CREDIT CARD DEFAULT PREDICTION USING ML
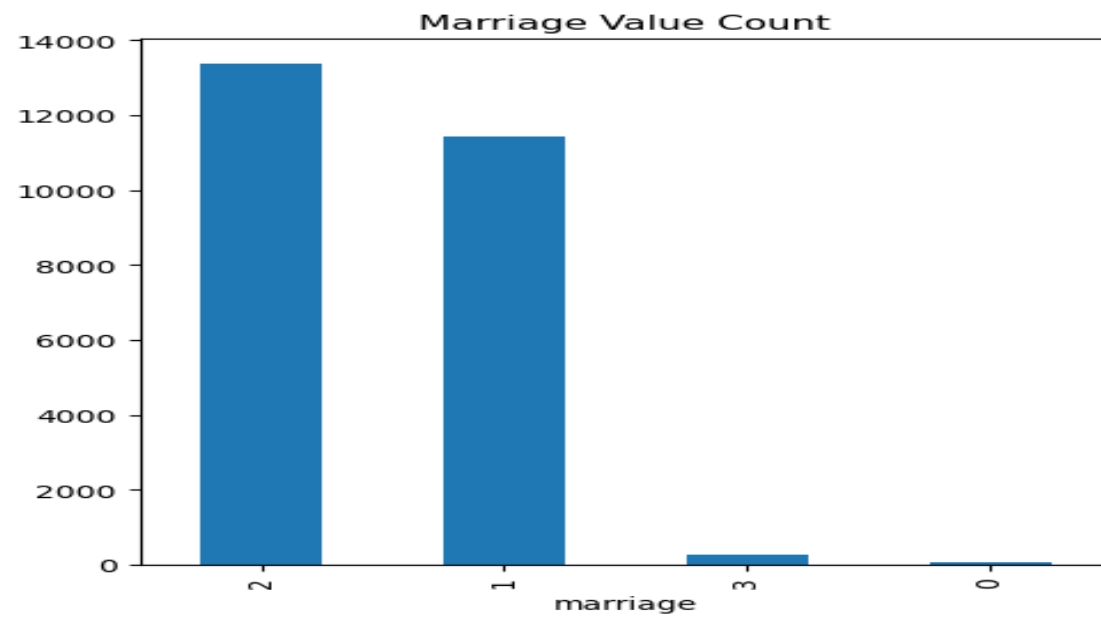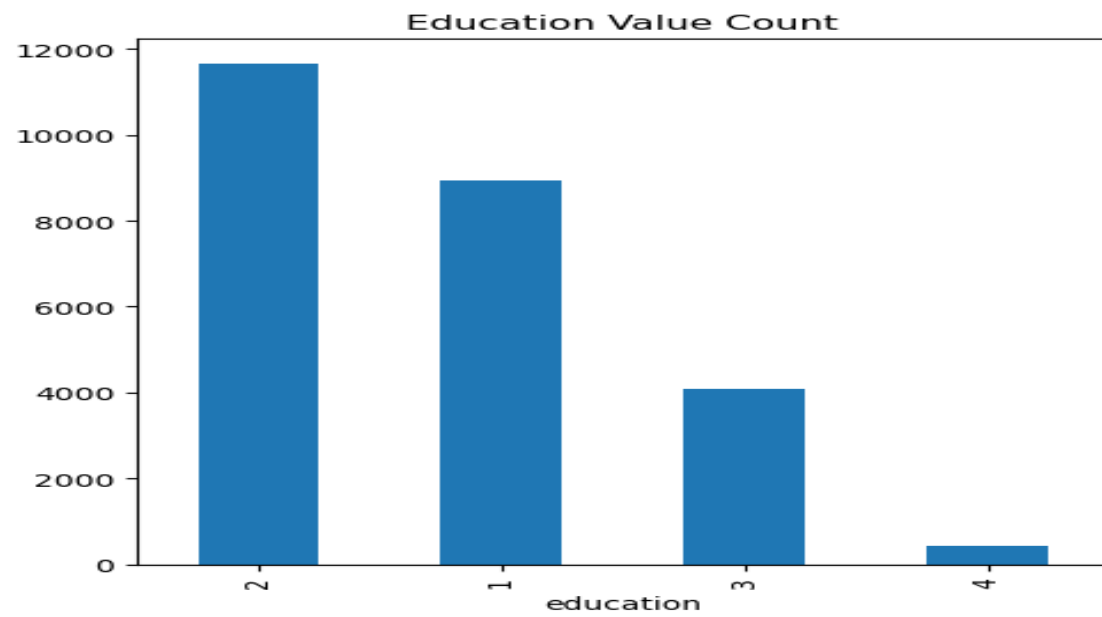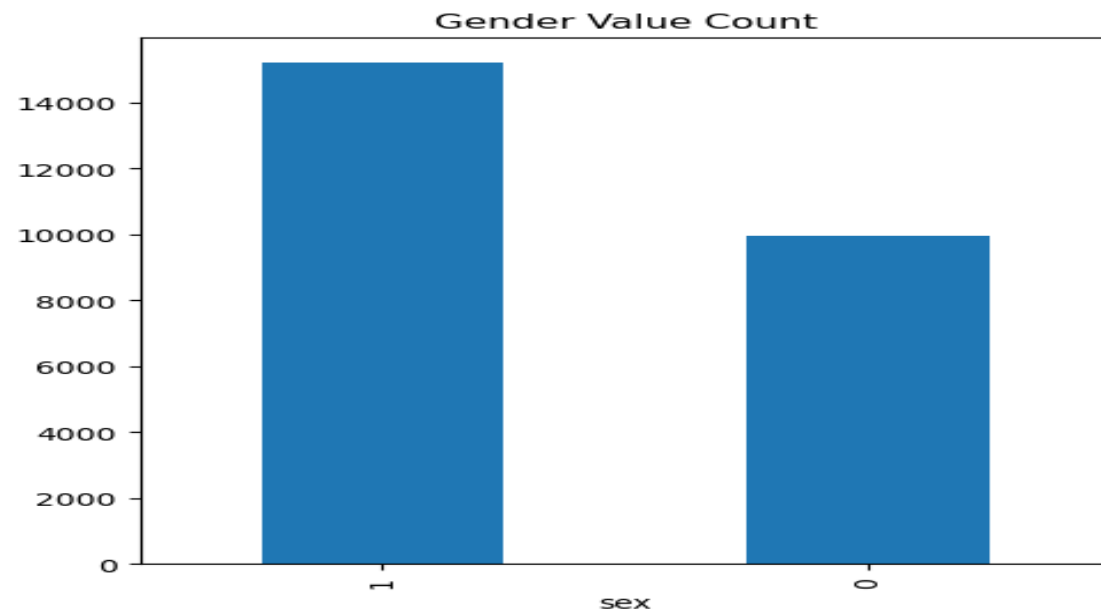
VIKRANT SINGH [22112114]

DATA EXPLORATION & CLEANING

- Dropped 0.499069% missing values in age as they were missing at random.

