Credit Card Default Prediction Report

Akash Jaiswal
21411003
Geophysical Technology

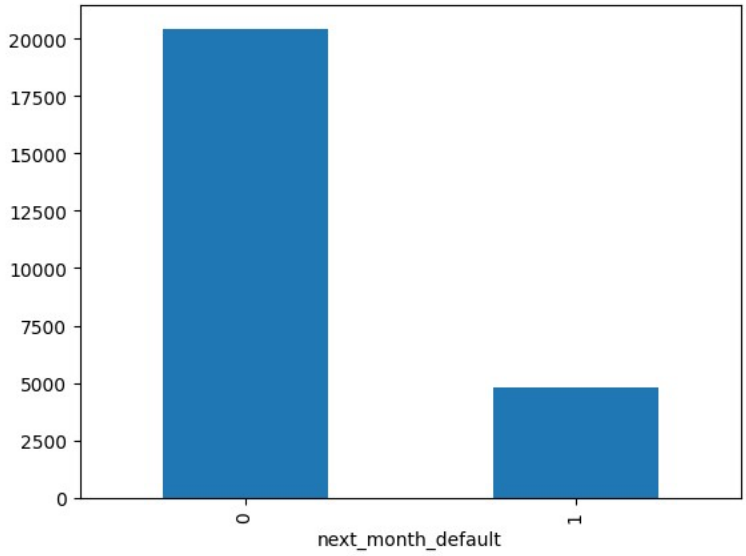
1. Data Preparation

- **Data Import and Initial Checks**
- Imported all necessary libraries and uploaded the dataset.
- Checked dataset size, structure, and presence of missing values.
- Only the *age* column had missing values, which were filled using the mean strategy.
- Verified for duplicate records.
- Analyzed the correlation between features and the target (*next_month_default*). The highest correlation was with *pay_0*.
- The correlation between the target and other features was negative, which is intuitive: credit limits are increased for those who pay bills on time.

2.Exploratory Data Analysis :

Univariate:

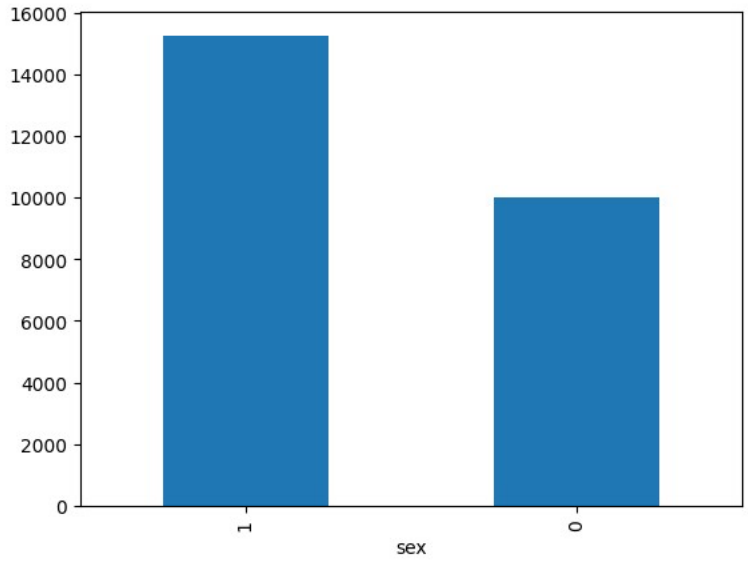
- Total number of Defaulters vs Non-Defaulters:



next_month_default	count
0	20440
1	4807

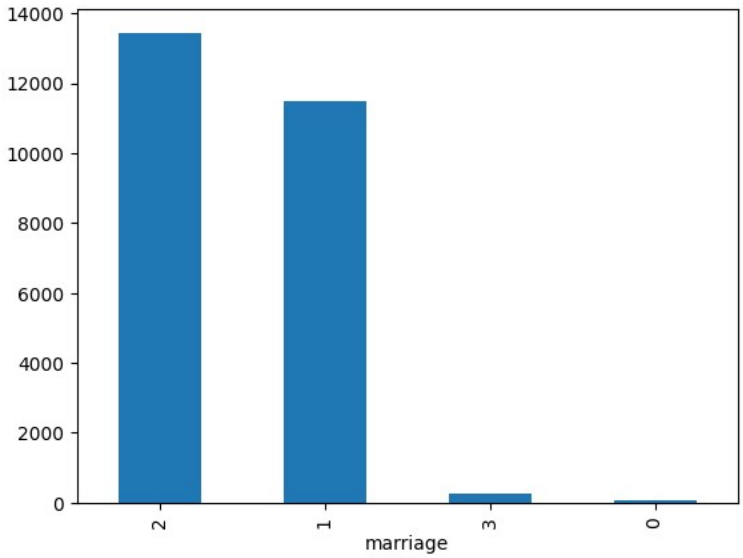
dtype: int64

- Based on sex males are more who are having credit card than females.



