Risk Radar :
Forecasting Credit Card Default

Ting Siou Chen

Outline

1. Motivation & Goal
2. Flowchart
3. Exploratory Data Analysis (EDA)
4. Result

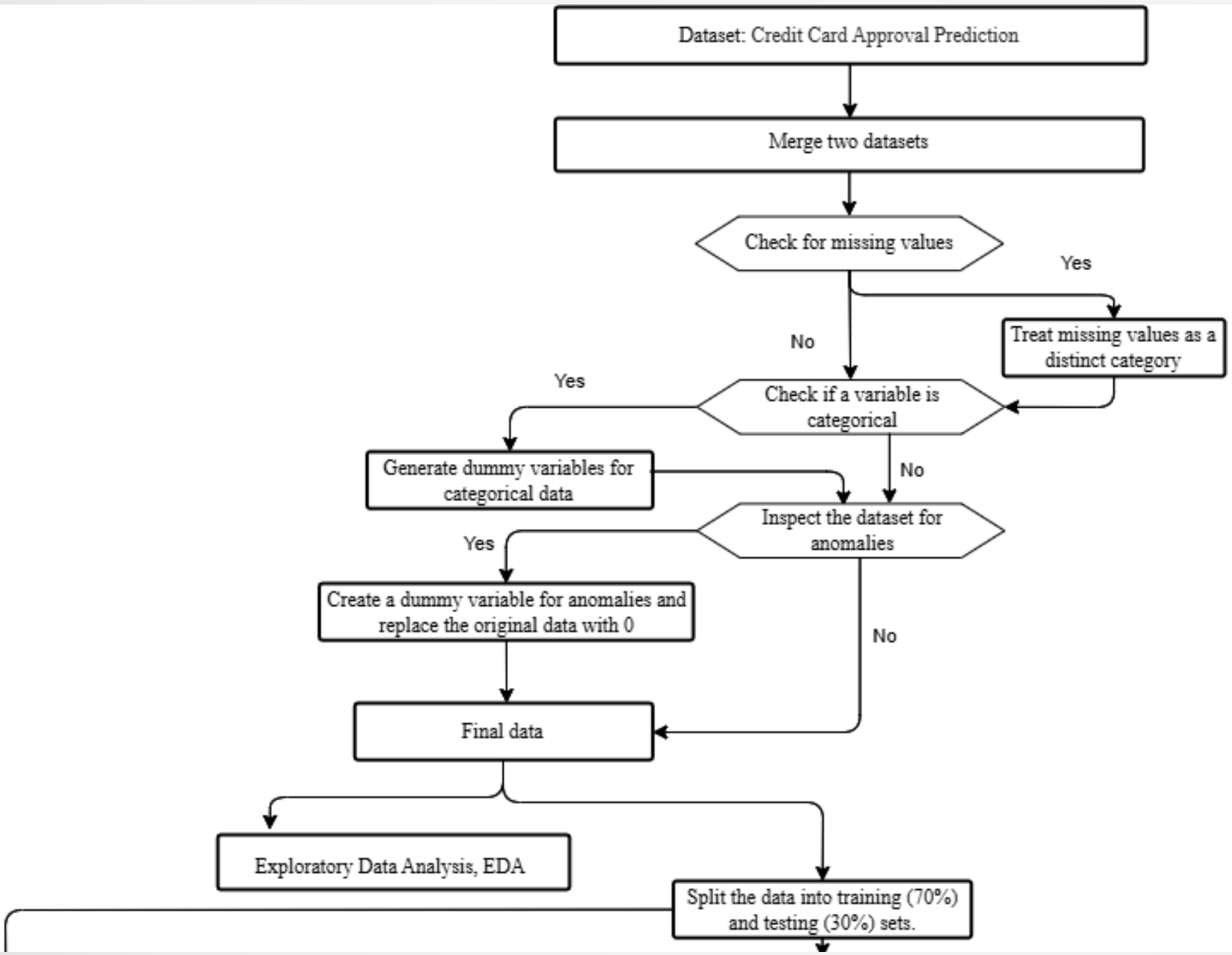
Motivation

- Risk management and loss prevention
- Efficient resource allocation
- Customer segmentation and tailored services

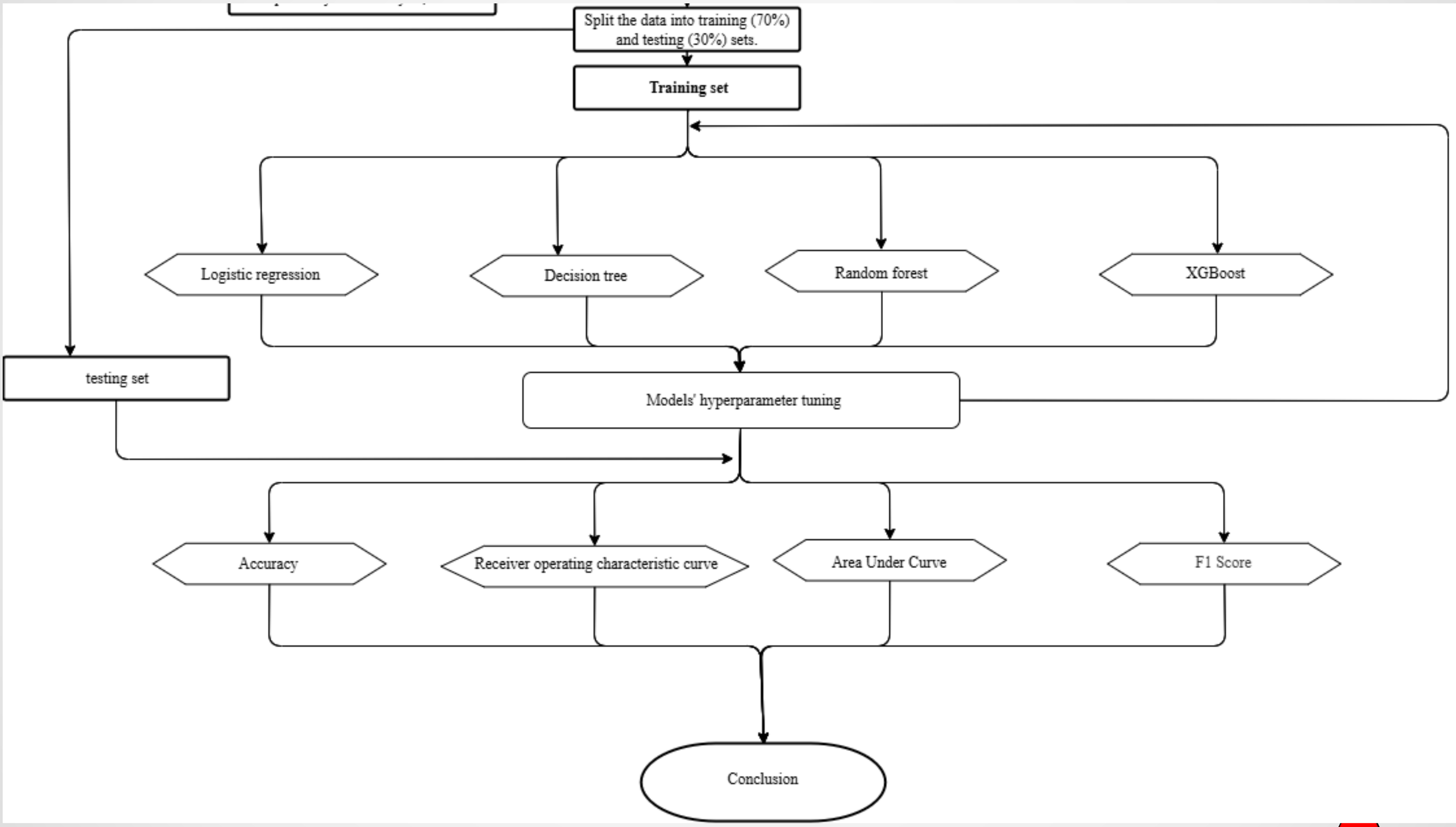
Goal

- Develop a predictive model
- Accurately identifies customers with a high risk of default
- Allowing financial institutions to minimize losses and optimize credit strategies.