- No duplicated rows found in the dataset.

- Density plot of age remained unchanged before and after dropping missing values.
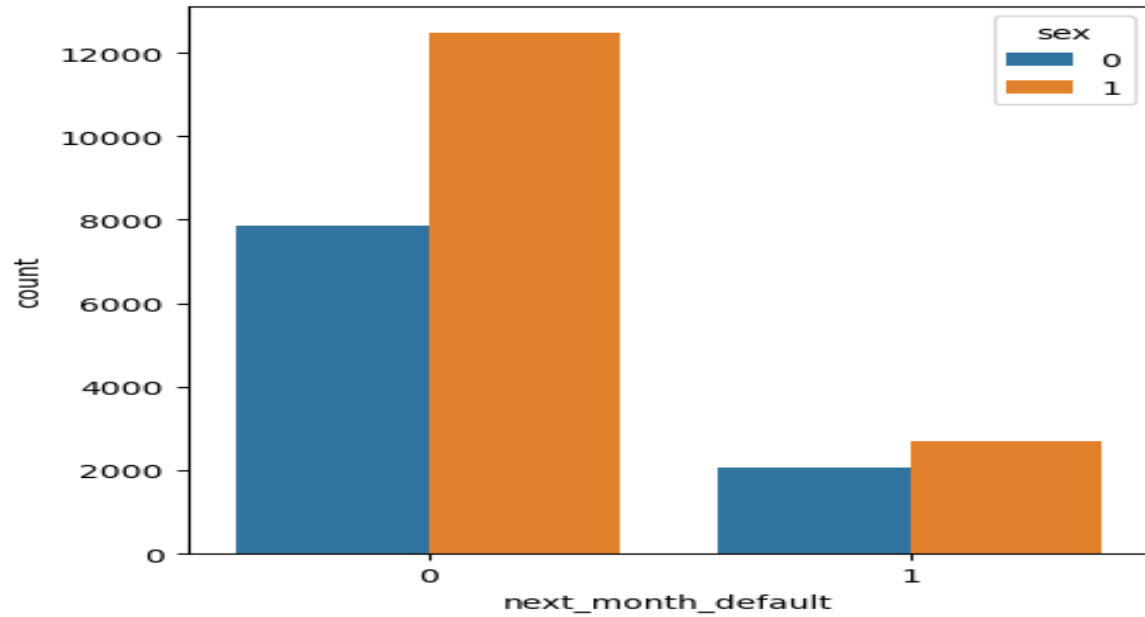
- Merged invalid education values (0, 5, 6 → 4 = Others)
