CREDIT CARD DEFAULT PREDICTION

DETAILED PROJECT REPORT

AHAMED AJAS S

AFRIN S

iNeuron

PROJECT DETAIL

|	
Project Title	Credit Card Default Prediction
Technology	Machine Learning Technology
Domain	Banking
Project Difficulty level	Intermediate
Programming Language Used	Python
Tools Used	NumPy, Pandas, R, Tableau, MS-Excel, MS-Power BI

OBJECTIVE

- The Goal of credit card default prediction is to develop a predictive model that can accurately identify whether a credit card holder is likely to default on their payments in the upcoming month.

PROBLEM STATEMENT

- Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. The goal is to predict the probability of credit default based on credit card owner's characteristics and payment history.

ARCHITECTURE
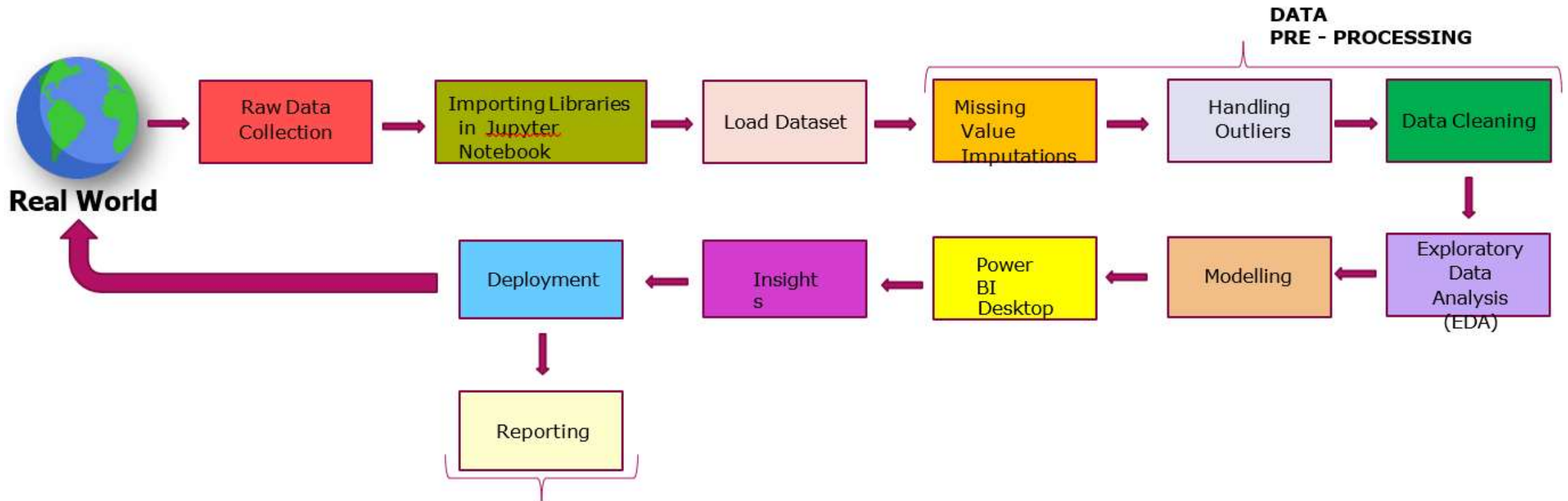
DATASET INFORMATION

ID: ID of each client

LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit

SEX: Gender (1=male, 2=female)

EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

**MARRIAGE:** Marital status (1=married, 2=single, 3=others)

AGE: Age in years

PAY_0 to PAY_6: Repayment status in September, 2005 (-1=pay duly,1=payment delay for one month, 2=payment delay for two months, 8=payment delay for eight months,9=payment delay for nine months and above)

BILL_AMT 1 to BILL_AMT 6: Amount of bill statement for the last 6 months.

PAY_AMT 1 to PAY_AMT 6: Amount of previous payment made for the last 6 months.