Flowchart


Dataset: Credit Card Approval Prediction
Merge two datasets
Check for missing values
Yes
Treat missing values as a distinct category
No
Check if a variable is categorical
Yes
Generate dummy variables for categorical data
No
Inspect the dataset for anomalies
Yes
Create a dummy variable for anomalies and replace the original data with 0
No
Final data
Exploratory Data Analysis, EDA
Split the data into training (70%) and testing (30%) sets.

Flowchart


Split the data into training (70%) and testing (30%) sets.
Training set
Logistic regression
Decision tree
Random forest
XGBoost
testing set
Models' hyperparameter tuning
Accuracy
Receiver operating characteristic curve
Area Under Curve
F1 Score
Conclusion

EDA-data

- Dataset

  Contains demographic and credit behavior data from credit card applicants, divided into : **application_record.csv**, **credit_record.csv**

- Source

  Kaggle - Credit Card Approval Prediction Dataset

EDA-data

- Sample Size (n): 36457
- Number of Features (p): 41 features
- Data Categories

  Demographics, Financial status, Past credit behaviors

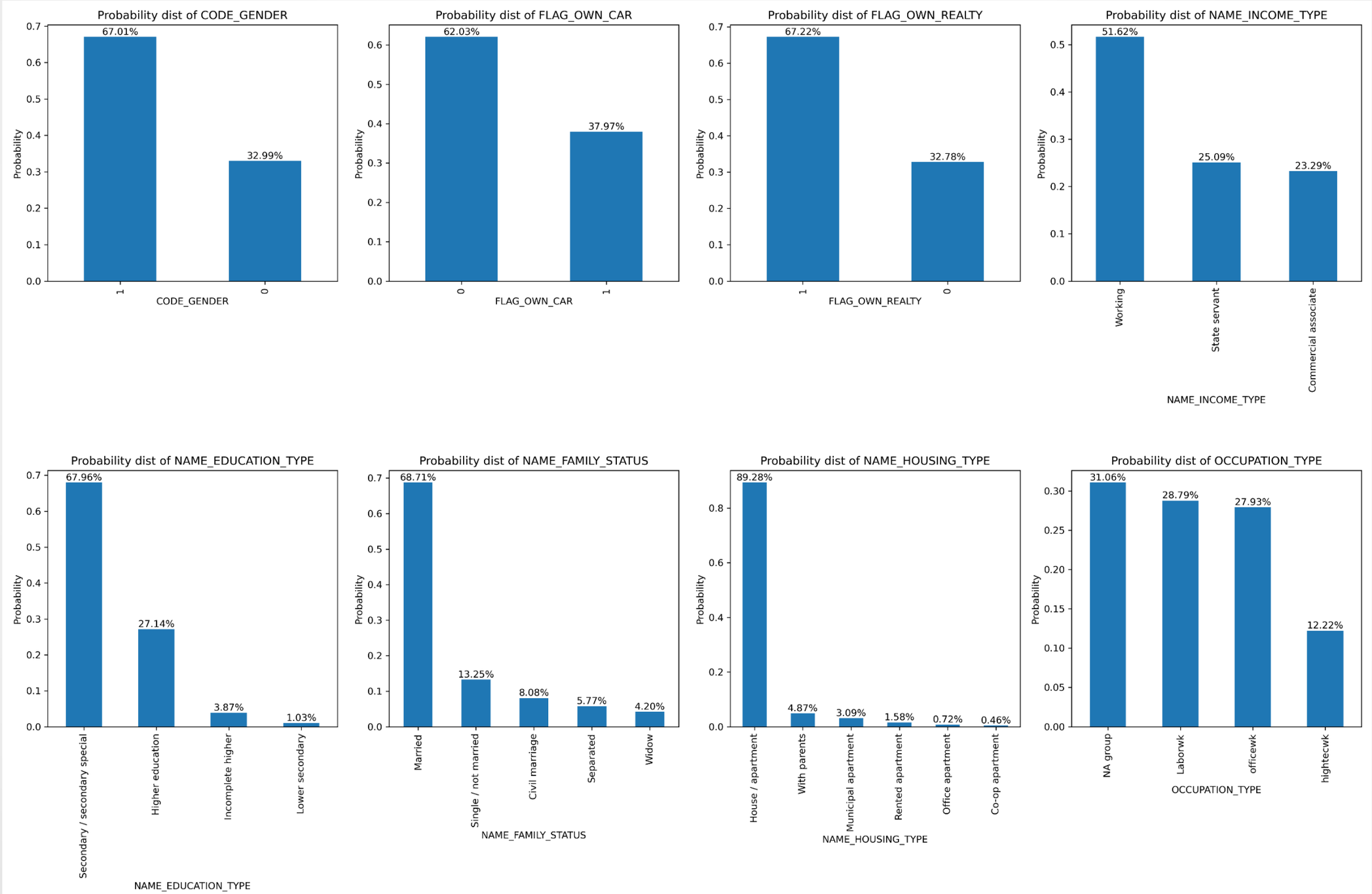
EDA-data

Table 1: Description of Variables

Variable	Description
ID	Client number
CODE_GENDER	Gender (Male=0, Female=1)
FLAG_OWN_CAR	Is there a car (No=0, Yes=1)
FLAG_OWN_REALTY	Is there a property (No=0, Yes=1)
AMT_INCOME_TOTAL	Annual income
LOG_AMT_INCOME_TOTAL	The logarithm of annual income
NAME_INCOME_TYPE	Income category
NAME_EDUCATION_TYPE	Education level
NAME_FAMILY_STATUS	Marital status
NAME_HOUSING_TYPE	Way of living
AGE	Clients' age
LOG_AGE	The logarithm of clients' age
YEARS_EMPLOYED	Duration of the most recent job in years
LOG_YEARS_EMPLOYED	The logarithm of duration of the most recent job in years
FLAG_WORK_PHONE	Is there a work phone (No=0, Yes=1)
FLAG_PHONE	Is there a phone (No=0, Yes=1)
FLAG_EMAIL	Is there an email (No=0, Yes=1)
OCCUPATION_TYPE	Occupation
CNT_FAM_MEMBERS	Family size
MONTHS_BALANCE	Record month (The month of extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, etc.)
STATUS	Status (-1: No loan for the month; 0: paid off that month;1: 1-29 days past due; 2: 30-59 days past due; 3: 60-89 days overdue; 4: 90-119 days overdue; 5: 120-149 days overdue; 6: Overdue or bad debts, write-offs for more than 150 days)
DEFAULT	Whether a customer is in default(DEFAULT = 1 when STATUS $\geq$ 1 (1+ days overdue); otherwise, DEFAULT = 0)