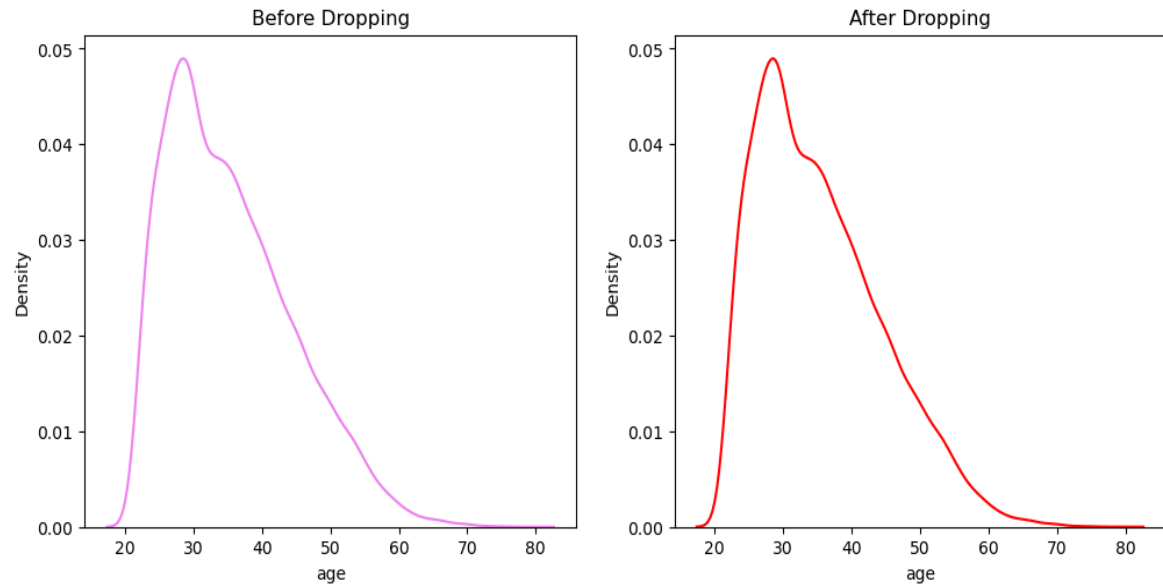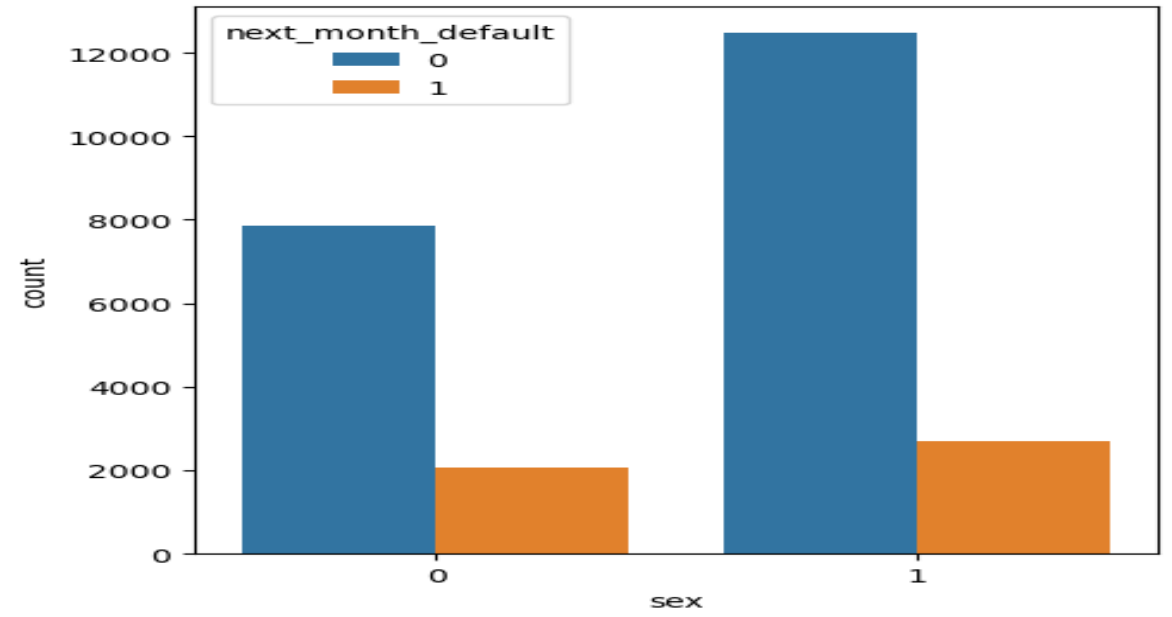
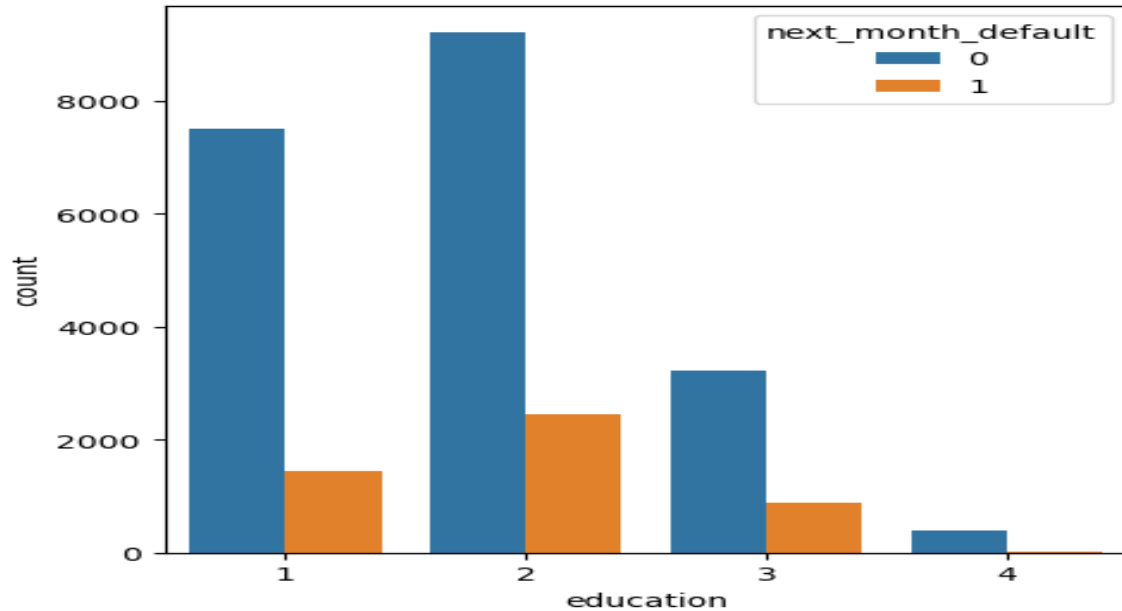# EXPLORATORY DATA ANALYSIS (EDA)
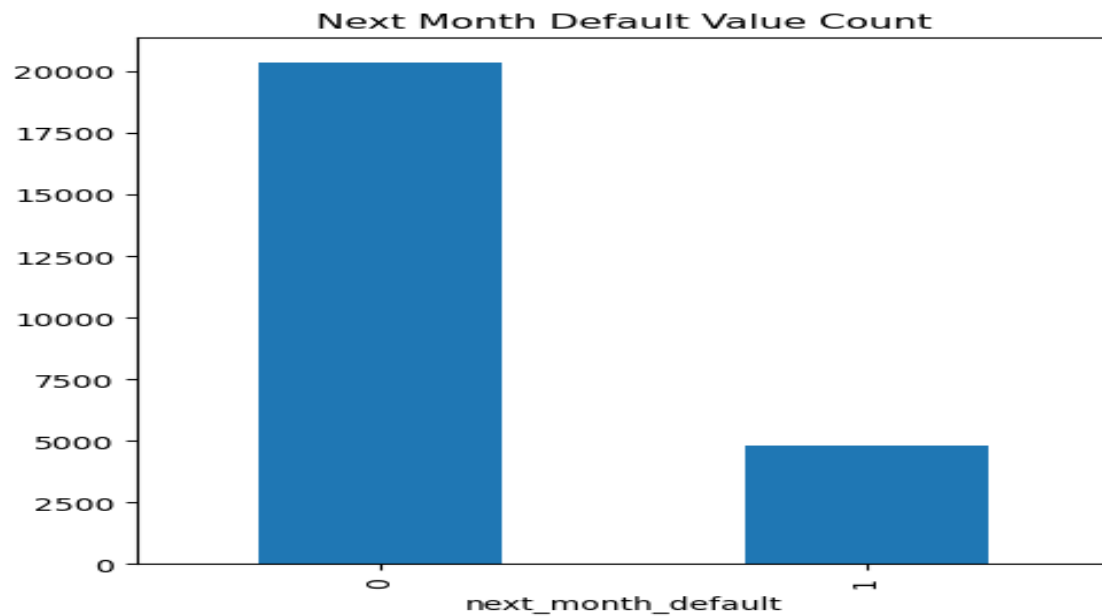
EXPLORATORY DATA ANALYSIS (EDA)