Default.payment.next.month: Target variable indicating whether the individual will default on the credit card payment next month (1 = yes, 0 = no).
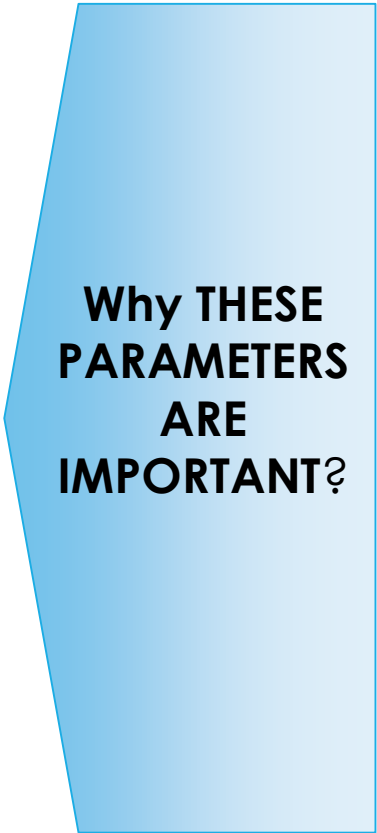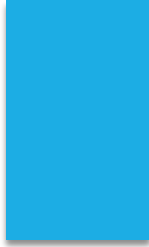
ID: While the ID itself might not directly contribute to predicting defaults, it serves as a unique identifier for each record, helping in data management and referencing.

Limit Balance: The credit limit or balance is a crucial factor as it represents the maximum amount a credit card holder can borrow. Higher limits may indicate higher risk, especially if not managed responsibly.

Sex: Gender can be a demographic factor that influences financial behaviors. Some models consider it as a feature, although it's important to handle such variables ethically and ensure they don't contribute to bias or discrimination.

Education: The level of education can influence financial literacy and stability. It may provide insights into a person's ability to manage credit responsibly.

Marriage: Marital status can be a factor in financial stability. For example, married individuals may have shared financial responsibilities, potentially impacting their credit behavior.

Age: Age is a critical factor as younger individuals may have less financial experience and stability. Older individuals may have a more established credit history, providing insights into their creditworthiness.

Bill Amount: The bill amounts across different months indicate the credit card holder's spending behavior. Higher bill amounts relative to the credit limit may suggest financial stress.

Pay Amount: Payment amounts reflect how much of the billed amount the cardholder is repaying. Consistently low payments or delayed payments may indicate financial distress and an increased risk of default.

Default Payment Next Month: This is the target variable that the model aims to predict. It indicates whether a customer will default on their payment in the next month. Understanding and predicting this outcome is the primary objective of credit card default prediction models.

Why THESE PARAMETERS ARE IMPORTANT?