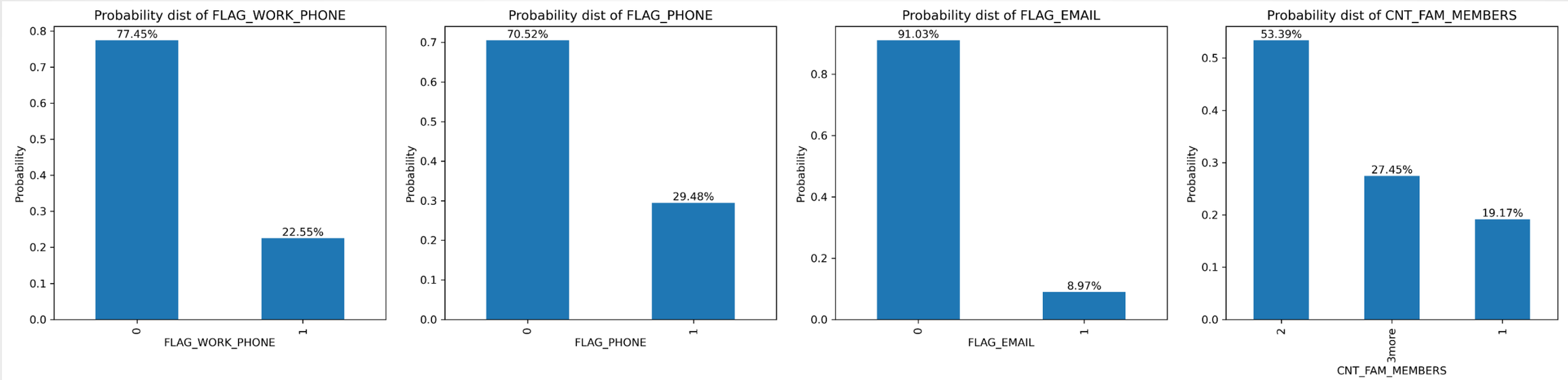
EDA

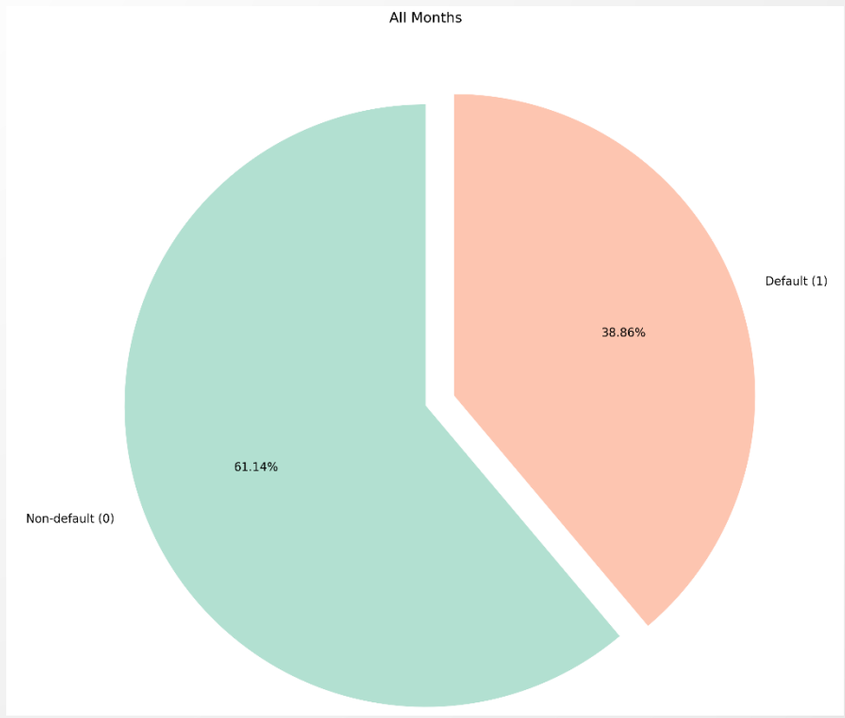
Categorical variable



EDA

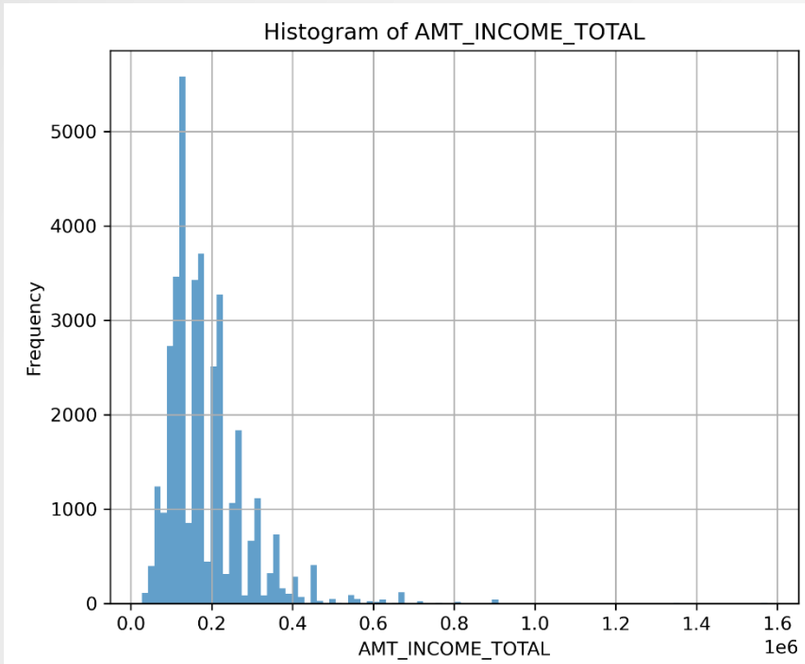
Categorical variable

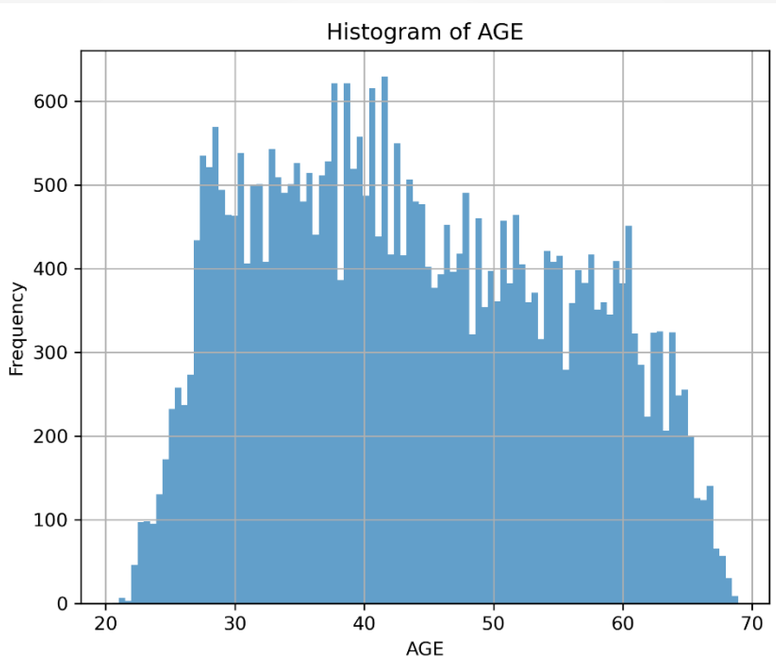


