Capstone Project-3

CREDIT CARD DEFAULT PREDICTION

Submitted by- Shreesha K

A brief introduction

- A credit card is a financial instrument issued by banks with a pre-set credit limit, which helps to make cashless transactions. The banks determine the credit limit by credit score, credit history, and several other factors.
- Banks issue credit cards to their customers without checking their backgrounds to improve their market share. Even some customers use credit cards beyond their repayment capacity, resulting in debt accumulation.
- Credit card default usually happens if a customer doesn't make minimum payments for six months in a row.
- It is important to predict whether a customer is risky or non-risky by analyzing his past transactions, current payment status, etc.

Contents

- Describing the problem.
- About the dataset
- Exploratory Data Analysis
- Feature Engineering
- Modelling different ML models
- Comparing the different algorithms
- Conclusion

Problem statement

The objective is to build a machine-learning model to predict whether a customer will default the payment.

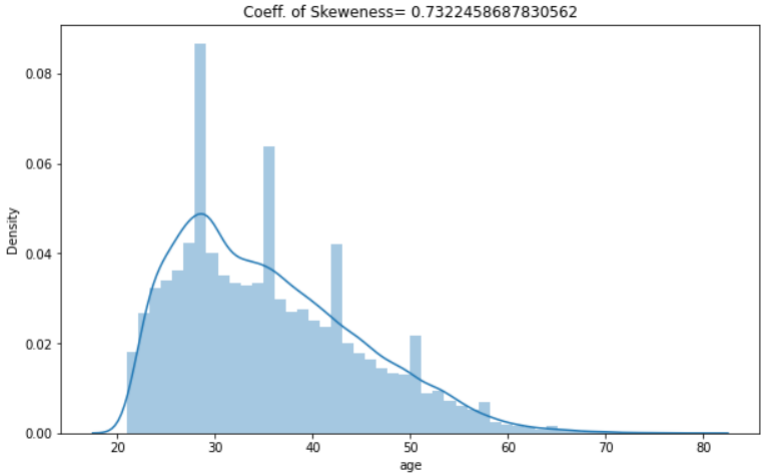
About the Dataset

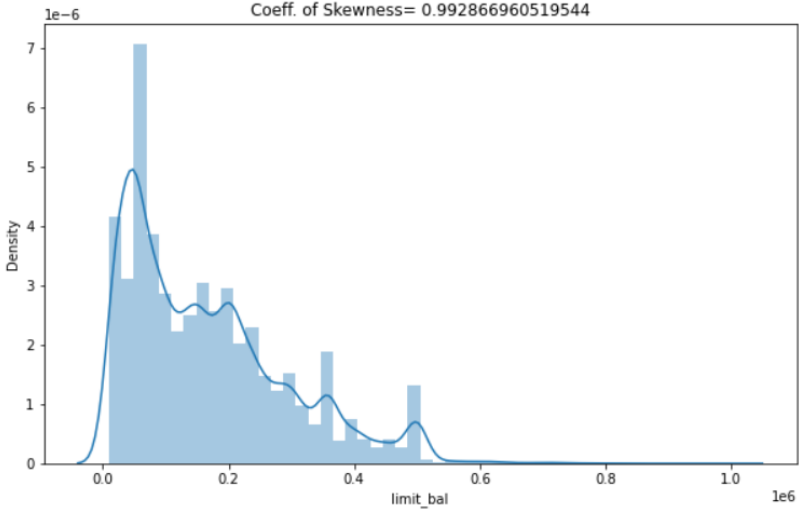
- This dataset contains information on the customer default in Taiwan.
- The dataset contains 30000 rows and 26 columns.
- There were no null values in the data indicating that the data is already cleaned.
- There were no duplicate entries as well.
- The columns in the dataset were: 'ID', 'LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6', 'defaulted'.

About the Dataset

- In the dataset, PAY_0, PAY_1, etc., represent the repayment status in various months.
- BILL_AMT1, BILL_AMT2, etc., represent the amount of bill statement in various months.
- PAY_AMT1, PAY_AMT2, etc., represent the amount repaid in various months.
- Gender number 1 denotes the males and 2 denotes the females.
- The education status- 1 indicates University, 2- Graduate, 3-Highschool, 4- Others.
- Marital status 1 represents married people, 2 represents single and 3 represents others.

