Name:Shambhavi Jha | Enrollment No:23117131

REPORT:CREDIT CARD DEFAULT PREDICTION

UNDERSTANDING OF PROBLEM STATEMENT

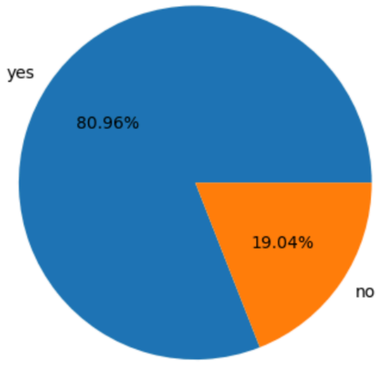
In this project, I was given two datasets by Bank A. The **training dataset** contained around 25,000 customer records with features like credit limit, payment history, bill amounts, and a target column called ***next_month_default***, which shows whether the customer defaulted on payment in the next month.

The **validation dataset** had about 5,000 similar customer records but **without the target column**. My task was to build a classification model that predicts whether a customer will default next month and generate predictions for the validation set.

The main goal was to help the bank identify potential defaulters early so they can manage credit risk better. The focus was on creating a model that is not just accurate, but also interpretable and aligned with financial risk — especially by maximizing the **F2 Score**, which gives more importance to identifying actual defaulters.

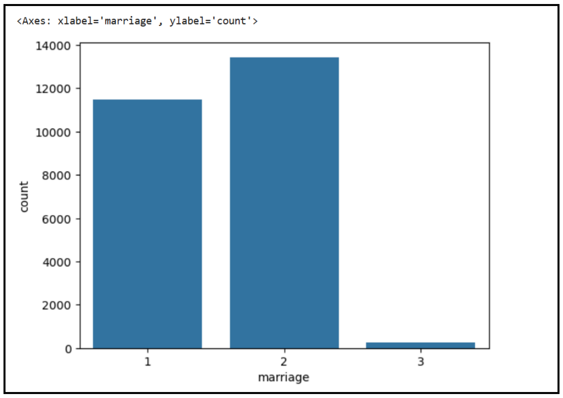
DATA QUALITY CHECK

- Initially inspected the dataset to check for missing values and data types.
- Found that most columns were non-null, but a few missing values were present in the **Age** column.
- Missing values were imputed using the **median** of the respective column to avoid skewing the data.
- Discovered that the **target variable (next_month_default) was highly imbalanced**, with far fewer defaulters than non-defaulters.
- The majority class accounts for **80.96%** of the data, while the minority class represents only **19.04%.**

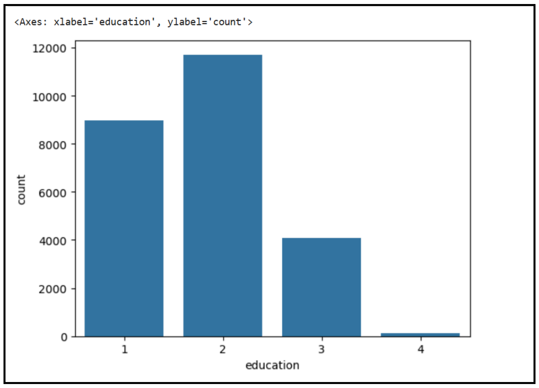


EXPLORATORY DATA ANALYSIS (EDA)

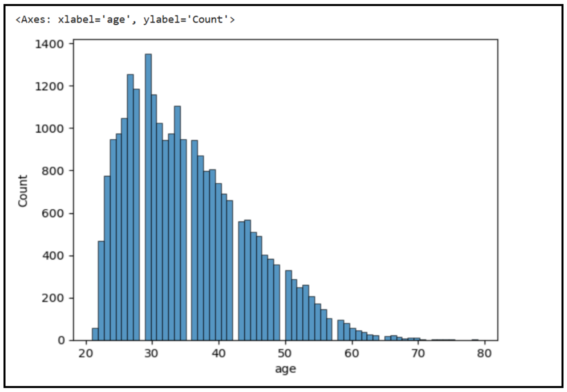
- Conducted univariate and bivariate analysis to understand distributions and relationships between variables.
- Boxplots were plotted for key numerical features like credit limit, bill amounts, payments, and age to visually inspect the presence of outliers. Although some outliers were observed, they were not removed or capped, as the majority of models used (such as Random Forest, XGBoost, and LightGBM) are tree-based and generally robust to outliers. Removing them could also risk losing important edge-case behaviors relevant to credit risk.
- For the univariate analysis-countplot for ‘age’,’marriage’,’sex’ and ‘education’ were plotted.
- Males are more represented in the dataset than females, but both genders are fairly well represented.
- The ages are most commonly concentrated between 30 to 40 years. This suggests that the dataset is dominated by middle-aged individuals, which may reflect the typical demographic of credit card holders.
- The majority of users have education level 3 (likely High School). Education levels 1 and 2 are present in much smaller proportions.
- Most clients are either married or single, with singles slightly more in number.
- Some rows consisted of the 0 value for marriage which is invalid and 0,5,6 value for education, which is also invalid. So these rows were later filtered to keep only the meaningful ones.
- Below are the countplots of marriage, education and sex and the histogram of age of filtering.