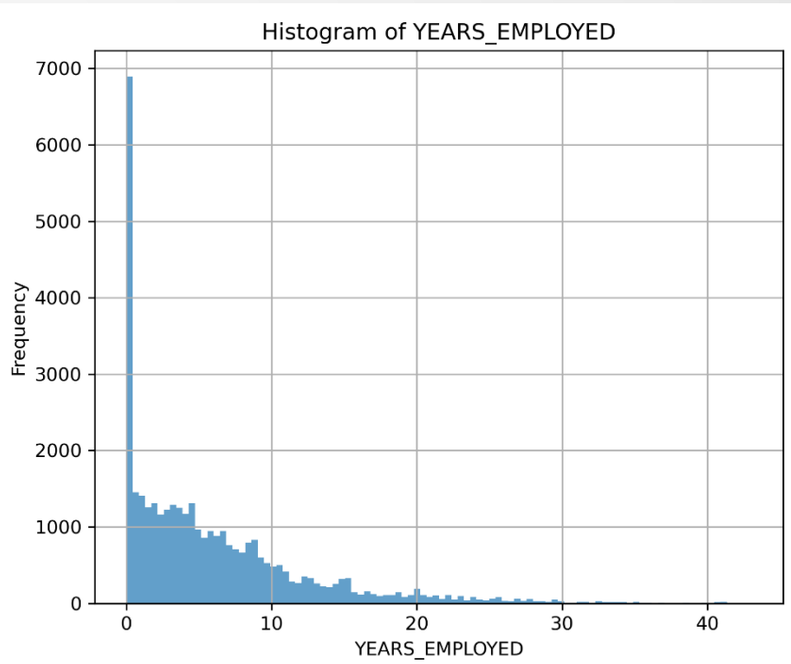


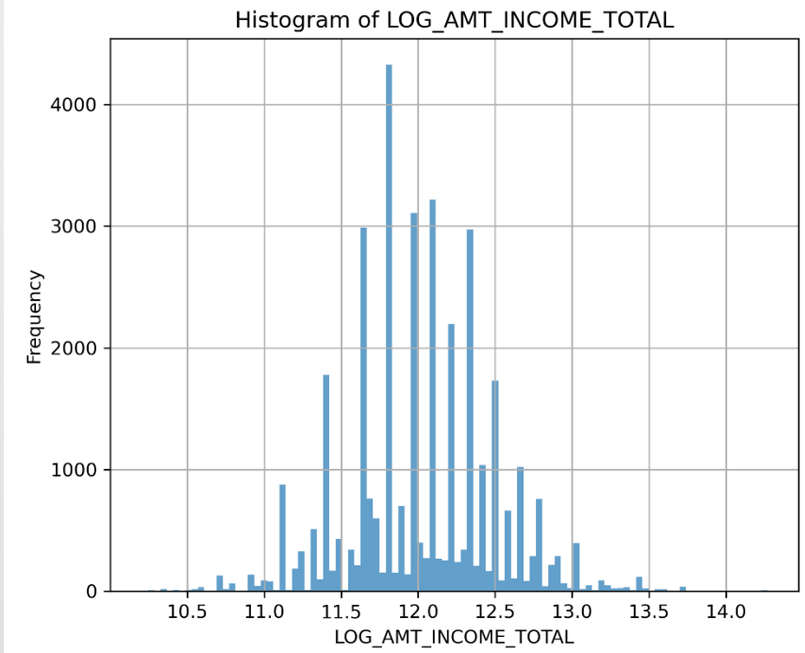
EDA

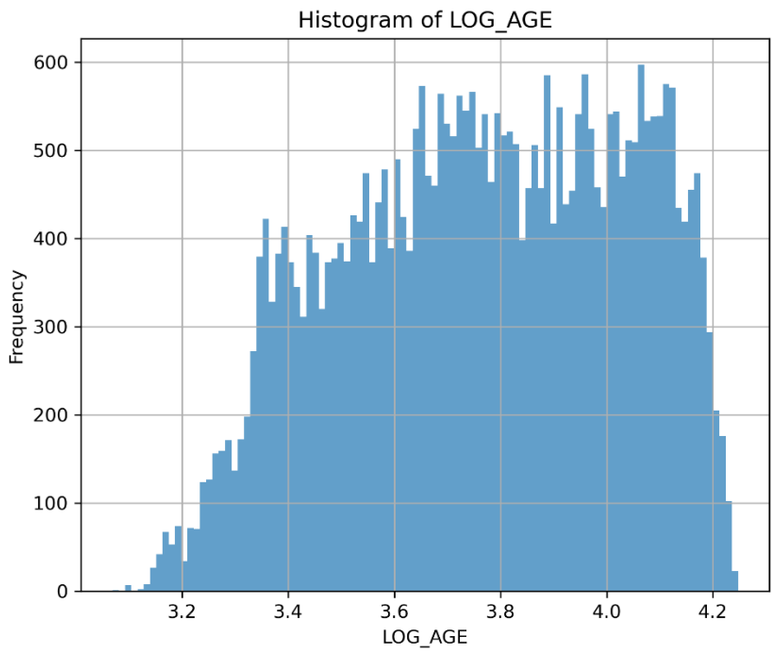
Continuous variable

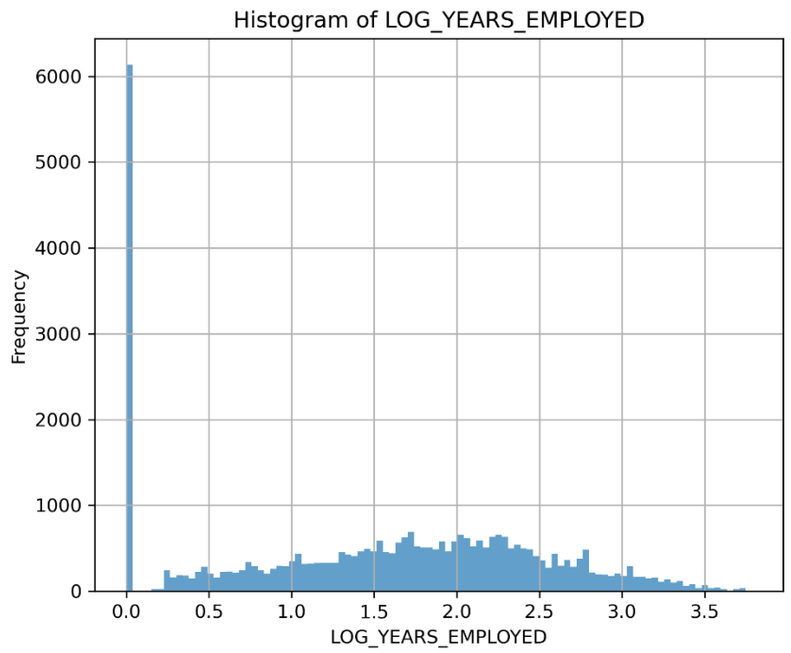