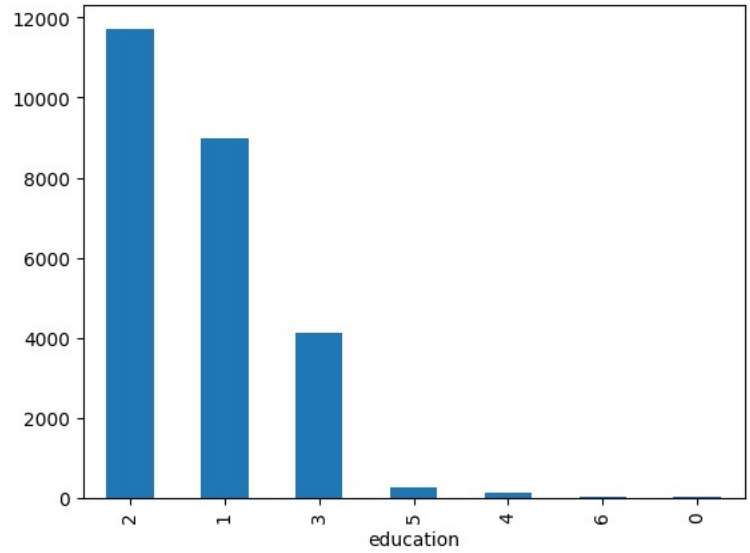
sex	
1	15252
0	9995

- Based on Marriage:
- People who are single are having more credit card.
- Here 2 denotes single and 1 denotes married remaining are others



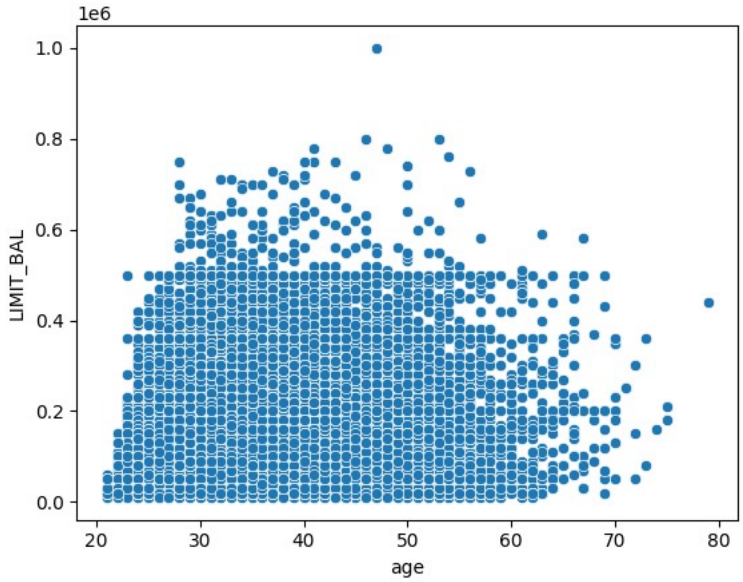
marriage	count
2	13441
1	11480
3	273
0	53

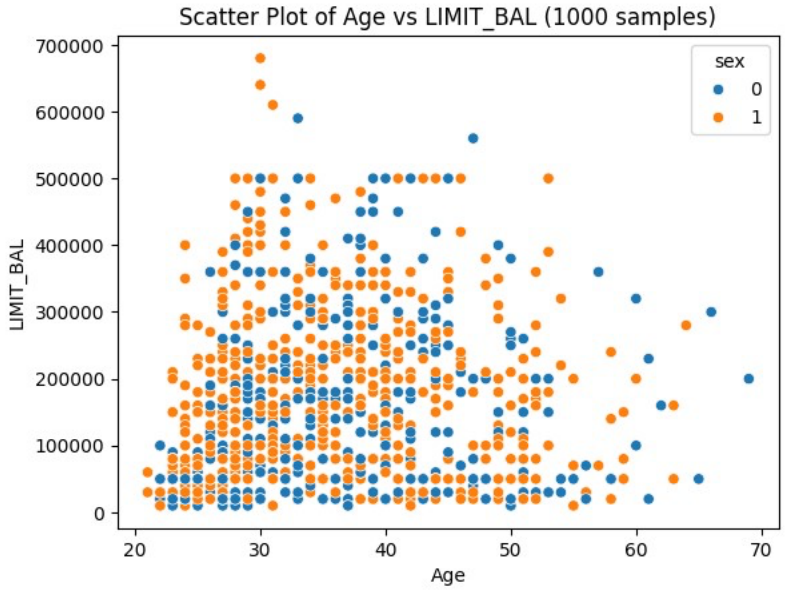
- Base on Education :
- Maximum customers are of University
- Here 2 denotes university



Bivariate Analysis:

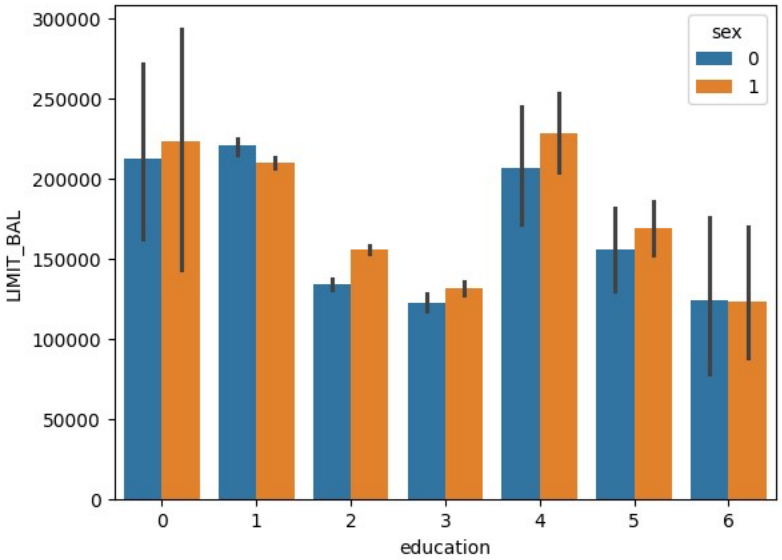
- It is analysis between age and limit of credit card which is sort of linear because as age increases source of earnings increases hence their limit is also increase as they become valuable customers for bank .