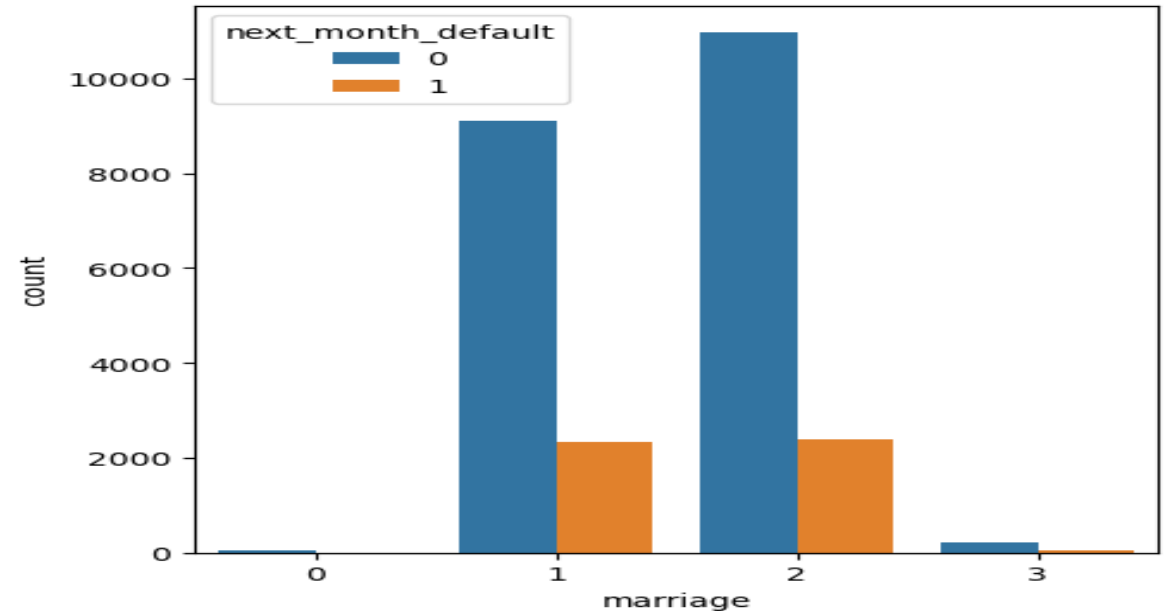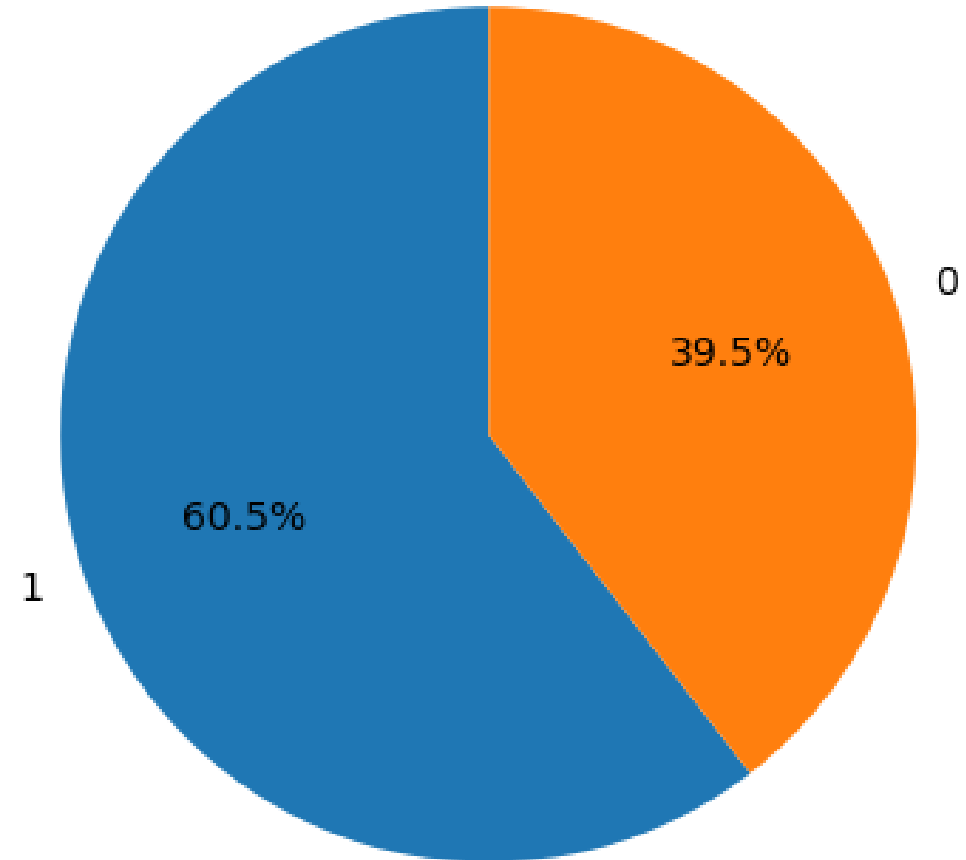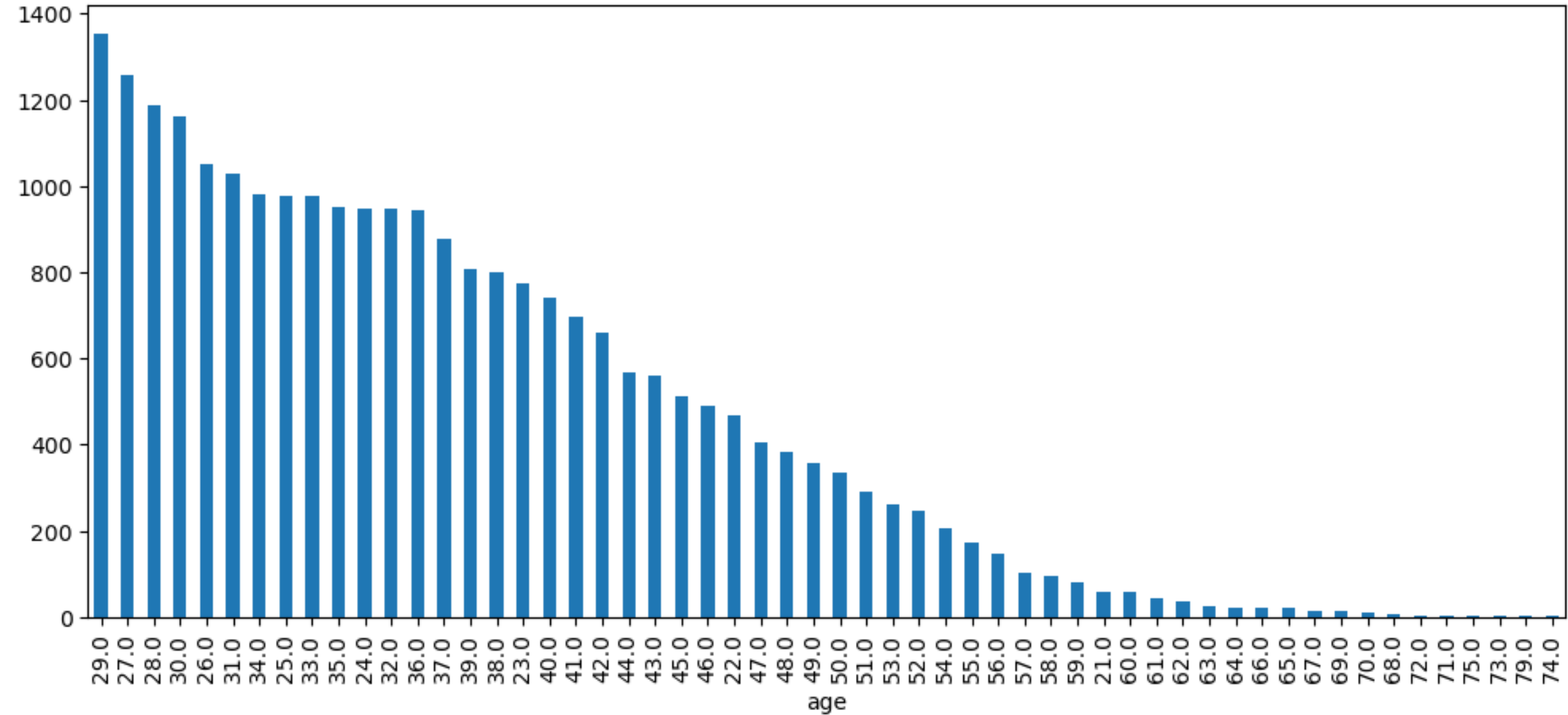# EXPLORATORY DATA ANALYSIS (EDA)