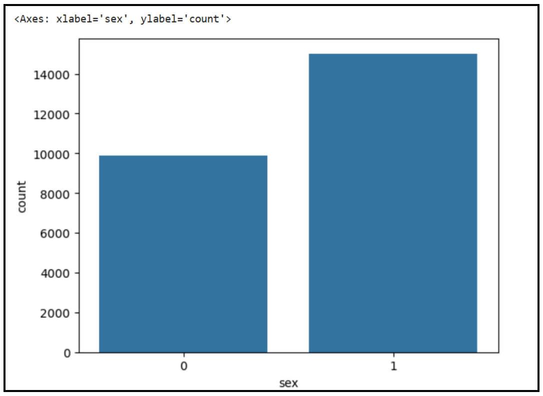
```
<Axes: xlabel='marriage', ylabel='count'>
```



```
<Axes: xlabel='education', ylabel='count'>
```

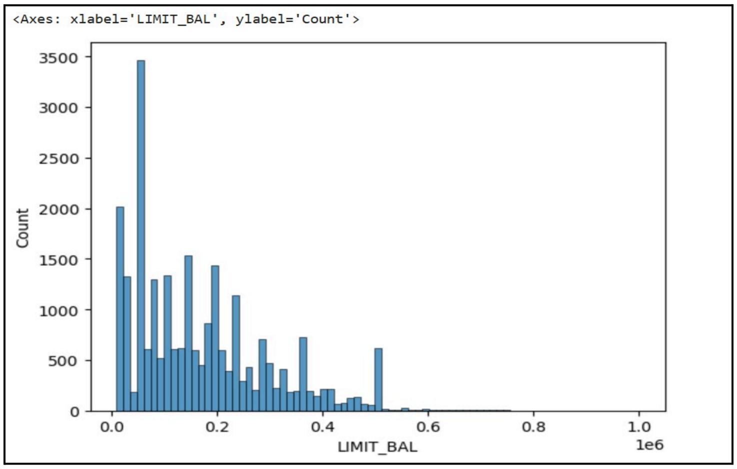


```
<Axes: xlabel='age', ylabel='Count'>
```

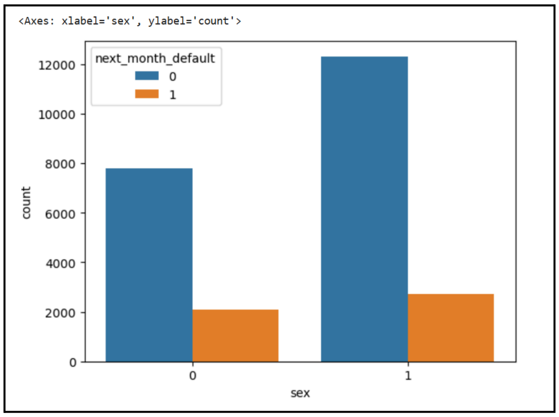


```
<Axes: xlabel='sex', ylabel='count'>
```

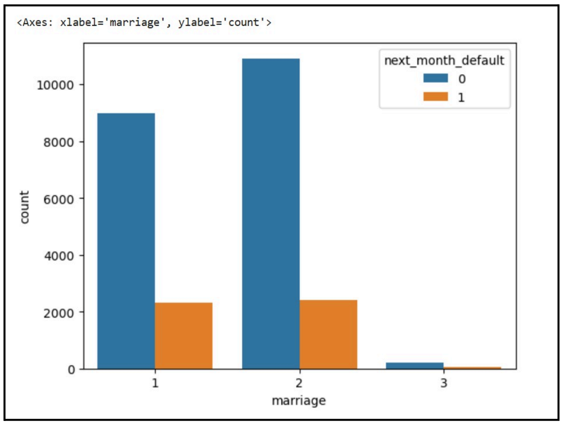
- This histogram below shows the distribution of customers based on their credit limit (LIMIT_BAL). Most customers have credit limits below 200,000, with the frequency decreasing as the credit limit increases. The distribution is right-skewed, indicating that only a small number of customers have very high credit limits. This suggests that the majority of clients have access to relatively modest credit amounts.



