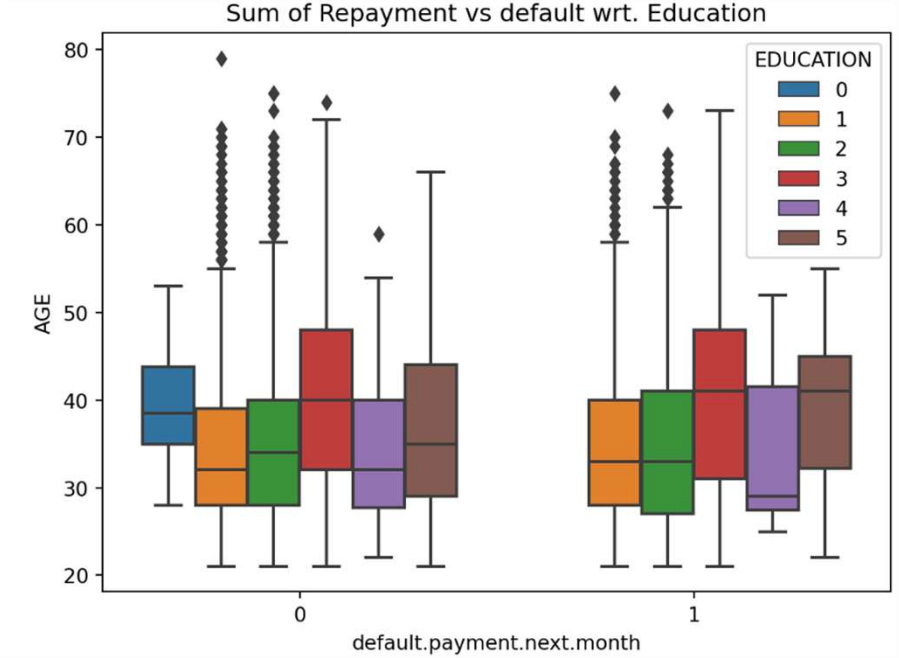
Credit Card default is a common issue that banks face each month. This project focuses on using the data provided by one such bank and working on that data. During this project, the data was cleaned and then trained to be used to create an effective model identifying defaulters. A number of models were applied to train the dataset, including random forest, logistic regression, SVM, and naive Bayes. This project helps to find which user will default in the upcoming month based on the given demographics.