INSIGHTS
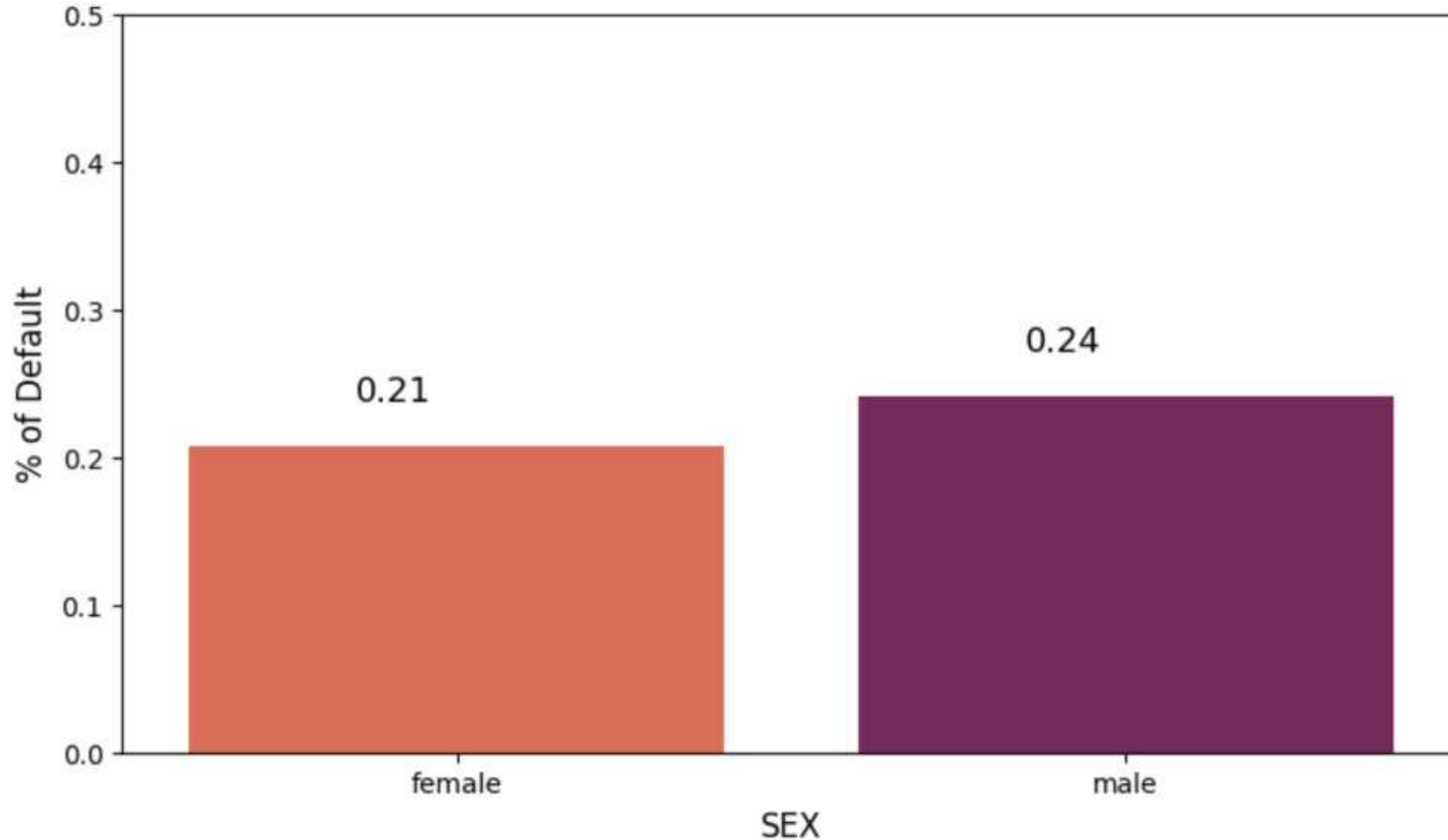
What Kind Of Population do we have?
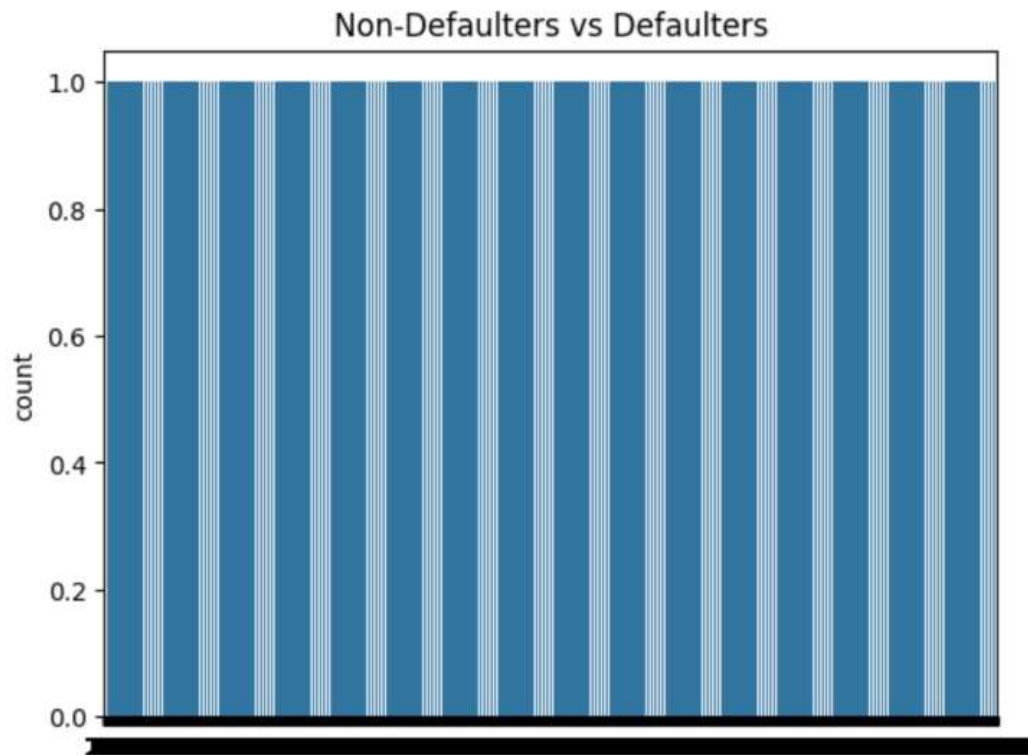