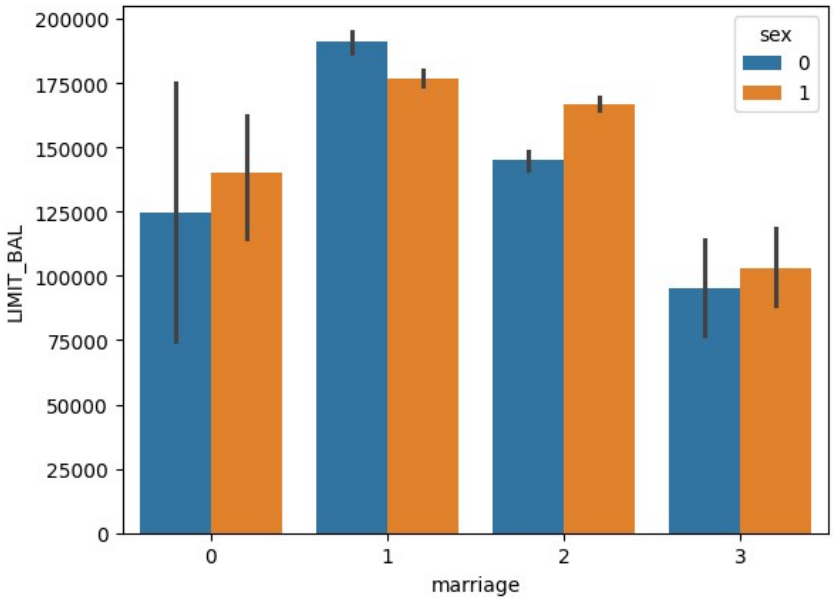
- In this scatterplot orange ones are males and blue ones are females, outliers customers are males .

- Sex and education Vs Limit_balance:



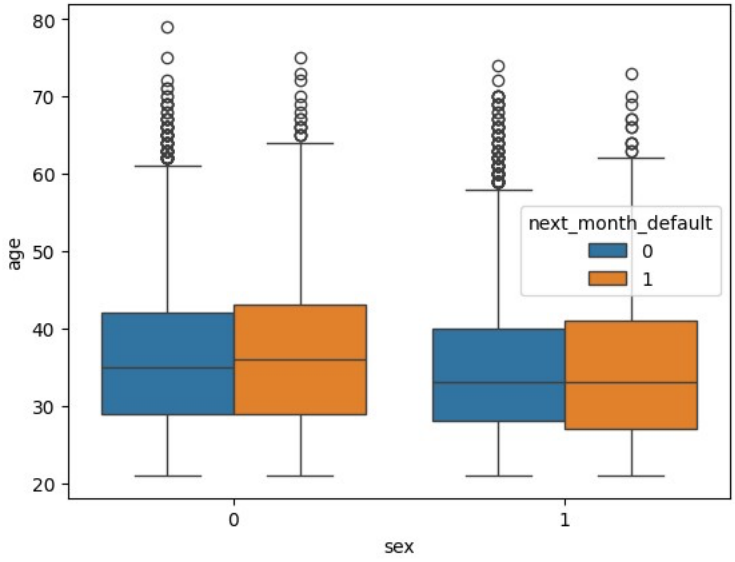
- This plot tells how each category of educated customer have their limit balance including sex also as each educated customer maybe male or female.

- Marriage ans sex vs Limit_balance:



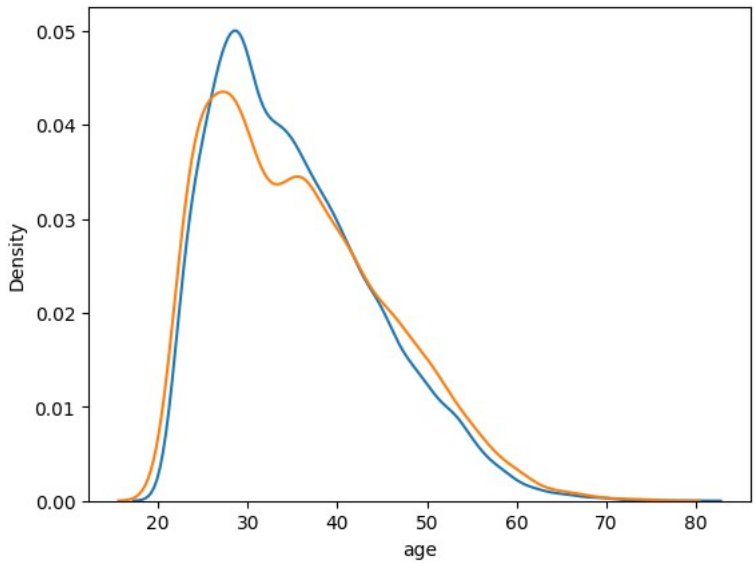
- This plot tells how each category of married customer have their limit balance including sex also as each married customer maybe male or female

Box plot:



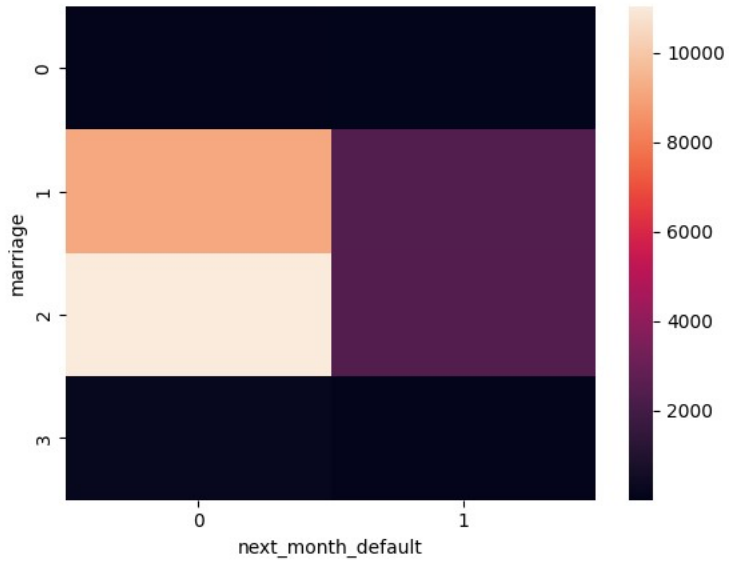
- This above plot tells which range age customer will default , as customer may be male or female , so from above plot it is concluded that females range of age is more for defaulting than males range.
- Small circles are outliers which can be removed while training our model.

Distplot:

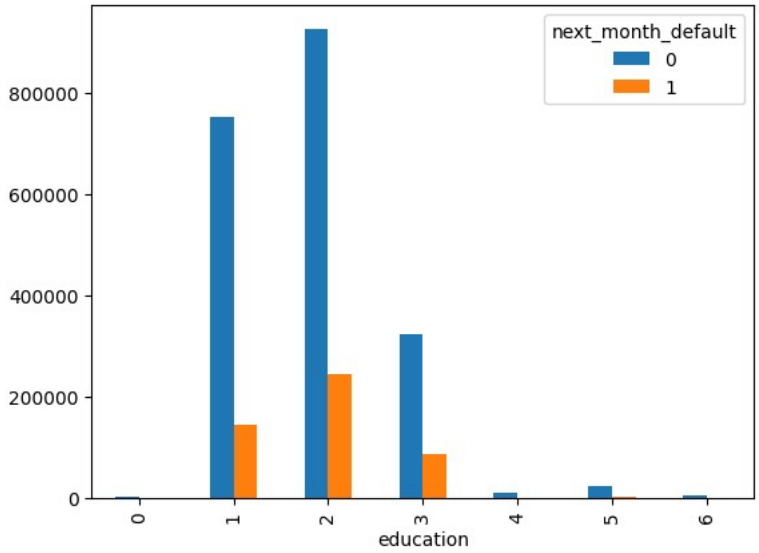


- This curve if pdf(probability density function) of defaulters vs non-defaulter based on age.
- Orange: non-defaulter
- Blue: defaulter
- If we analyse this curve then, pdf of defaulters of age range 22 to 38 lying above pdf of non defaulters of age range 22 to 38 ,it means this range people has higher probability of defaulting.
- Age range 40 and above pdf of non-defaulter is more than pdf of defaulter ,means older people having less chance of defaulting as their in come may be more.

Heatmap Analysis:



- Heatmap telling number of customers defaulted based on marriage.



- This plot tells how each category of educated customer defaulted including sex also as each educated customer maybe male or female.

Preperation of data for model training :

- Checked how a column feature is correlated to our output column.
- Droped Customer_ID from dataframe as it does not contribute to output column.
- Putting missing values through SimpleImputer class using strategy as mean.
- Splitiing Data for training the model is done.
- Checked for class-Imbalance,it was found that there was class imbalance so SMOTE technique was used to handle class imbalance
- Applied MinMax scaling on data(numerical cols)to bring data in same range for only logistic regression.

Threshold for This Dataset: ○ Given the imbalance (assuming defaults are rare), a threshold **lower than 0.5** (e.g., 0.3–0.4) may be optimal to capture more true defaults while accepting some false positives.

- ○ **Example**: If the bank prioritizes reducing defaults, a threshold of 0.3 might achieve higher recall (e.g., 80%) while keeping precision reasonable (e.g., 50%).

- **Metrics to Monitor**: Focus on **recall** (minimize missed defaults) while keeping **precision** at an acceptable level to avoid excessive false alarms.

All the score in all models are based on threshold of 0.4

Model -1(logistic regression)

- Accuracy of the Logistic Regression model: 0.7013861386138613
- Precision: 0.3327205882352941
- Recall: 0.5644490644490644
- F1 Score: 0.41865844255975326

From here on wards no scaling in other models as they don't have impact on scaling

Model-2(Decision-Tree):

- Accuracy of the Decision Tree model: 0.744950495049505
- Precision of the Decision Tree model: 0.3742283950617284
- Recall of the Decision Tree model: 0.5041580041580042
- F1 Score of the Decision Tree model: 0.4295837023914969

Model-3(Random-Forest) :

- Accuracy of the Random Forest model: 0.8057425742574258
- Precision of the Random Forest model: 0.4890929965556831
- Recall of the Random Forest model: 0.44282744282744285
- F1 Score of the Random Forest model: 0.46481178396072015