FINDINGS

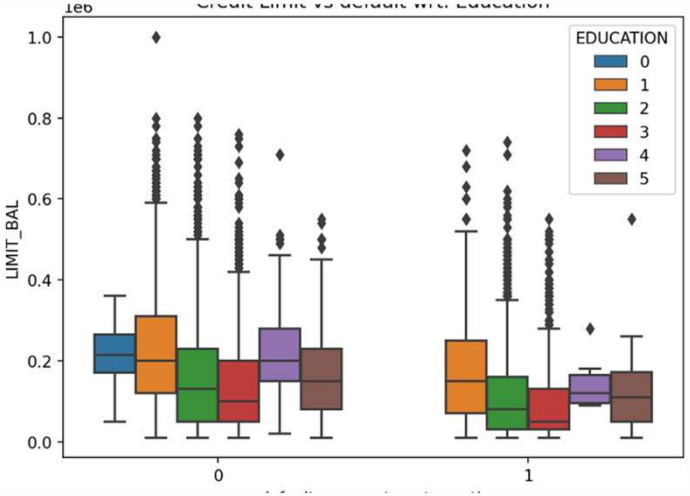


01 Sample Graph indicating that the data given is skewed.

02 People in high school belong to the higher age group of around 40.

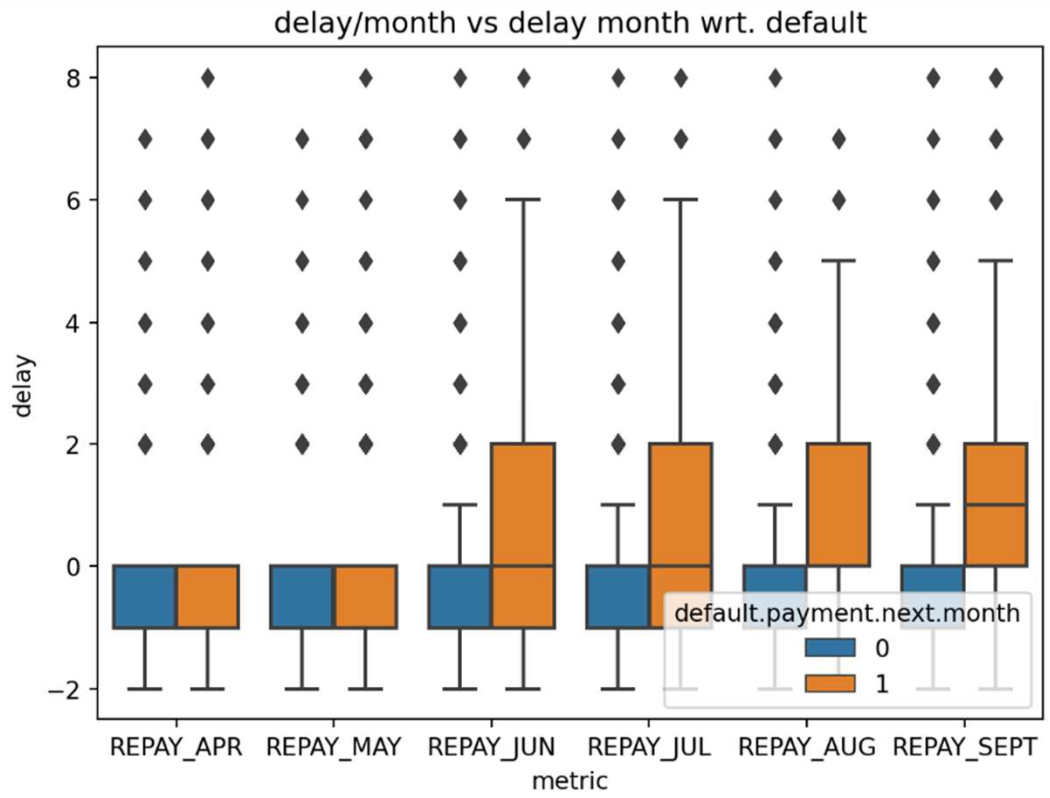


FINDINGS



03 People in high school have the least credit limit.

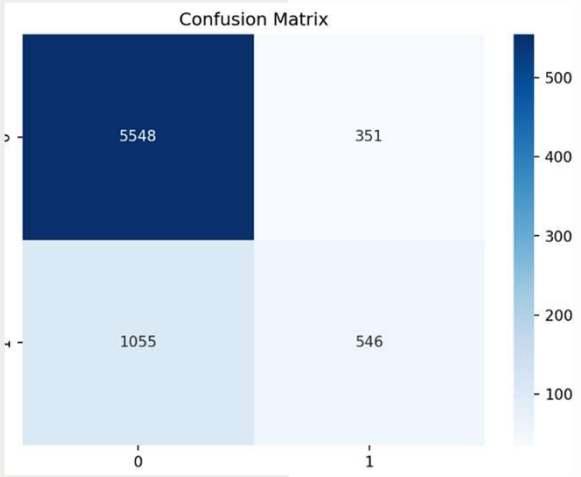
04 Repayment in September has the most impact on delay.



MODELLING

x1 = df[['LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'Repay_mean', 'BILL_APR', 'BILL_MAY', 'BILL_JUN', 'BILL_JUL', 'BILL_AUG', 'BILL_SEPT', 'PAID_APR', 'PAID_MAY', 'PAID_JUN', 'PAID_JUL', 'PAID_AUG', 'PAID_SEPT']]
y1 = df['default.payment.next.month']