Next Month Default
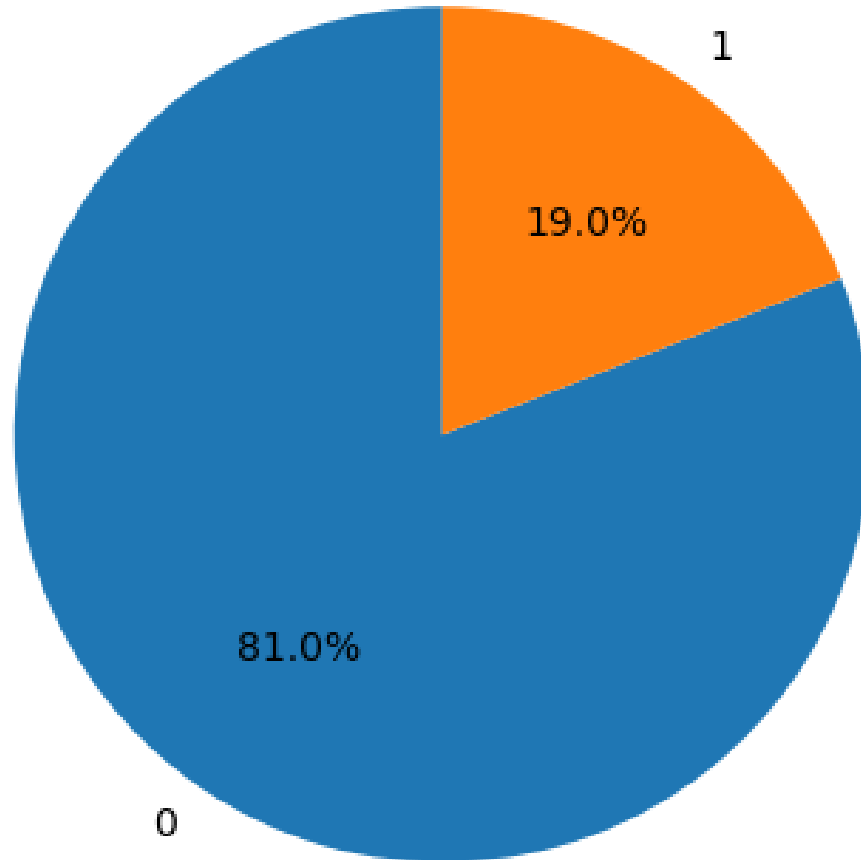
Sex

NUMBER OF PEOPLE VS AGE

FEATURE ENGINEERING

SOME FEATURES ARE CONSTRUCTED FOR BETTER RESULT

- **BILL_AMT**: AVERAGE BILL AMOUNT ACROSS PAST 6 MONTHS — REFLECTS OVERALL CARD USAGE BEHAVIOUR.

- **LIMIT_BAL**: CREDIT LIMIT ASSIGNED TO THE CUSTOMER — PROXY FOR INITIAL RISK ASSESSMENT BY THE BANK.

- **PAY_TO_BILL_RATIO**: RATIO OF TOTAL PAYMENTS MADE TO TOTAL BILLED AMOUNTS — INDICATES REPAYMENT DISCIPLINE.

- **AGE**: AGE OF THE CUSTOMER — CAPTURES LIFE STAGE AND POTENTIAL EARNING STABILITY.

- **AVG_MONTHLY_PAYMENT**: MEAN OF PAST 6 MONTHS' PAYMENTS — INDICATES REPAYMENT CAPACITY AND CONSISTENCY.

- **DELINQUENCY_PROBABILITY**: PREDICTED PROBABILITY OF PAST DELINQUENCY FROM AUXILIARY MODEL — EARLY WARNING SIGNAL.

- **DUE_1 TO DUE_6**: DUES FOR EACH OF THE PAST 6 MONTHS — TRACKS ROLLING OVERDUE PATTERNS.

- **EDUCATION**: EDUCATIONAL BACKGROUND — USED AS A SOCIOECONOMIC INDICATOR.

- **MARRIAGE**: MARITAL STATUS — POTENTIAL PROXY FOR FINANCIAL DEPENDENTS OR OBLIGATIONS.

- **PAY_0 TO PAY_6**: HISTORICAL PAYMENT STATUS VARIABLES — KEY SIGNALS OF PAST DELINQUENCY BEHAVIOUR.

- **SEX**: GENDER — INCLUDED AS A DEMOGRAPHIC VARIABLE FOR BEHAVIUORAL SEGMENTATION.