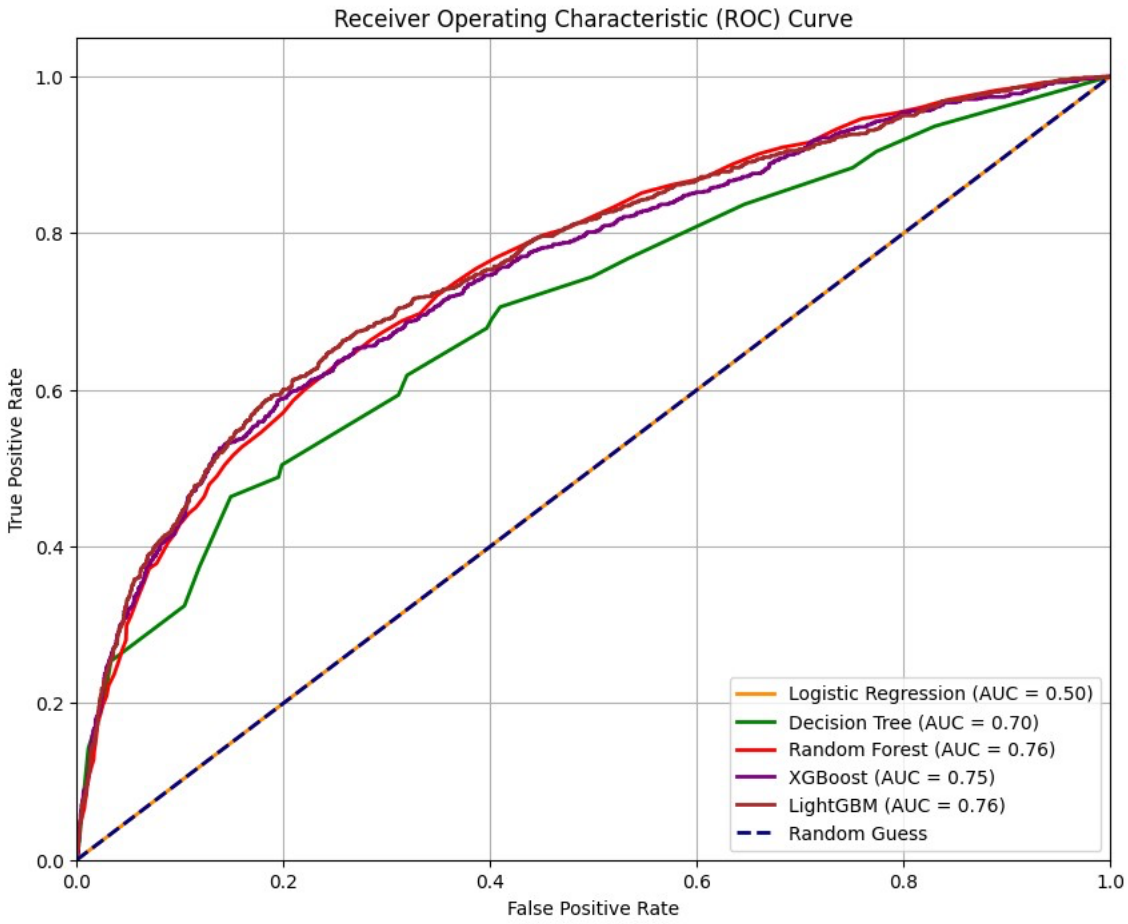
Model-4(XGBoost)

- Accuracy of the XGBoost model: 0.6128712871287129
- Precision of the XGBoost model: 0.29825274278748476
- Recall of the XGBoost model: 0.762993762993763
- F1 Score of the XGBoost model: 0.42886356996786446

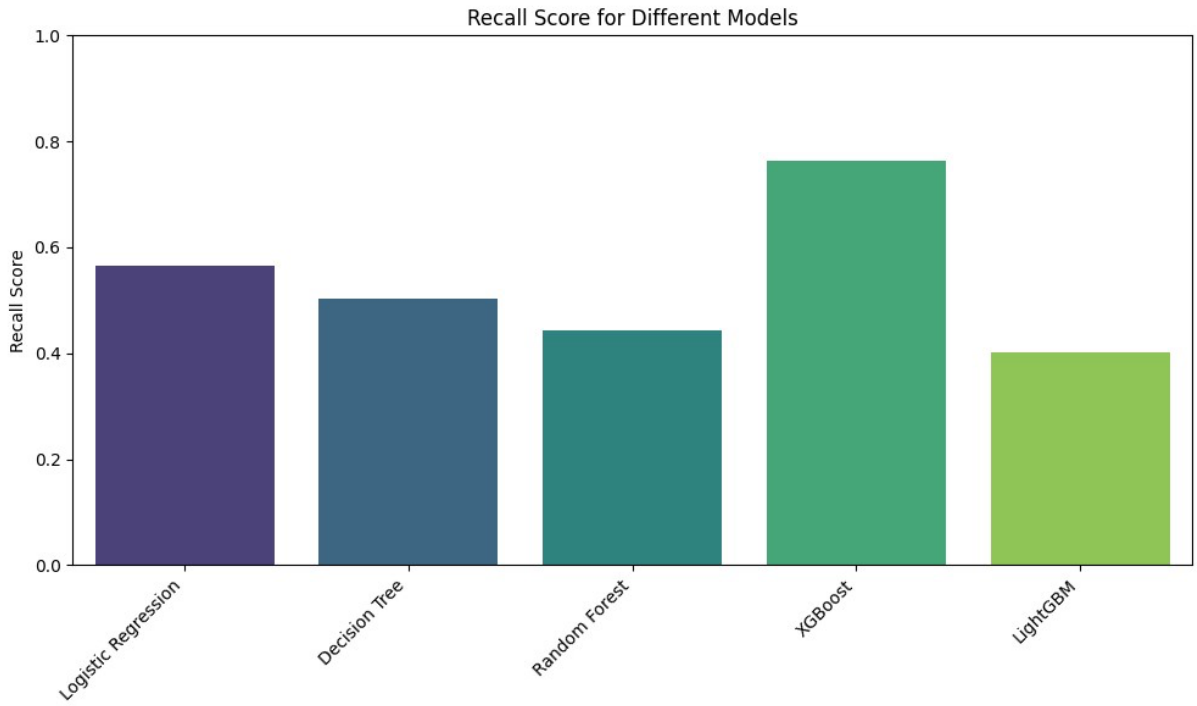
Model-5(Lightgbm)