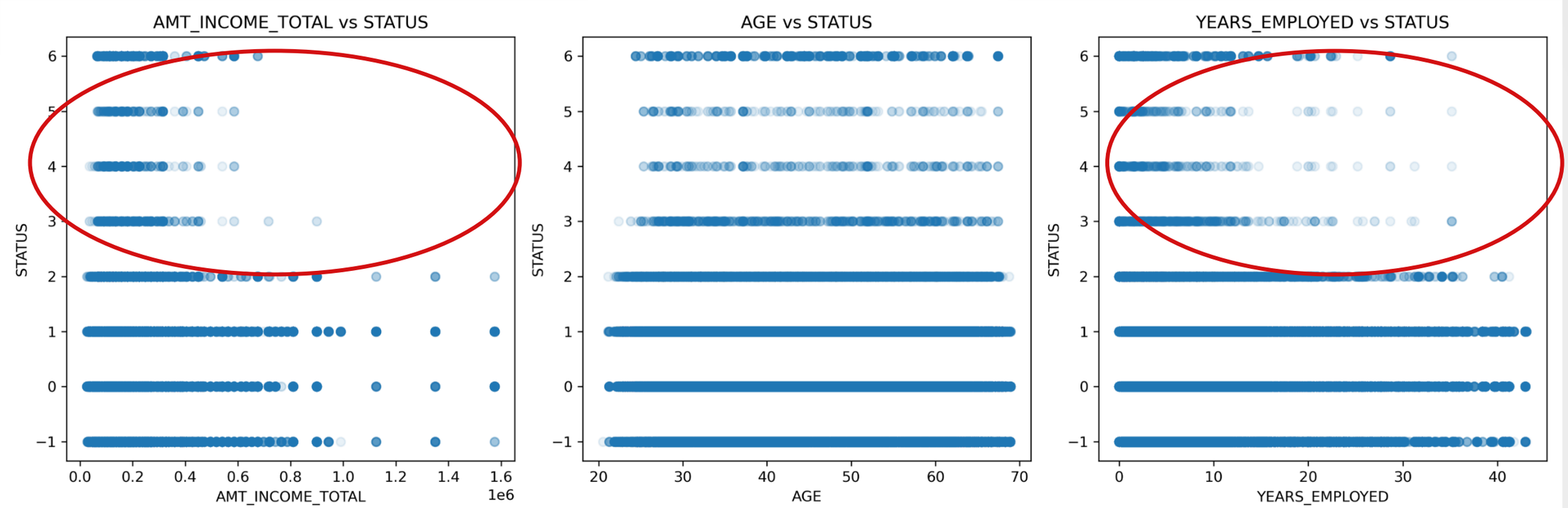


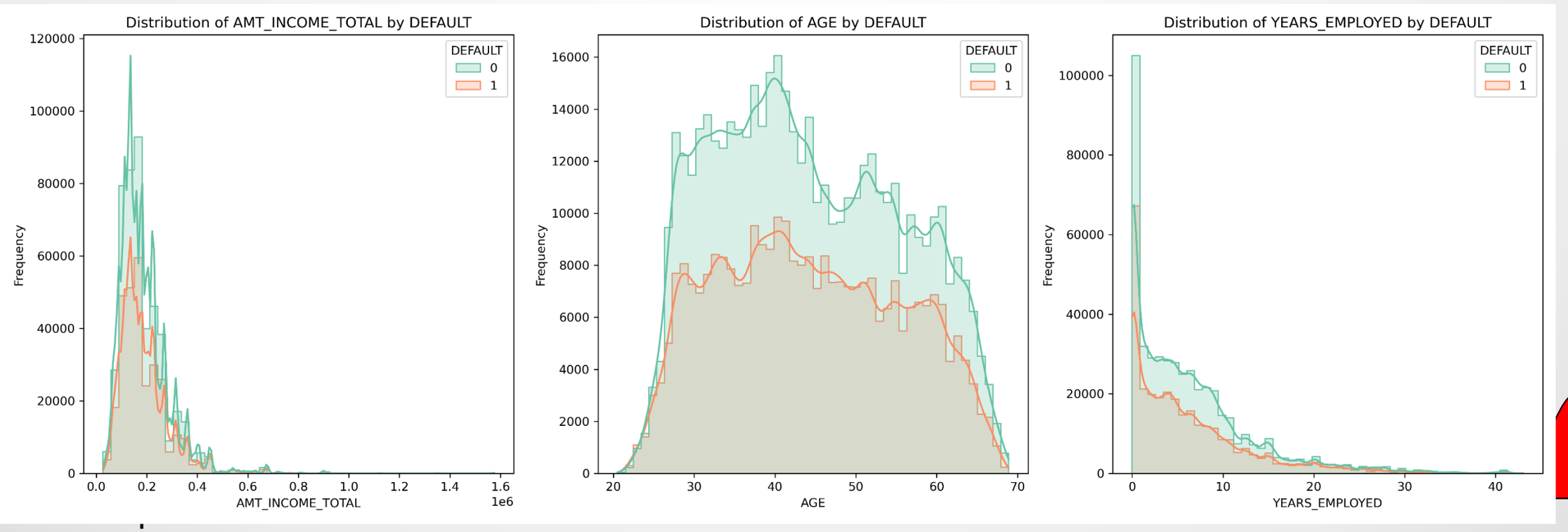




EDA

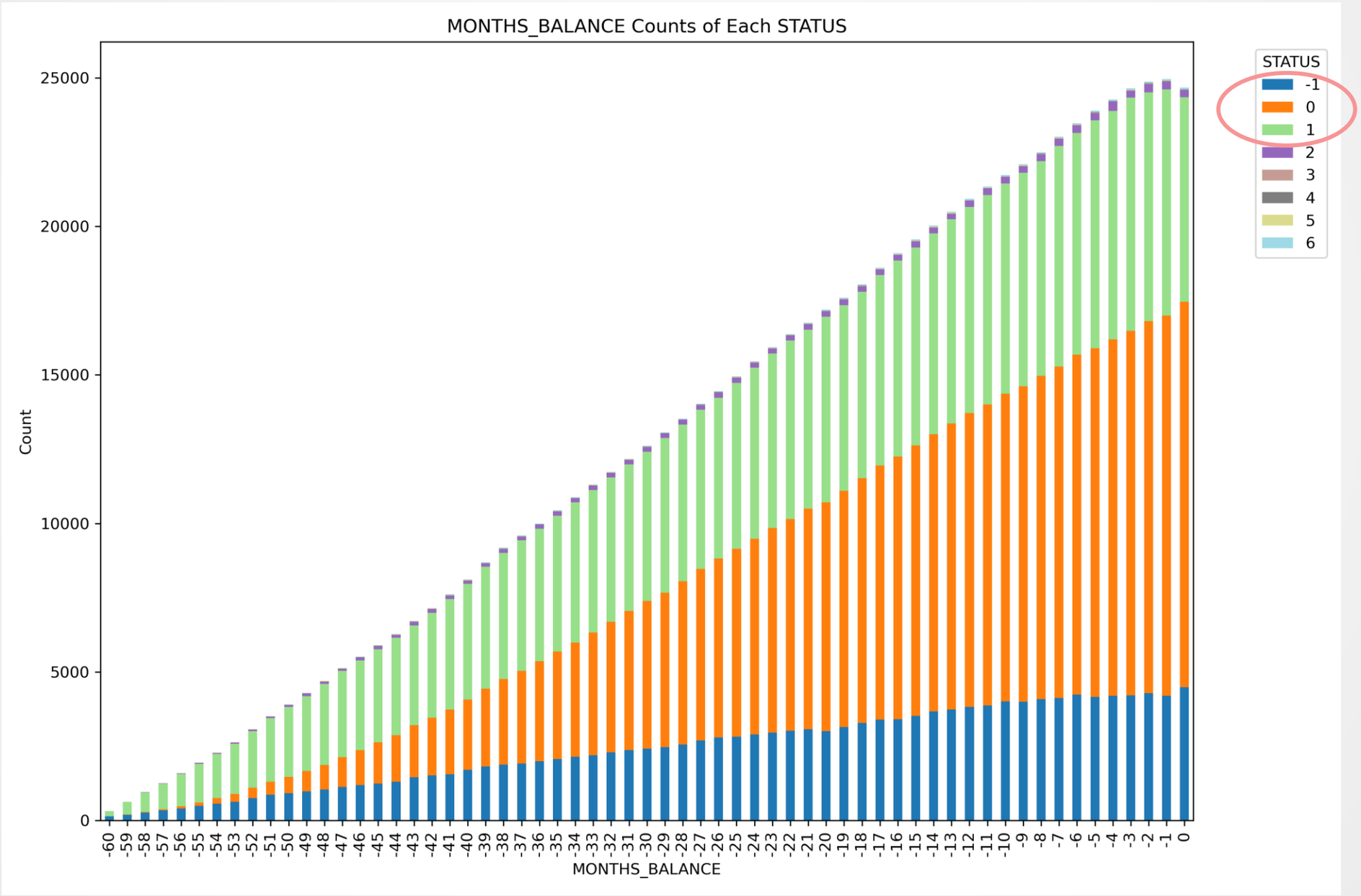
- STATUS/DEFAULT v.s. continuous variables.