- **TOTAL_DUES**: TOTAL OUTSTANDING BALANCE AT CURRENT TIME — REFLECTS CREDIT RISK EXPOSURE

DEALING WITH CLASS IMBALANCE
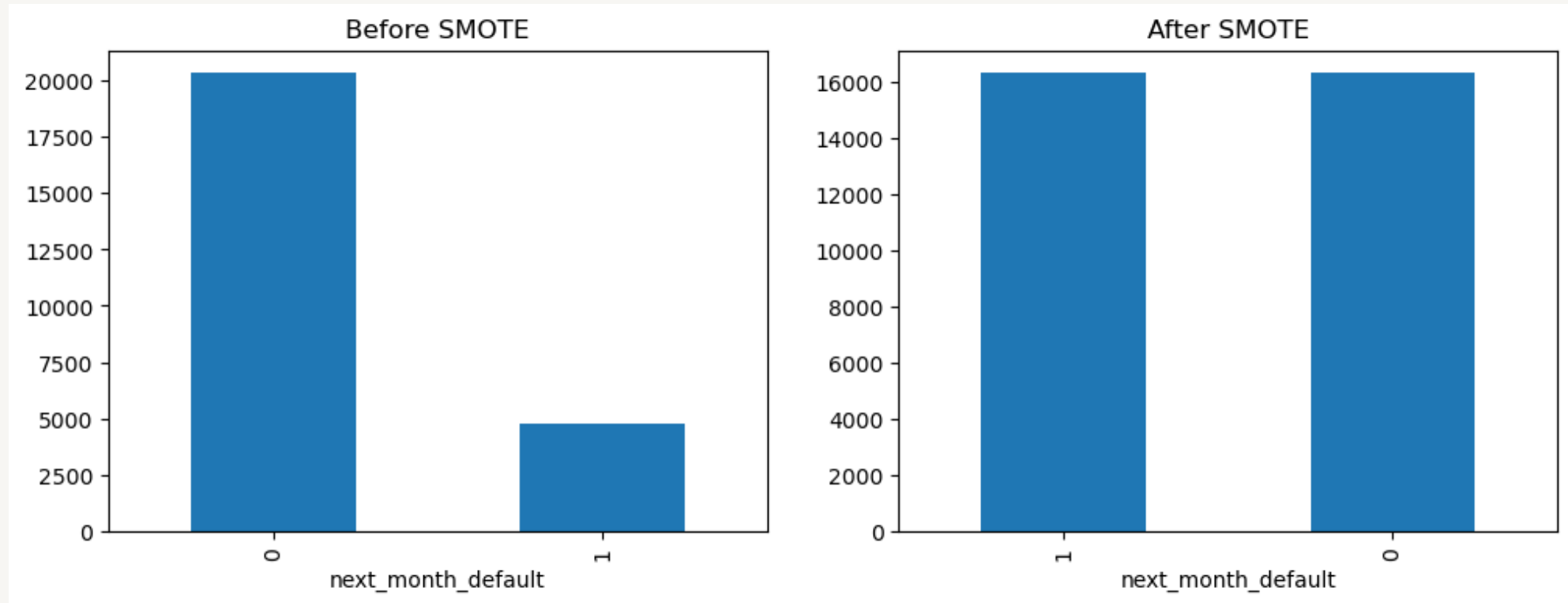
MODEL SELECTION & HYPERPARAMETER TUNING

ALGORITHM CHOSEN FOR TUNING

Random Forest, AdaBOOST, Gradient Boosting, Logistic Regression, XGBOOST, LIGHTBGM

MODEL FOR TUNING

GridSearchCV is used for Hyperparameter Tuning

NUMBER OF CV

5 Cross Validations are performed for each candidate

SCORING

F1 Score is used as scoring metric to perform Cross Validations because it balances the trade-off between Precision & Recall

RANDOM FOREST

RANDOM FOREST PERFORMED BEST OUT OF ALL MODEL BASED ON F1 SCORE

CLASSIFICATION REPORT ON TEST DATA

	precision	recall	f1-score	support
0	0.88	0.88	0.88	4048
1	0.51	0.52	0.51	977
accuracy			0.81	5025
macro avg	0.70	0.70	0.70	5025
weighted avg	0.81	0.81	0.81	5025

CLASSIFICATION REPORT ON TRAIN DATA

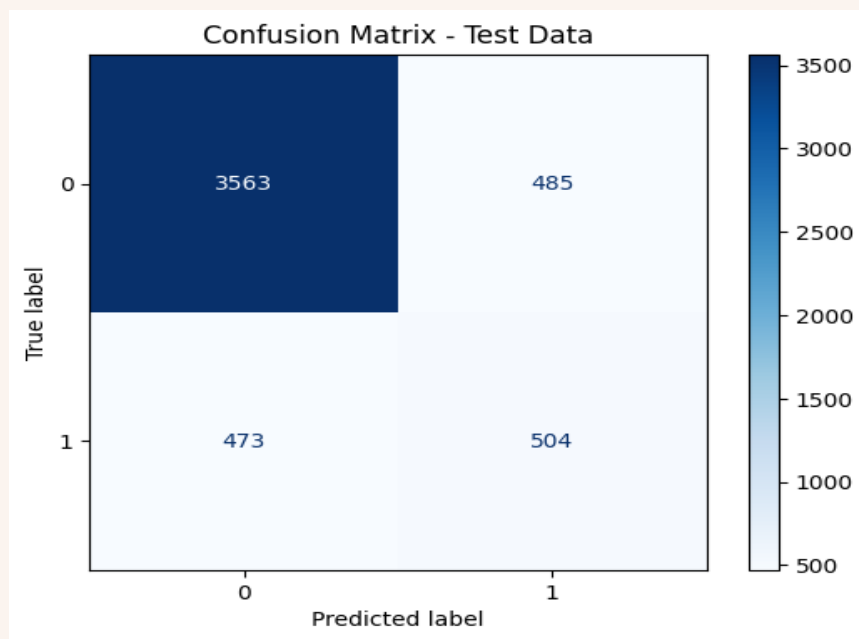|	precision	recall	f1-score	support
0	0.91	0.90	0.91	16289
1	0.59	0.62	0.61	3807
accuracy			0.85	20096
macro avg	0.75	0.76	0.76	20096
weighted avg	0.85	0.85	0.85	20096