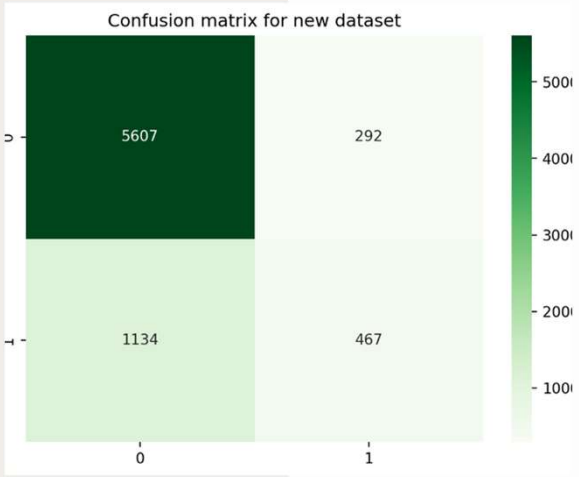
Using all the data given in the dataset we see that the correct predicted defaults are 546.



MODELLING

x2 = df[['SEX', 'EDUCATION', 'MARRIAGE', 'Repay_Sum']]
y2 = df['default.payment.next.month']

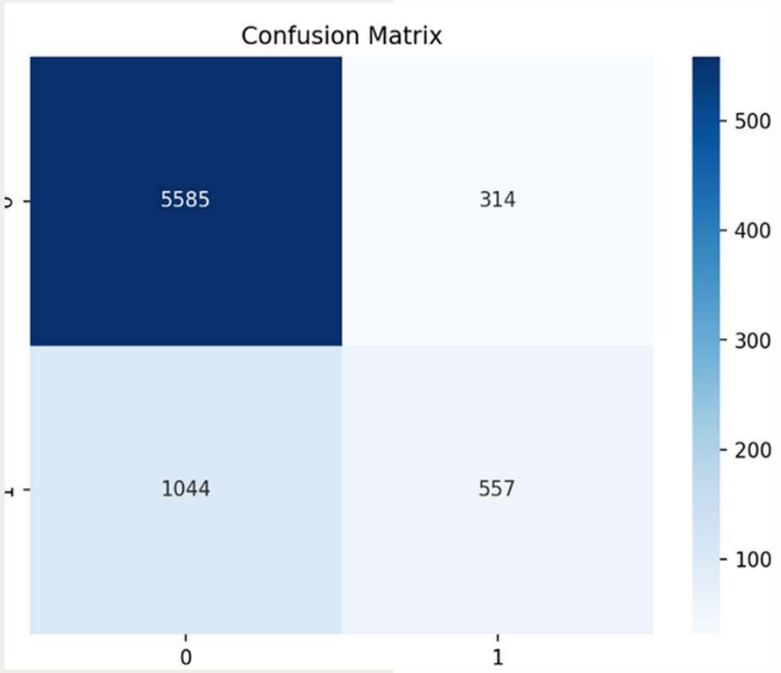
Using all the data that we analyzed to be useful during Data Analysis in the dataset we see that the correct predicted defaults have dropped which means we missed some of the important columns.



MODELLING

x1 = df[['REPAY_SEPT', 'BILL_APR', 'BILL_MAY', 'BILL_JUN', 'BILL_JUL', 'BILL_AUG', 'BILL_SEPT', 'PAID_APR', 'PAID_MAY', 'PAID_JUN', 'PAID_JUL', 'PAID_AUG', 'PAID_SEPT']]
y1 = df['default.payment.next.month']

Finally used the skewed data after removing skewness, along with the other columns analyzed to be important during Data Analysis, and we get 557 correct predictions, which is higher than the initial.



FINAL CONCLUSIONS

The data that is skewed should also be always analyzed and should not be assumed to be unimportant. As I modeled my data using multiple models the best results were given after the skewed data was used after removing the skewness, for each model.

1.
The random forest model is the most effective model for our dataset.

2.
We achieved an accuracy of 82% with the random forest model.

3.
The paid amount and bill amount along with the repayment status of the previous month have the most effect on the chances of default of a user.

CONTACT INFO

E-mail anshikajain2405@gmail.com

Report https://anshikajain2405.github.io/OrionDataAnalyticsInternshipJul23/docs.html

Website https://github.com/AnshikaJain2405/OrionDataAnalyticsInternshipJul23