EDA

Stacked Bar Chart of STATUS by MONTHS BALANCE

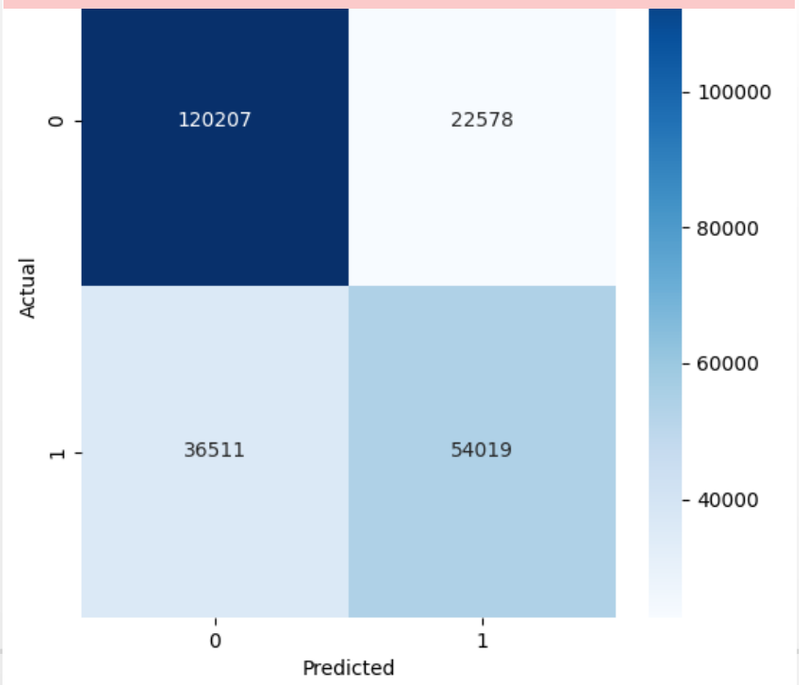


Result - Confusion matrix

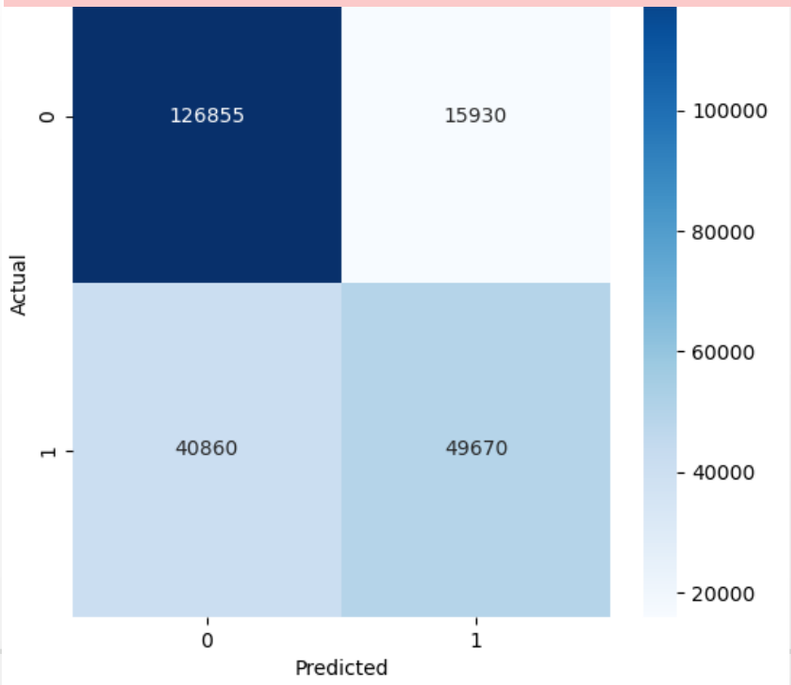
Logistic regression

Decision tree

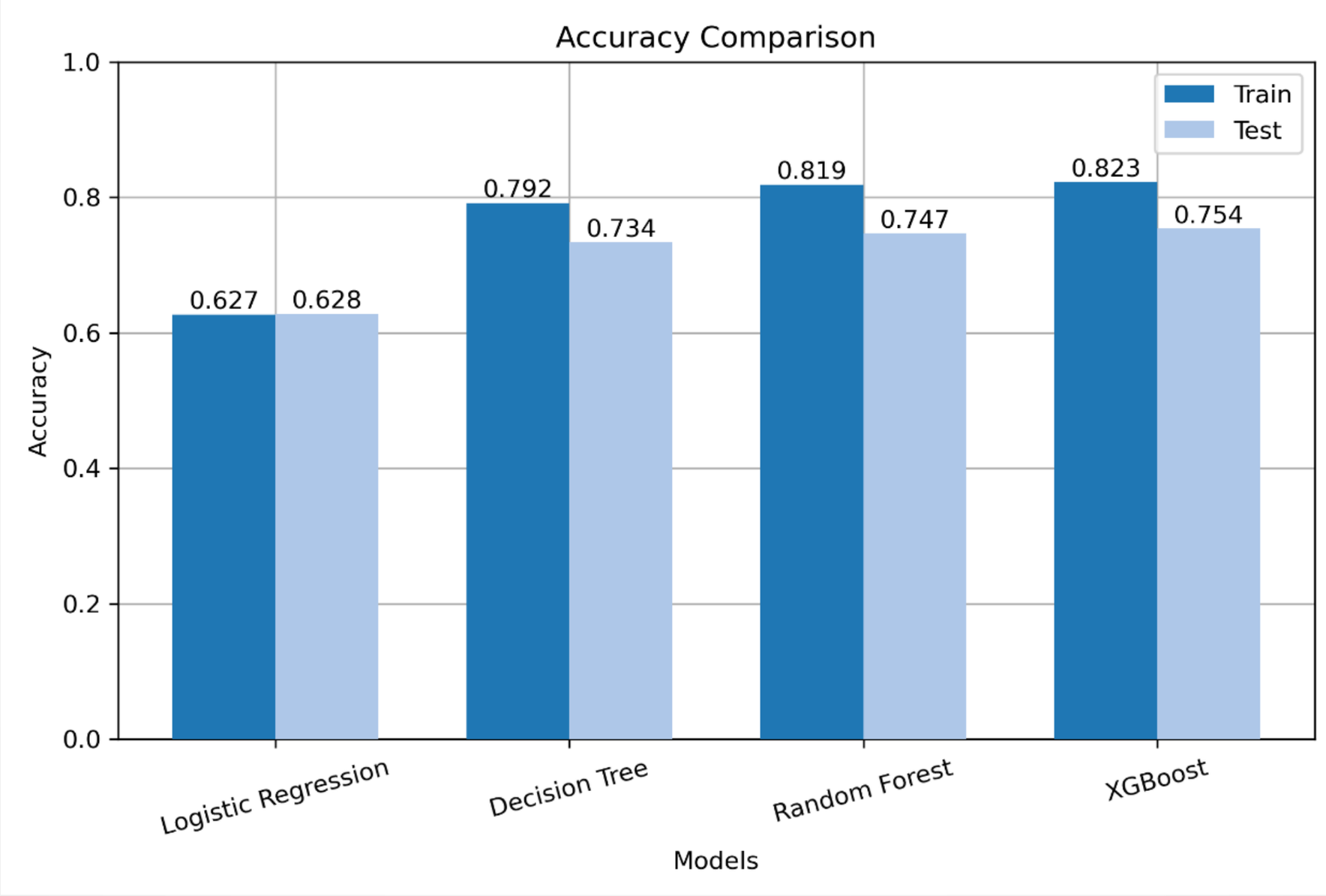
Random forest