F2 SCORE

F2 Score (Test Set): 0.5146

F2 Score (Train Set): 0.6149

CONFUSION MATRIX

TEST DATA

TRAIN DATA

ROC CURVE

Receiver Operating Characteristic - Test Set
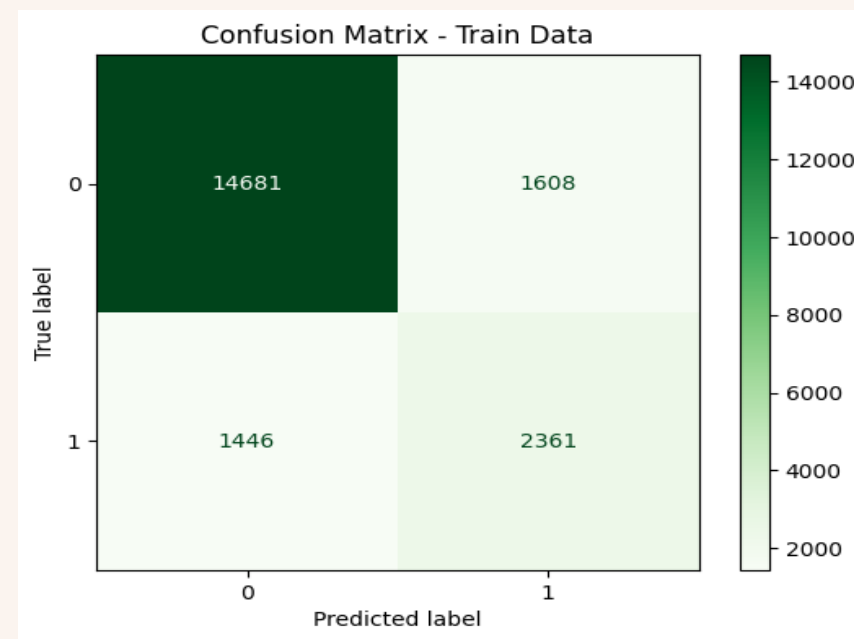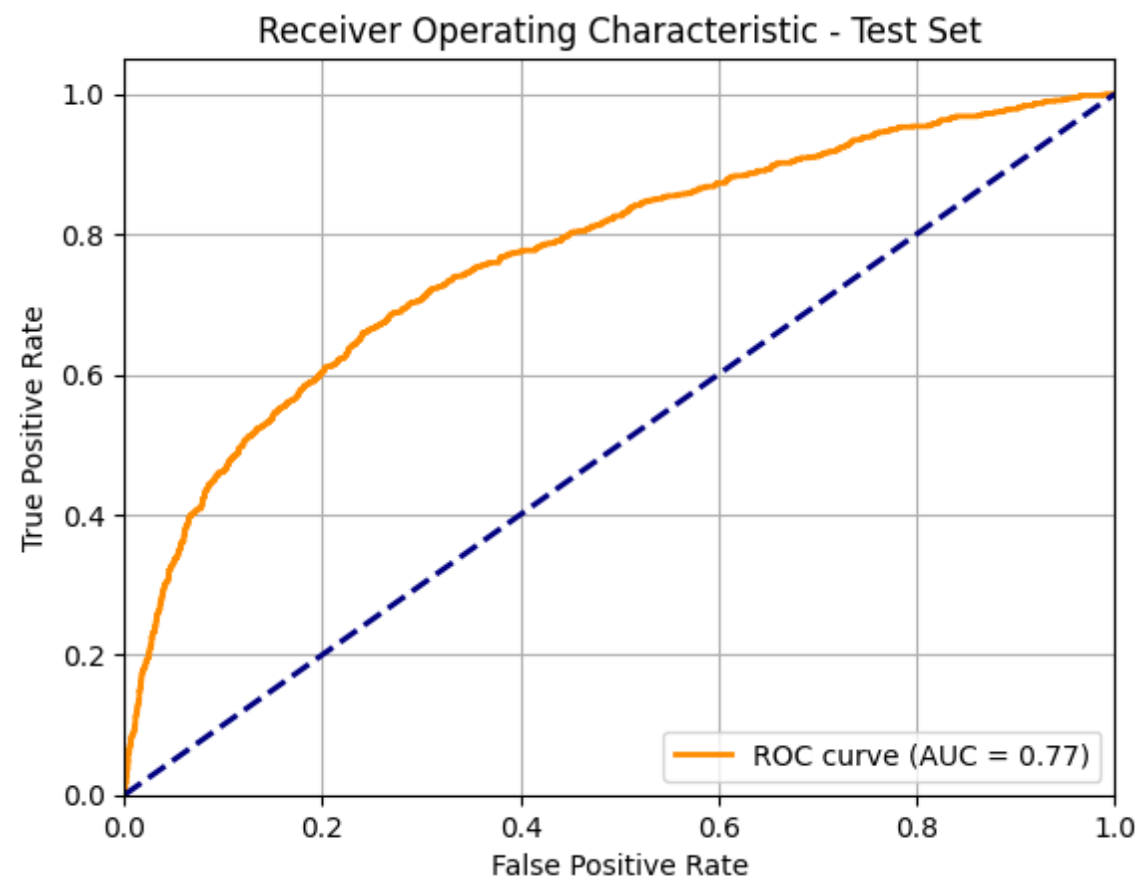
HYPERPARAMETER TUNING OF RANDOM FOREST FOR ACHIEVING BEST F2 SCORE

CLASSIFICATION REPORT ON TEST DATA

|	precision	recall	f1-score	support
0	0.90	0.78	0.84	4048
1	0.42	0.63	0.50	977
accuracy			0.76	5025
macro avg	0.66	0.71	0.67	5025
weighted avg	0.80	0.76	0.77	5025