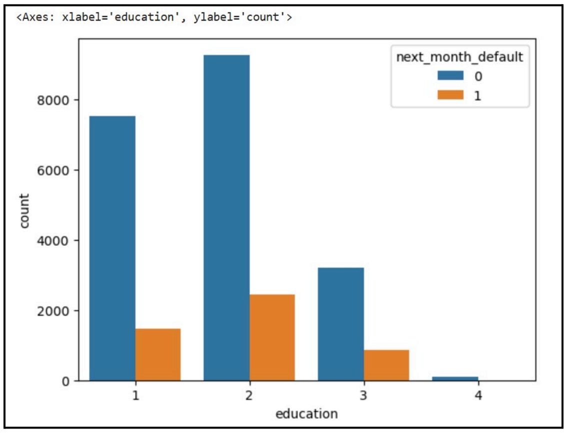
```
<Axes: xlabel='LIMIT_BAL', ylabel='Count'>
```



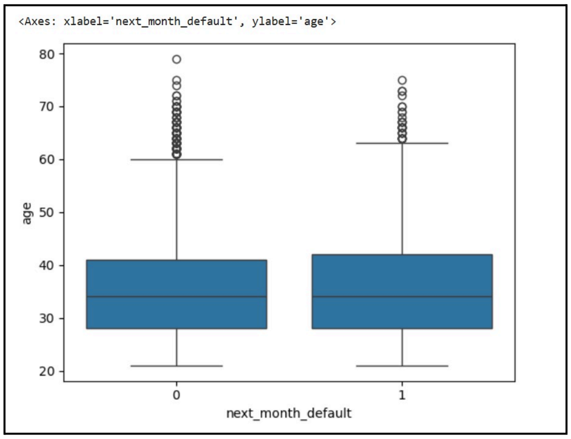
```
<Axes: xlabel='sex', ylabel='count'>
```



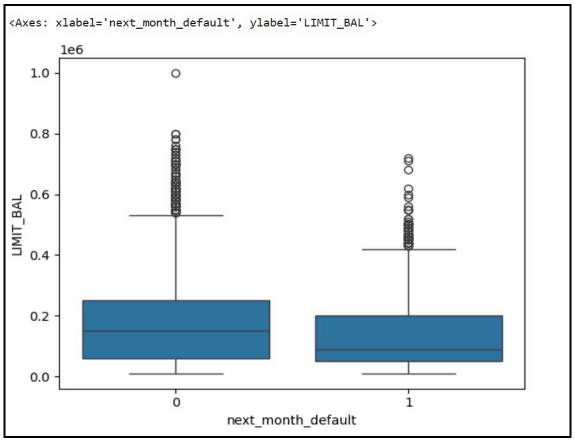
```
<Axes: xlabel='marriage', ylabel='count'>
```



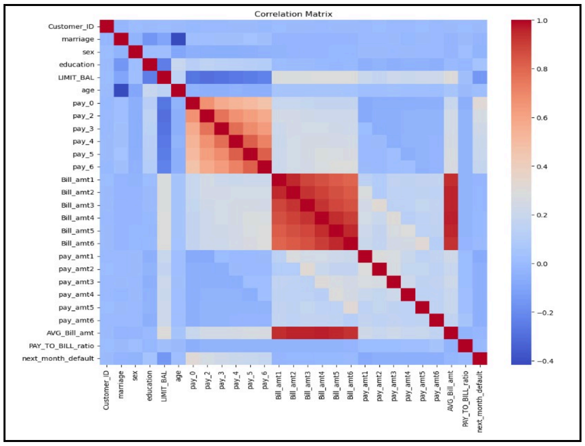
```
<Axes: xlabel='education', ylabel='count'>
```



```
<Axes: xlabel='next_month_default', ylabel='age'>
```



```
<Axes: xlabel='next_month_default', ylabel='LIMIT_BAL'>
```



- **Marriage vs. Default**: Most customers are either single (1) or married (2), with similar default rates across both groups. Default is relatively rare among 'others' (3), likely due to fewer data points.
- **Education vs. Default**: Customers with university (2) and graduate (1) education levels form the majority. Default rates are fairly similar across education levels, but those with lower education (3) show a slightly higher proportion of defaults.
- **Gender vs. Default**: Males (1) are more represented than females (0) and show a slightly higher count of defaults, though the difference is minor.
- **Age vs Default**: Age distribution is similar for both default and non-default groups, mostly between 25–45 years.
- **Limit Balance vs Default**: Defaulters tend to have lower credit limits on average compared to non-defaulters.
- The correlation matrix shows there is no single strong predictor of default, but payment delays and lower payment ratios are moderately linked to defaults.

FEATURE ENGINEERING

To improve model performance and make the predictions more financially interpretable, I engineered the following new features:

1. **Credit Utilization Ratio**
   `credit_util_ratio = AVG_BILL_AMT / LIMIT_BAL`
   This ratio measures how much of the available credit is being used on average. A high utilization ratio can indicate potential repayment stress and is a strong predictor of credit risk.