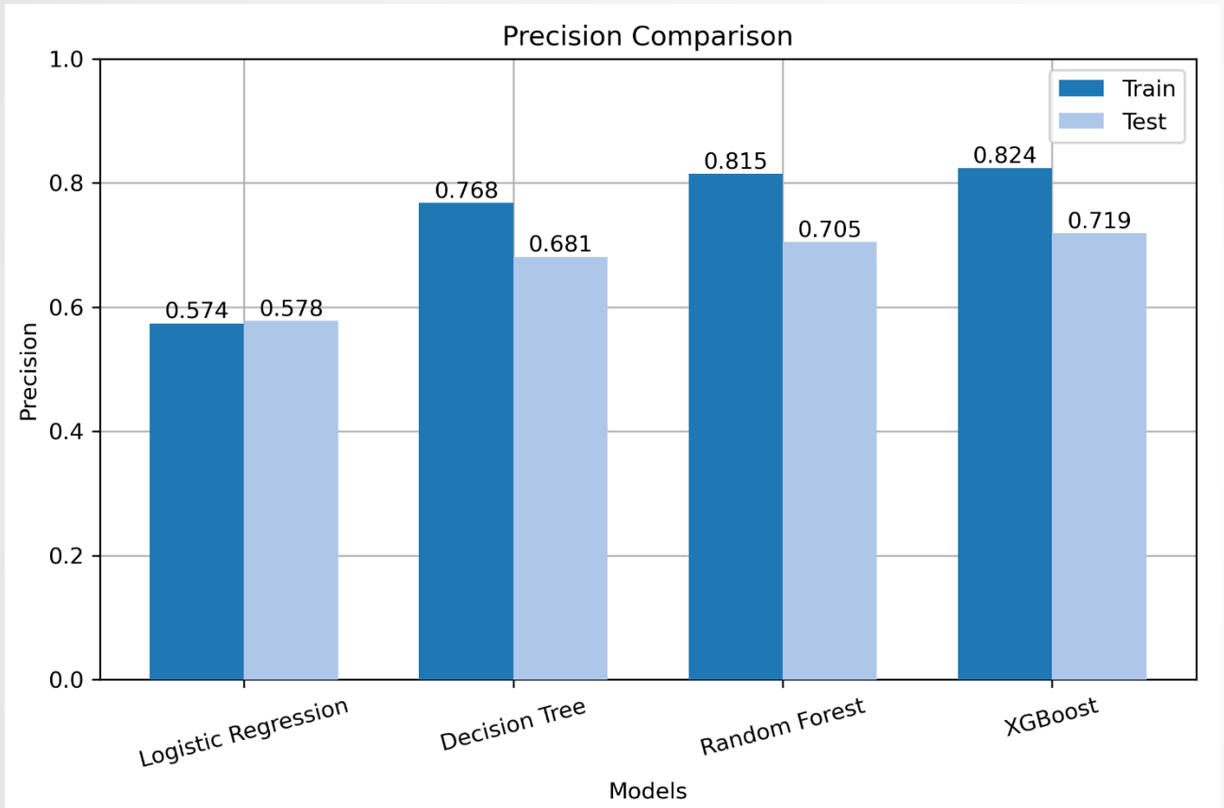
XGBoost

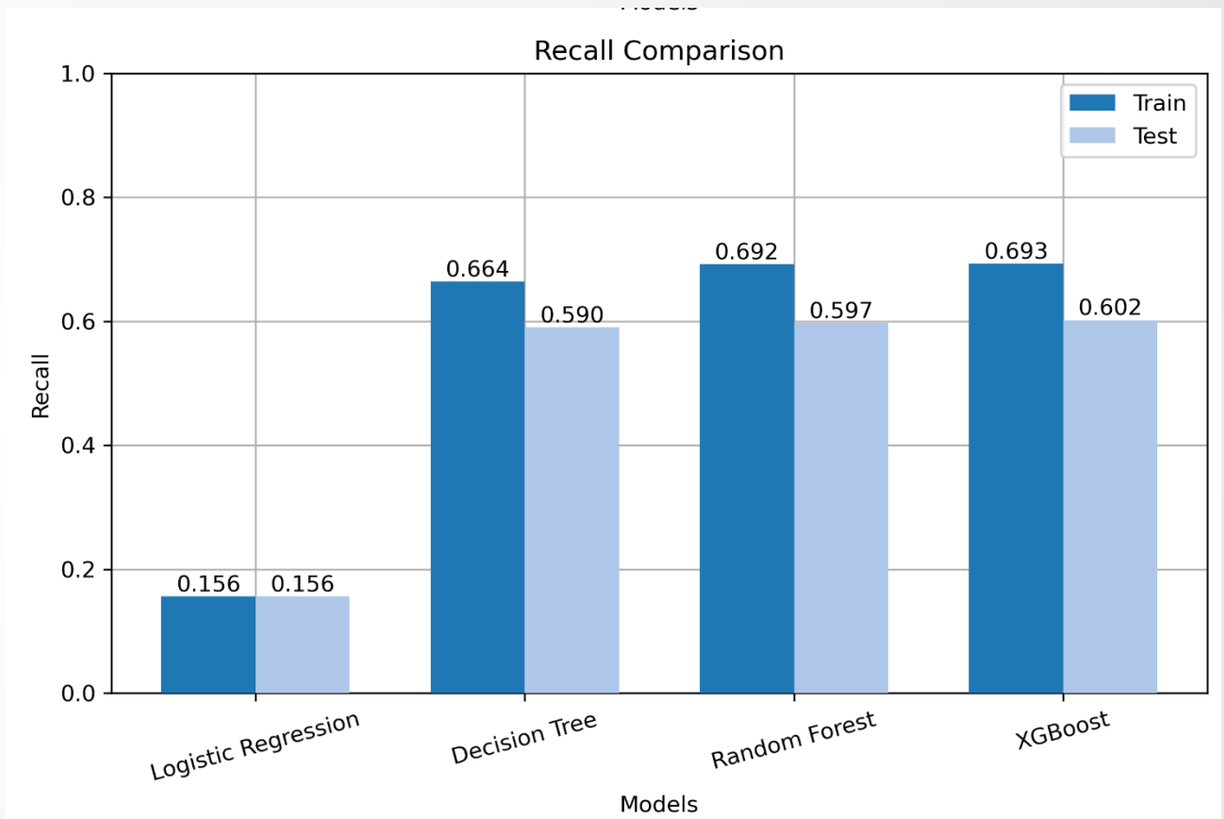


Result - Accuracy

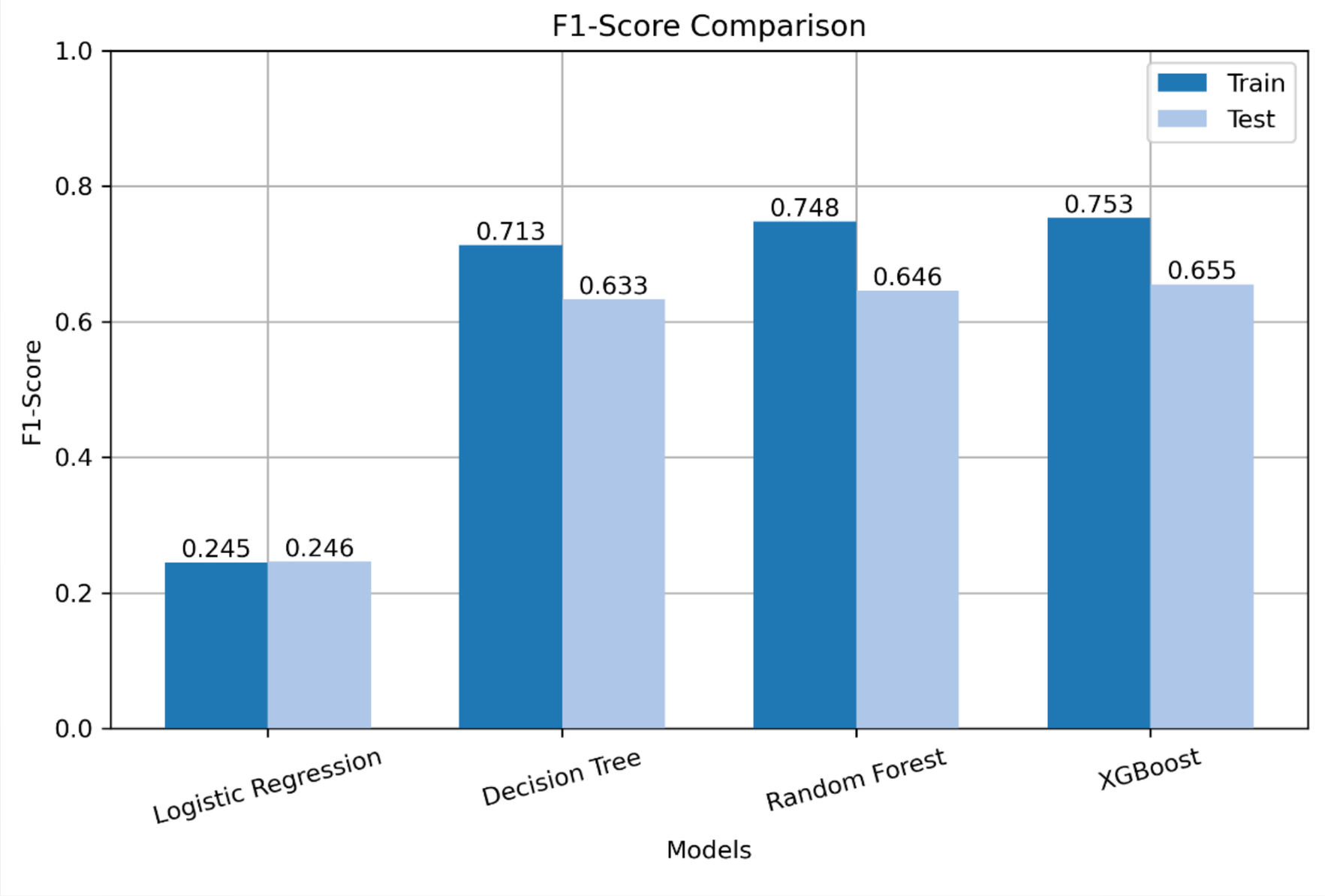


Result – Precision , Recall

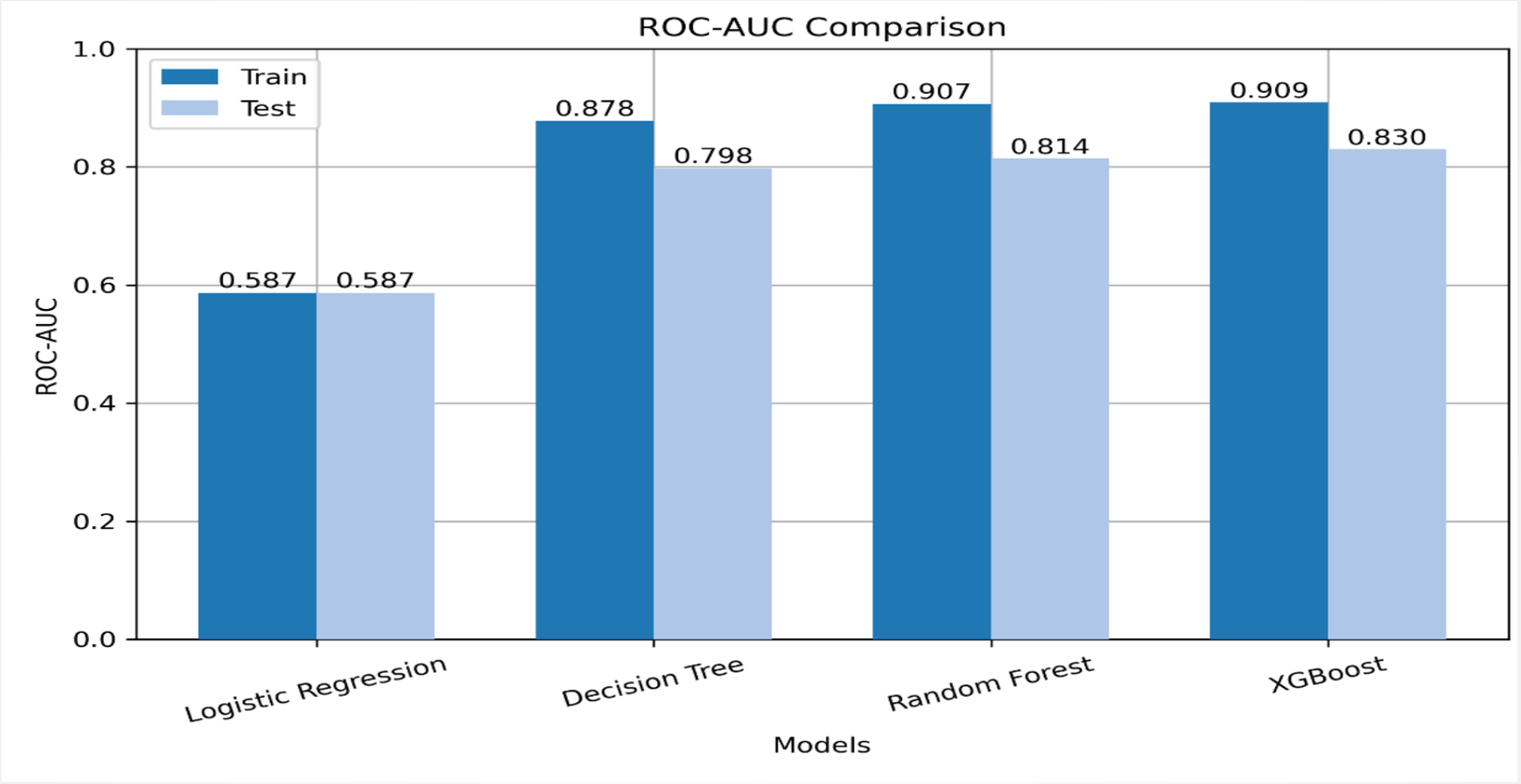




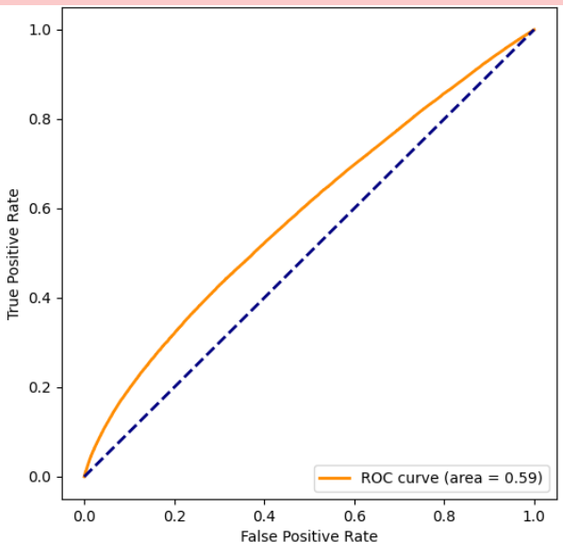
Result – F1 score