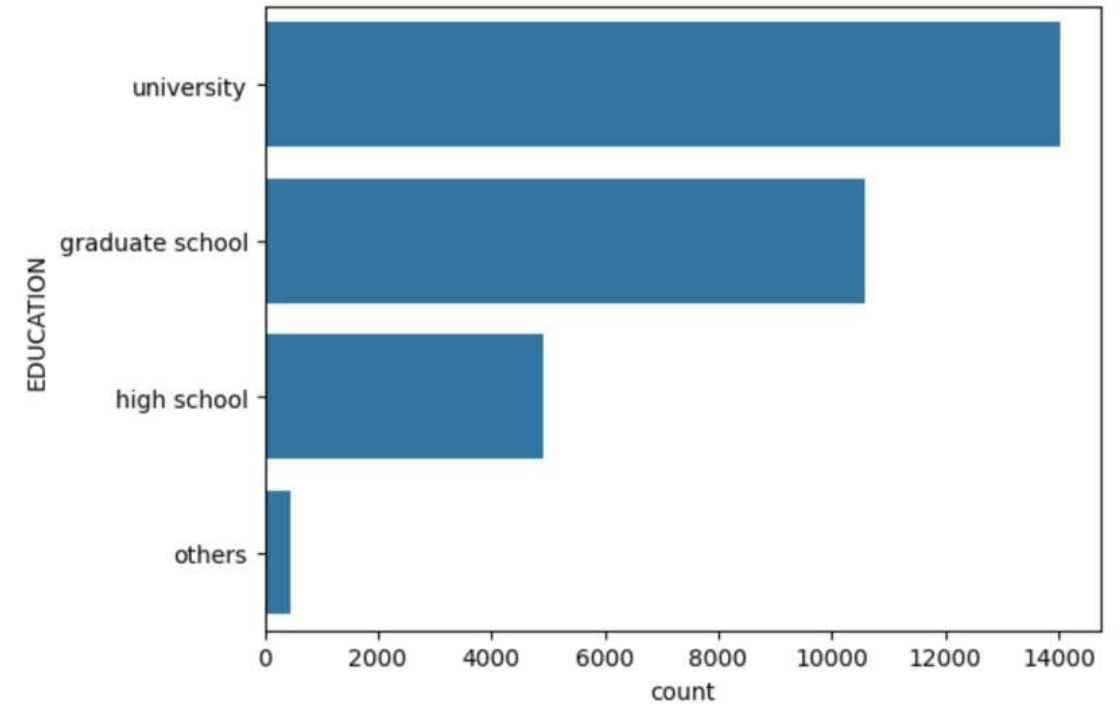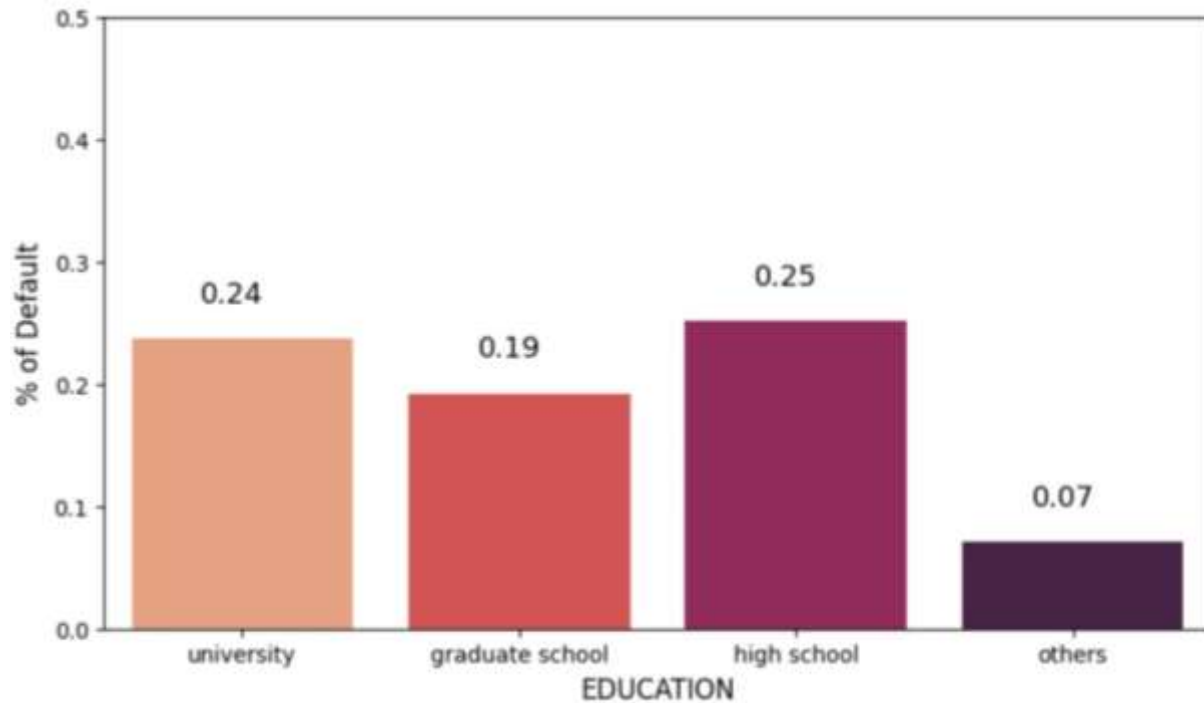- ❑ 0.5 Non