CLASSIFICATION REPORT ON TEST DATA

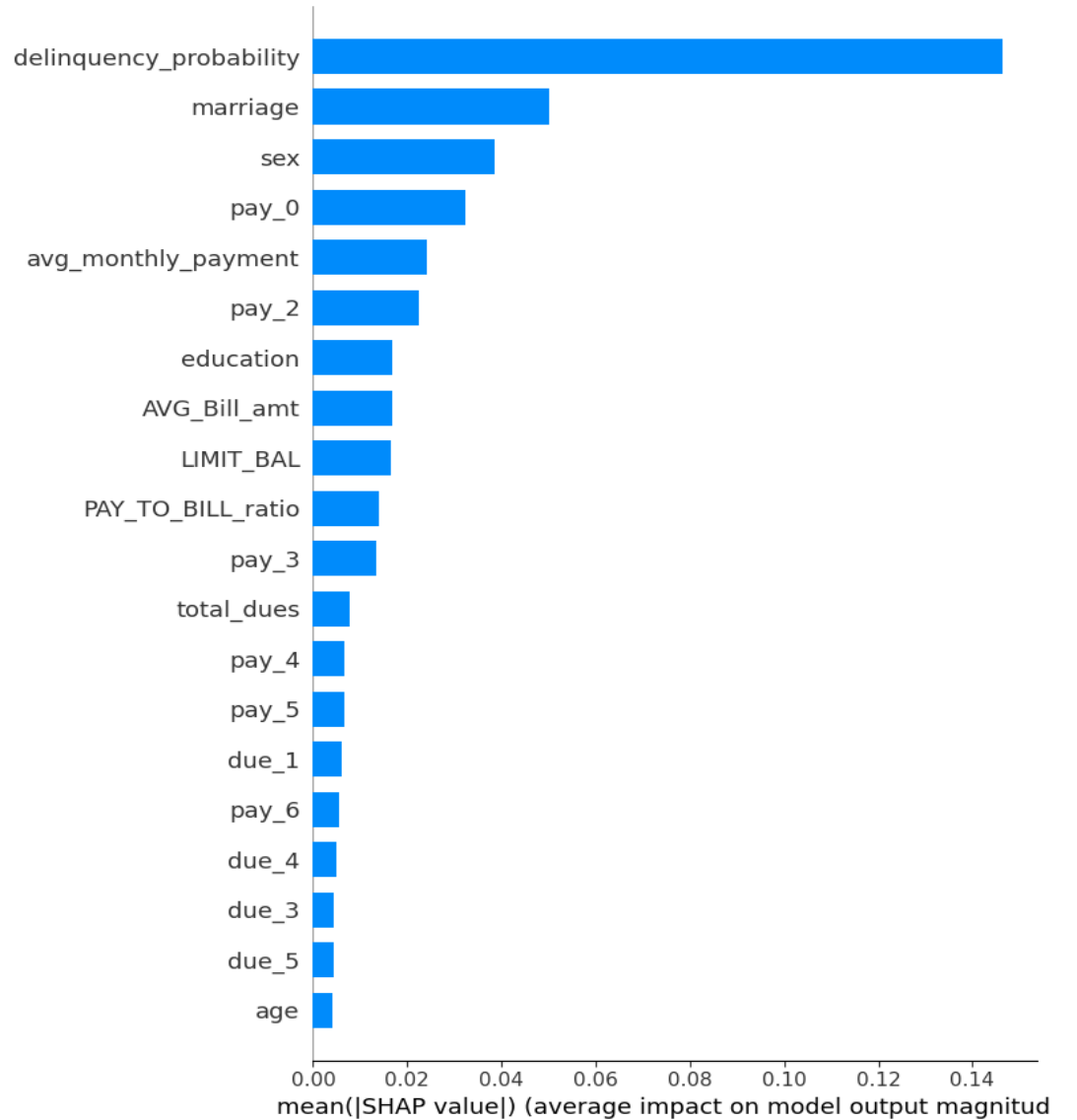|	precision	recall	f1-score	support
0	0.90	0.77	0.83	16289
1	0.39	0.64	0.48	3807
accuracy			0.74	20096
macro avg	0.65	0.70	0.66	20096
weighted avg	0.80	0.74	0.76	20096

F2 SCORE

F2 Score On Test Set: 0.5740740740740741

F2 Score On Test Set:0.7882184005873373

SHAP CURVE
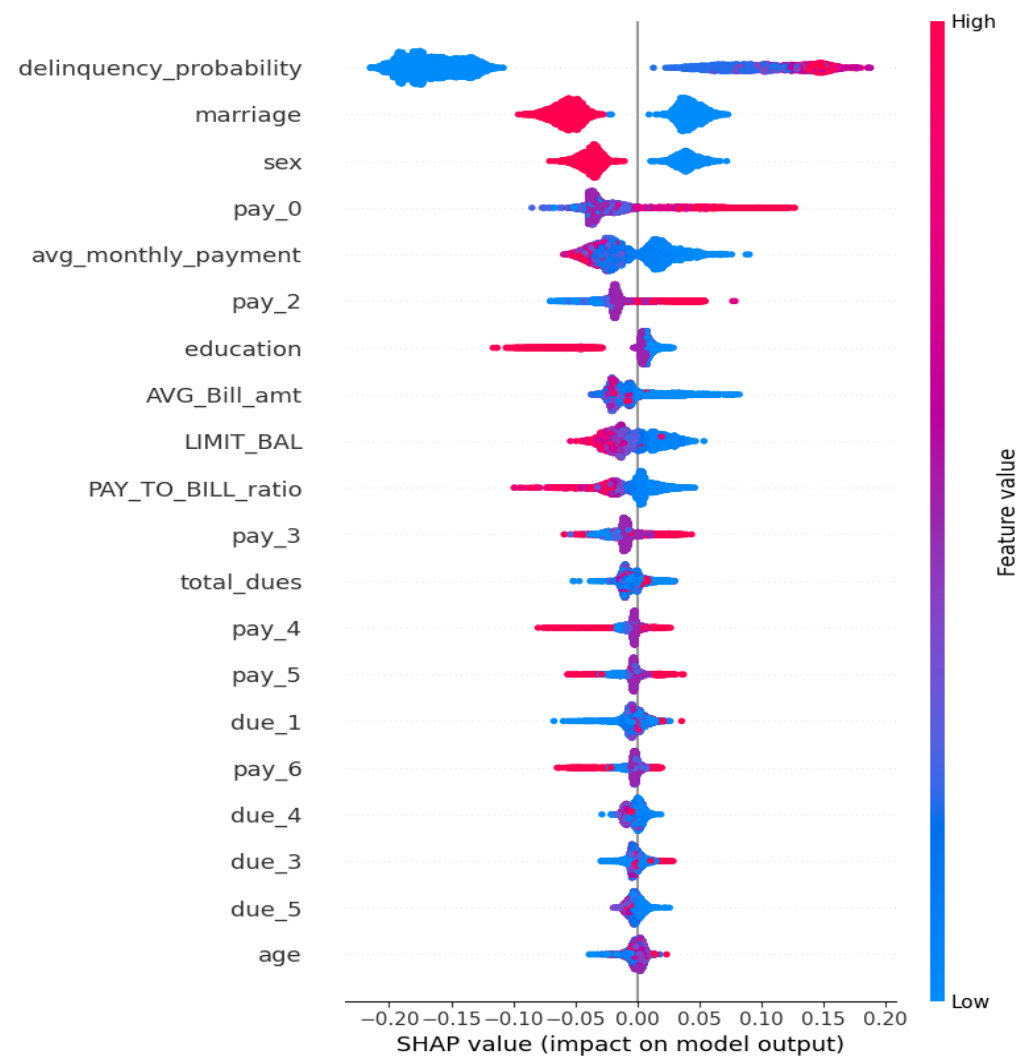
BAR PLOT FOR FEATURE IMPORTANCE

BUSINESS IMPLICATIONS

- Enables early identification of high-risk customers to reduce credit losses.

- Supports targeted interventions like credit limit reductions and payment reminders.

- Improves profitability by minimizing defaults while preserving good customer relationships.

- Informs personalized credit strategies based on customer behaviour.

- Strengthens regulatory compliance through data-driven credit risk modelling.

THANK YOU

VIKRANT SINGH

22112114

CHE 4Y [B.TECH.]

vikrant_s@ch.iitr.ac.in