2. **Total Payment Overdue Months**
   `total_overdue_months = count of PAY_m values ≥ 1`
   This feature sums the number of months where payments were overdue. It helps capture chronic

repayment issues and long-term delinquency.

3. **Payment Coefficient of Variation (pay_amt_cv)**
   `pay_amt_cv = std(pay_amt1 to pay_amt6) / mean(pay_amt1 to pay_amt6)`
   This measures inconsistency in repayment amounts over the last 6 months. A high CV indicates erratic repayment behavior, which can be a sign of financial instability.

4. **Delinquency Streak**
   `max_streak = longest continuous sequence of PAY_m ≥ 1`
   This feature captures how long a customer remained continuously delinquent. Longer streaks are often linked to higher default probability.

5. **Bill Amount Trend**
   `bill_amt_trend = slope of linear regression line through bill_amt1 to bill_amt6`
   This captures whether the customer's spending is increasing, decreasing, or stable. An upward trend might signal rising financial pressure if not matched by payments.

MODEL COMPARISON

To develop a predictive model for credit card default, several classification algorithms were evaluated. The entire modeling process followed a structured pipeline:

- The data was split into training and testing sets using an 80:20 ratio. This means 80% of the data was used for model training, and the remaining 20% was used for evaluating model performance on unseen data.
- The target variable (default.payment.next.month) was highly imbalanced, with a significantly larger number of non-defaulters (class 0) compared to defaulters (class 1). To address this imbalance and improve model sensitivity towards the minority class, **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to the training data.
- Since many machine learning algorithms are sensitive to the scale of input data (especially Logistic Regression and XGBoost), **StandardScaler** was applied to normalize the numerical features. Scaling was performed **after** SMOTE to ensure synthetic samples were also scaled.

To build a robust credit default prediction system, I experimented with multiple classifiers: **Random Forest**, **Logistic Regression**, **XGBoost**, and **LightGBM**. All models were trained after applying **SMOTE** for class balancing and **StandardScaler** for feature scaling.

1. Random Forest

- Performed reasonably well with high accuracy (81%) and good performance on class 0 (non-defaulters).
- However, it struggled to capture the minority class, with class 1 recall = **0.48** and F1 = **0.5**.
- F2=0.485
- ROC-AUC Score=0.7732

2. Logistic Regression

- Also biased towards class 0.
- Achieved slightly better recall (0.59) for defaulters but had a low precision (0.36).
- Overall F1-score for class 1 was **0.45**.
- F2=0.523
- ROC-AUC Score=0.7276

3. XGBoost (Default Threshold)

- Focused more on maximizing recall for defaulters (0.78), but at the cost of precision (0.29), and hence low F1 = **0.43**.
- Accuracy dropped to **60%**, showing poor balance between classes.

4. XGBoost (Best Threshold = 0.42)

- Custom threshold tuning improved F2-score to **0.5943**, showing better alignment with recall-oriented objectives.
- Class 1 recall increased to **0.85**, but precision dropped to **0.27**, keeping F1 low at **0.41**.
- ROC-AUC Score=0.7499

5. LightGBM (Default Threshold)

- Class 1 precision = **0.56**, recall = **0.38**, and F1 = **0.46**.
- Accuracy = **83%**, matching Random Forest but with better defaulter handling.

6. LightGBM (Best Threshold = 0.11)

- After threshold tuning, LightGBM achieved an **F2-score** of **0.5915**.
- It maintained high recall (0.82) for defaulters.
- Though precision was 0.28, it was a reasonable trade-off for improved defaulter detection.
- ROC-AUC Score=0.7643

XGBoost with tuned threshold was selected as the final model because:

- It had the **highest F2-score** (0.5943), aligning with the project goal of prioritizing recall for high-risk predictions.
- Achieved strong class 1 recall (**0.85**) with acceptable precision.
- Outperformed other models in identifying defaulters while maintaining reasonable overall accuracy (**54%** after threshold tuning).
- Its speed and efficiency also make it practical for deployment.
- It also had a good roc-auc score of **0.7499.**

Validation Results

After selecting **LightGBM with a tuned threshold (0.13)** as the final model based on its superior F2-score on the test set, the model was evaluated on the separate **validation dataset** to estimate its real-world generalization performance.

Predictions on Validation Set

The model was used to predict the class labels on the validation set after applying the same preprocessing steps:

- The **validation features were scaled** using the `StandardScaler` fitted on the training set.
- The model's predicted **probabilities were thresholded at 0.42**, as this was previously determined to give the best F2-score.
- Final class predictions were generated based on this threshold.

Tools Used: Pandas,NumPy,Matplotlib,Seaborn,Scikit-learn,XGBoost,LightGBM,imbalanced-learn (SMOTE)

Conclusion: The final model prioritizes early identification of defaulters at the cost of some false positives, which is suitable for credit risk management.

.