Defaulters and 1.0 Defaulters
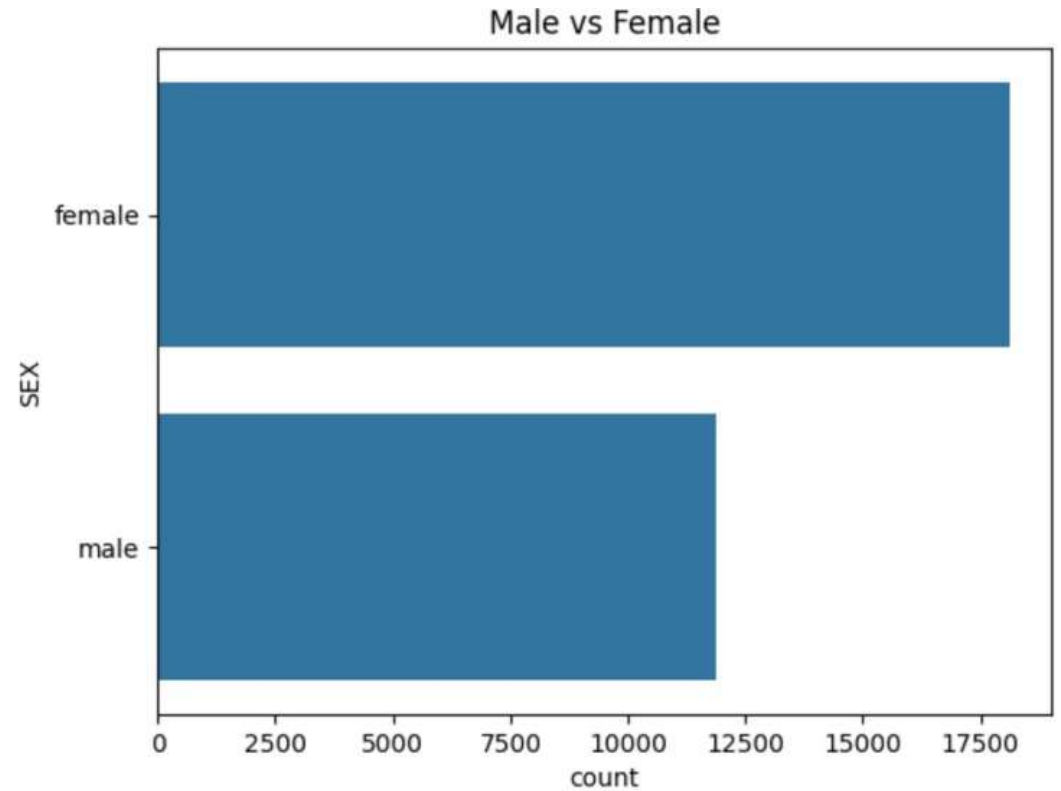
- ❑ Females have higher count than men.