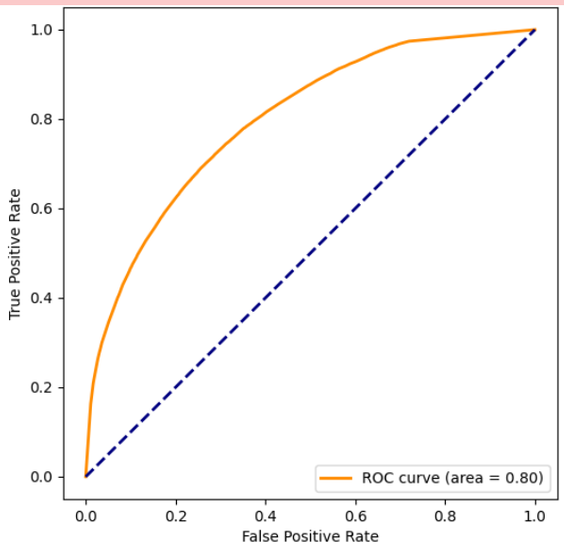
Result – ROC, AUC



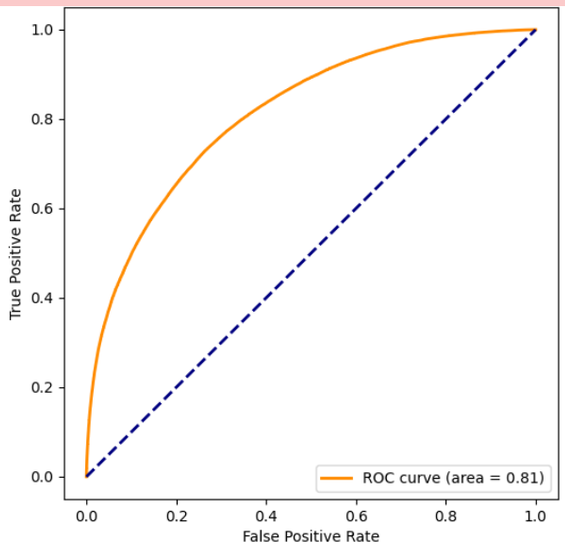
Logistic regression



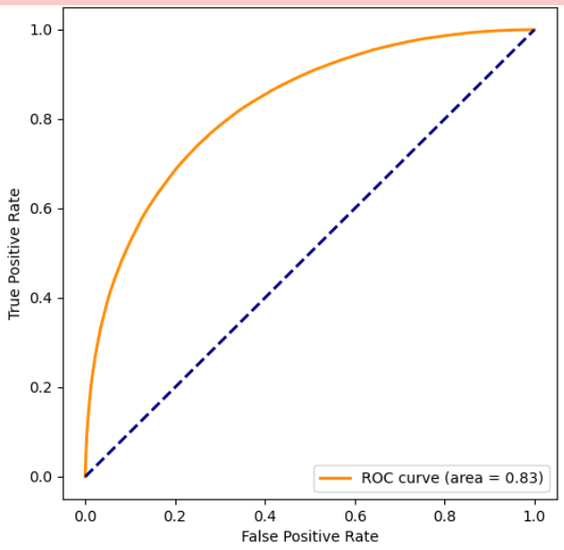
Decision tree



Random forest



XGBoost



Risk Radar :
Forecasting Credit Card Default

Ting Siou Chen