- Accuracy of the LightGBM model: 0.8239603960396039
- Precision of the LightGBM model: 0.5520684736091298
- Recall of the LightGBM model: 0.4022869022869023
- F1 Score of the LightGBM model: 0.465423932651834

Base on all above 5 models AUC-ROC Curve:



Base on all above 5 models Bar-Plot of Recall :

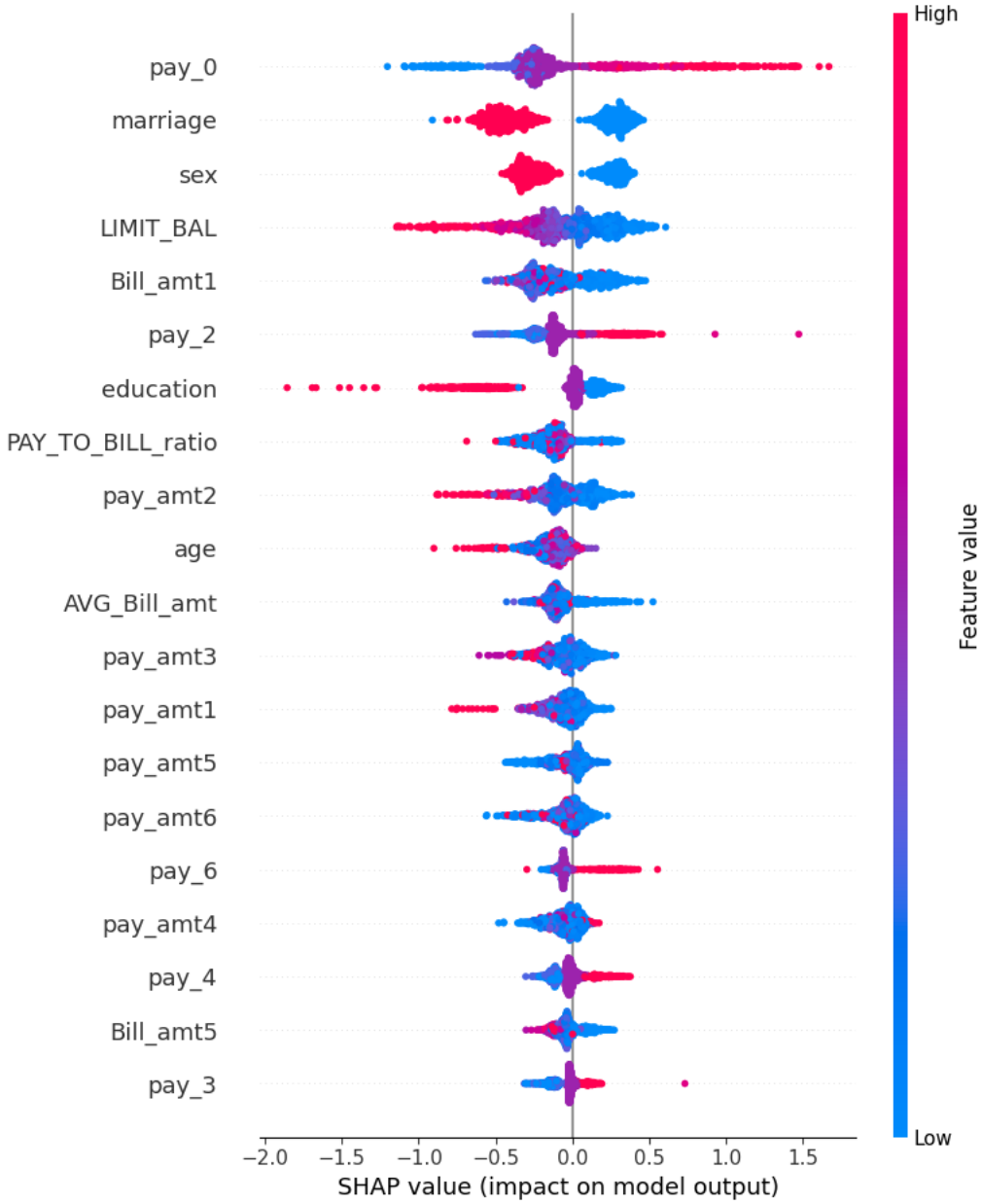


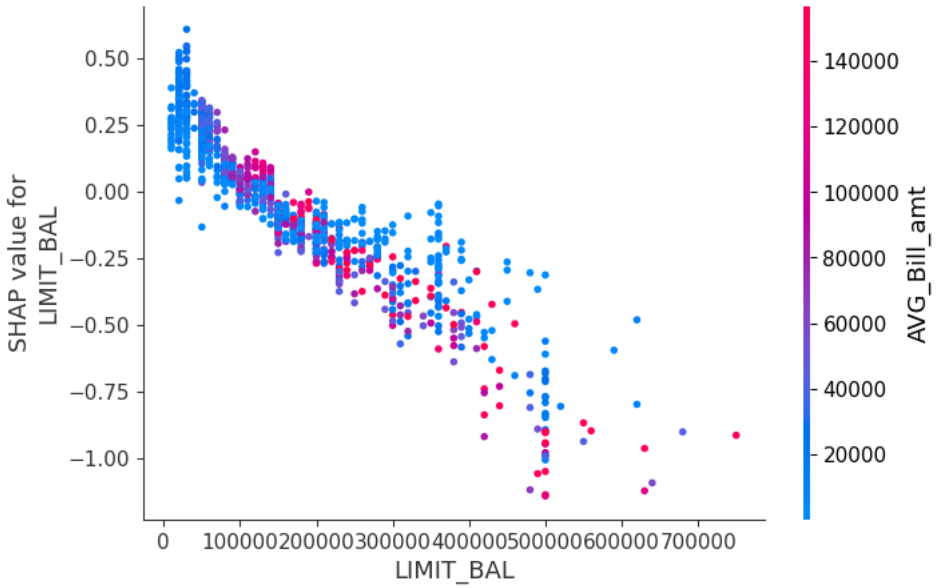
- Recall Scores:
- Logistic Regression: 0.5644
- Decision Tree: 0.5042
- Random Forest: 0.4428
- XGBoost: 0.7630
- LightGBM: 0.4023