Males or Females who are more likely to default next month ?

❑   So, from plot we can say Males are more likely to Default next month than females

Which level of education shows a higher likelihood of default ?

❑ The predominant level of education in our dataset is "University" , followed by ' Grade School', 'High School' and 'Others'.
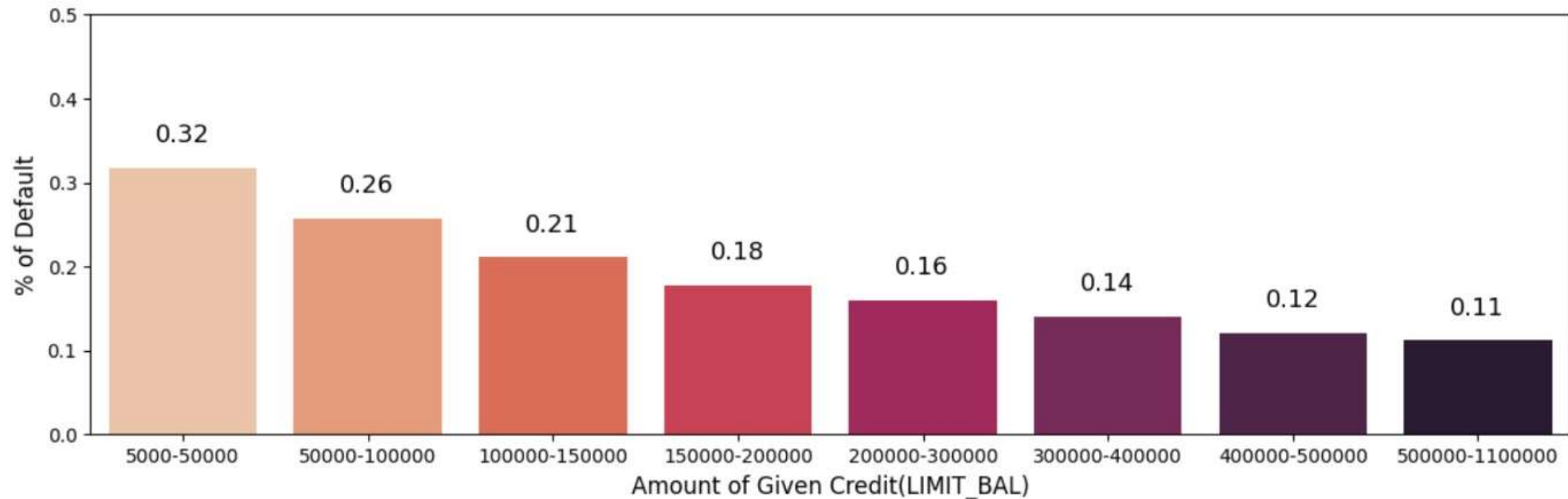
❑ From plot, Highschool and University candidates are more likely to Default.