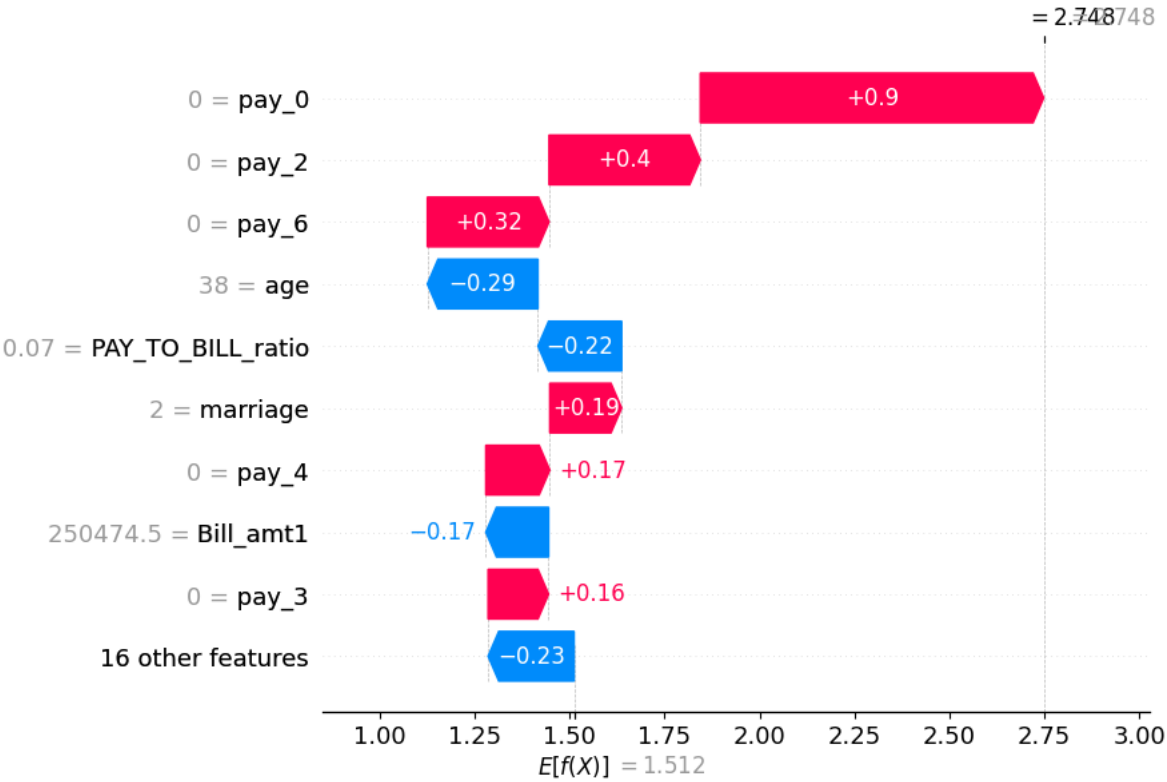
Model Selection

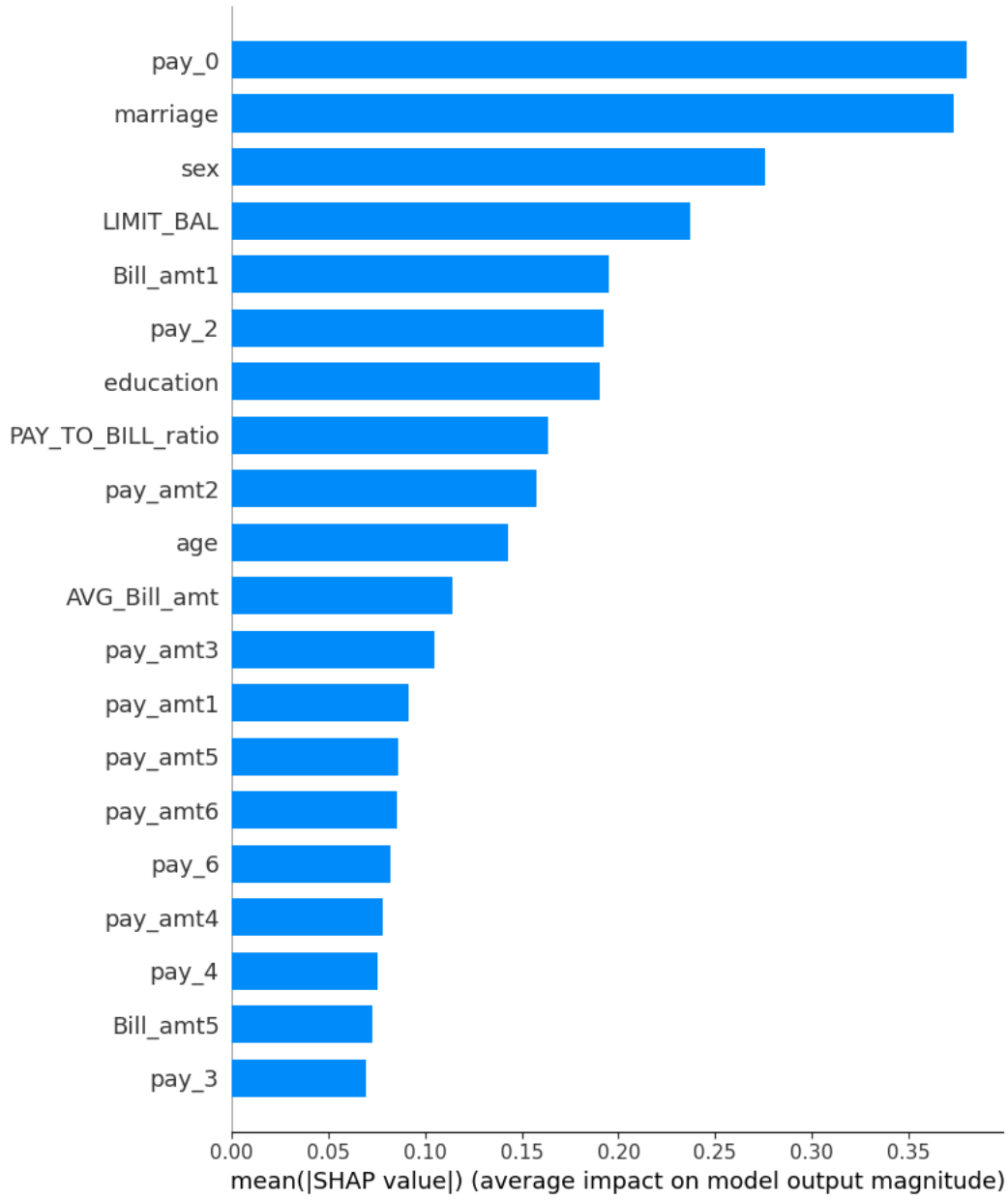
- **AUC-ROC Curve and Recall Comparison**
- XGBoost achieved the highest recall (0.7630), making it the best choice for minimizing Type-II errors.
- Recall is prioritized to reduce the risk of misclassifying defaulters as nondefaulters—a critical concern for banks.
- Hence based on this model I predicted prediction.csv file for validation set.

SHAP Analysis:(Plots) on Lightgbm model




SHAP value for
LIMIT_BAL
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
100000
200000
300000
400000
500000
600000
700000
LIMIT_BAL
140000
120000
100000
80000
60000
40000
20000
AVG_Bill_amt


= 2.748
0 = pay_0
0 = pay_2
0 = pay_6
38 = age
0.07 = PAY_TO_BILL_ratio
2 = marriage
0 = pay_4
250474.5 = Bill_amt1
0 = pay_3
16 other features
+0.9
+0.4
+0.32
−0.29
−0.22
+0.19
+0.17
−0.17
+0.16
−0.23
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
E[f(X)] = 1.512



Business Implications :

- **Risk Mitigation:**
- By prioritizing high recall, the bank minimizes the risk of failing to identify customers who are likely to default, thereby reducing potential financial losses.
- **Customer Segmentation:**
- Insights from EDA (such as age, gender, education, and marital status) can inform targeted marketing and risk assessment strategies.
- **Credit Policy Optimization:**
- Understanding which customer segments are more likely to default allows the bank to adjust credit limits or ofer tailored financial products.
- **Operational EJiciency:**

- Automated prediction models streamline the credit approval process and enable proactive risk management.

Summary of Key Learnings

- **Feature Importance:**
- The *pay_0* feature (payment status in the previous month) is the most correlated with default risk.
- **Demographic Insights:**
- Males and university graduates are the largest customer groups.
- Single individuals are more likely to hold credit cards than married individuals.
- **Age and Default Risk:**
- Customers aged 22–38 have a higher probability of defaulting, while those aged 40+ are less likely to default.
- **Model Selection:**
- XGBoost outperformed other models in recall, making it the best choice for minimizing Type-II errors.
- **Data Preparation:**
- Addressing class imbalance with SMOTE and proper feature engineering are crucial for model performance.

Conclusion

- **Best Performing Model:** XGBoost (based on recall and AUC-ROC curve).
- **Key Insight:** Recall is the most important metric for this use case, as it minimizes the risk of failing to identify defaulters.
- **Recommendation:** Use XGBoost for predicting credit card defaults on validation data