Visualizing the distribution of limit balances according to months.
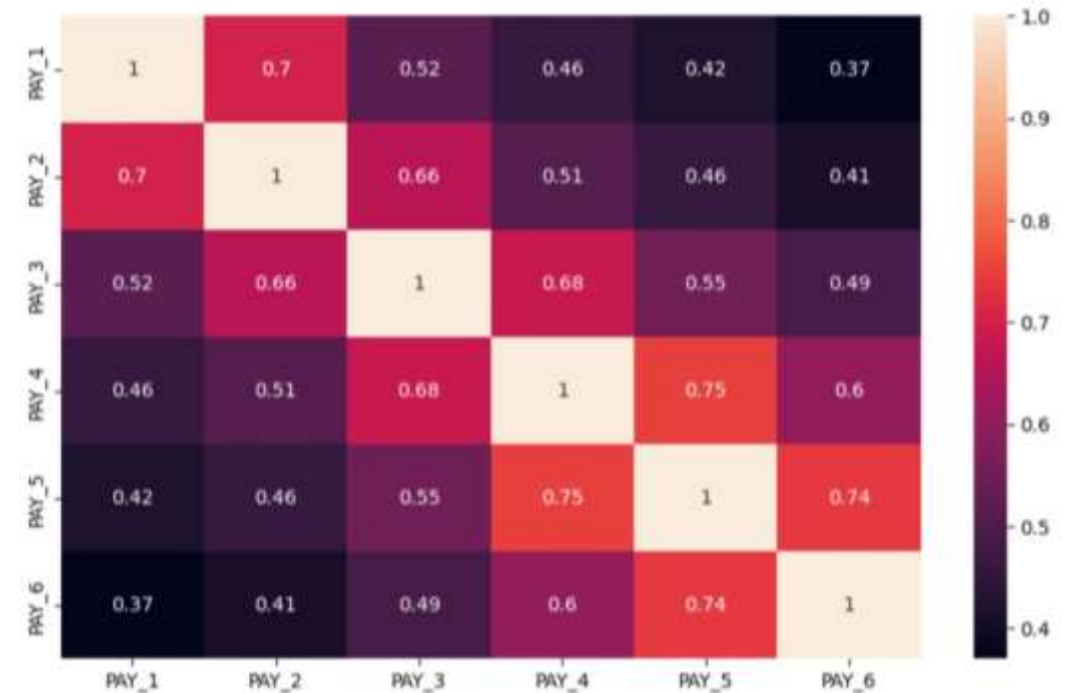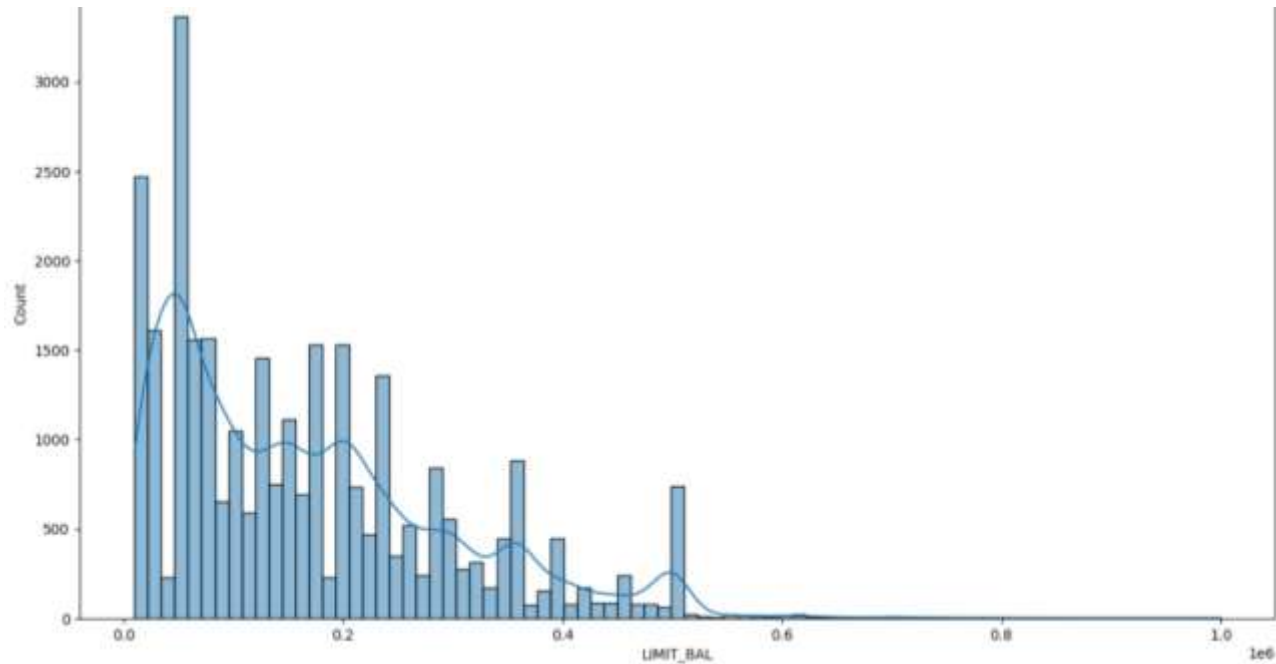
❑ It represents the maximum amount a credit card holder can borrow

Is there a noticeable trend in default rates based on credit limits in the dataset?

- ❑ Now, we got the clear picture that there is a significant rate of default (over 30%) from customers with 50k or less of credit limit.
- ❑ We also can say that the higher the limit, the lower is the chance of defaulting

KPIs (Key Performance Indicators)

- Impact of Credit Utilization on Default Probability
- Impact of Payment History on Default Probability
- Influence of Debt-to-Income Ratio on Default Probability
- Influence of Employment Status on Default Probability
- Influence of Age on Default Probability

CONCLUSION

- 0.5 Non Defaulters and 1.0 Defaulters

- Females have higher count than men.

- Males are more likely to Default next month than females

- The predominant level of education in our dataset is "University" , followed by ' Grad School', 'High School' and 'Others'.

- From plot, Highschool and University candidates are more likely to Default.

- It represents the maximum amount a credit card holder can borrow

- we got the clear picture that there is a significant rate of default (over 30%) from customers with 50k or less of credit limit.

- We also can say that the higher the limit, the lower is the chance of defaulting.

Q & A

Q1) What's the source of dataset ?
Ans) The Dataset was taken from iNeuron's Provided Project Description Document.
https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/data

Q2) What was the type of data ?
Ans) The data was combination of numerical and categorical values.

Q3) What's the complete flow you followed in this project ?
Ans) Refer Slide 5th for better understanding.

Q4) What Techniques were you using for data ?
Ans)
    - Visualizing relation of independent variables with each other & output variables.
    - Removing unwanted attributes.
    - Removing outliers.
    - Cleaning data imputing if null values are present.
    - Converting data into Categorical Values.

Q5) What were the libraries that you used in python ?
Ans) We used Pandas, NumPy and Matplotlib and Seaborn libraries in Pandas.

THANK YOU