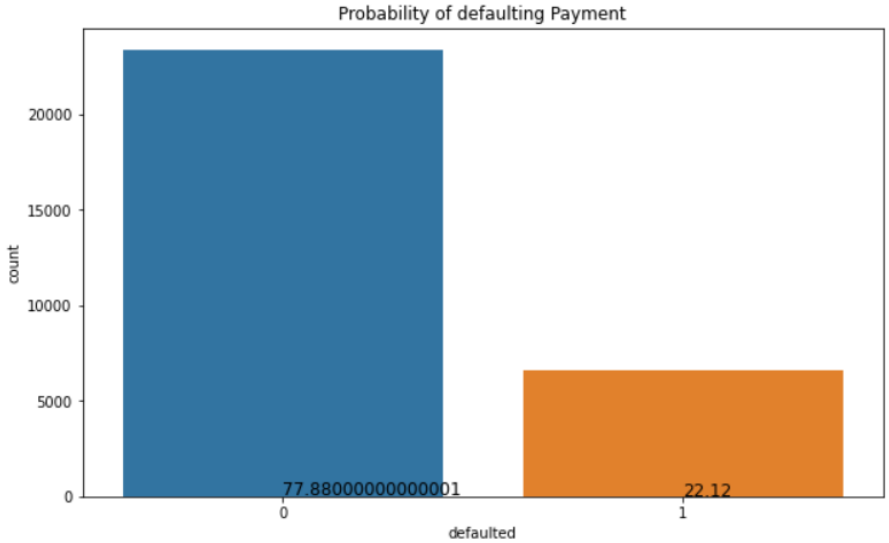
Distribution of age as well as credit limit





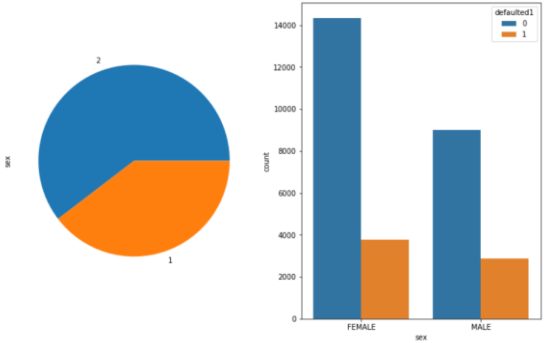
- We can see that most of our customers are in the age group of 25-40 and the number of senior citizens is too less.
- The credit limit allowed for most of the customers is between 10000 and 20000.

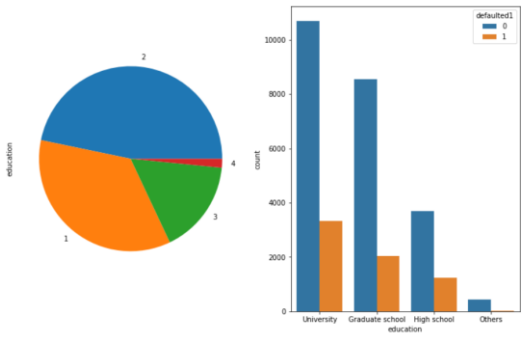
Percentage of Defaulters and Non- Defaulters

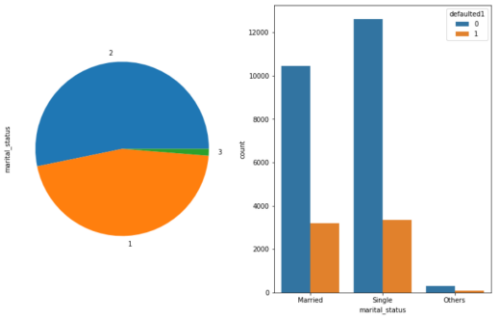


- We can observe that almost 78% of the customers are non defaulters and 22% customers are defaulters. Therefore, the dataset is imbalanced.

Number of customers category-wise

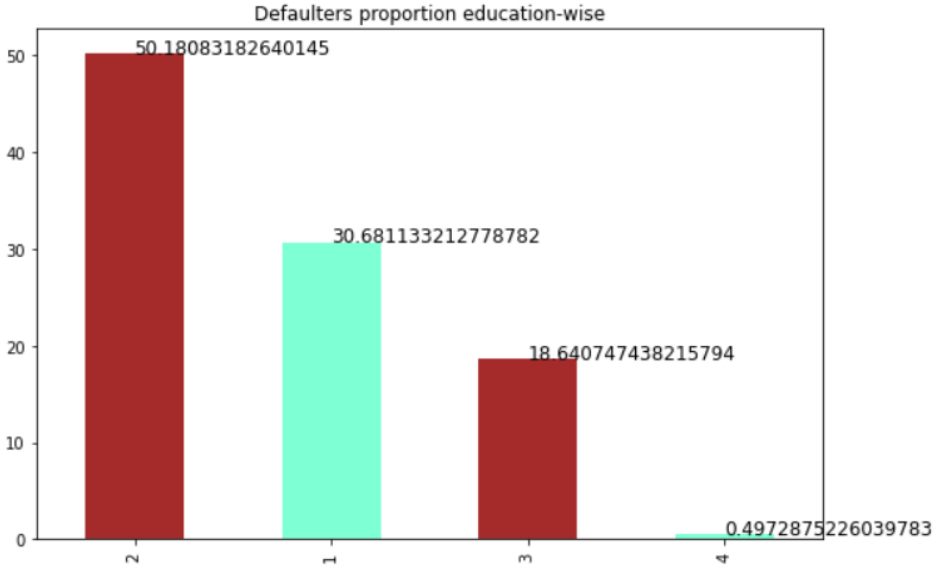






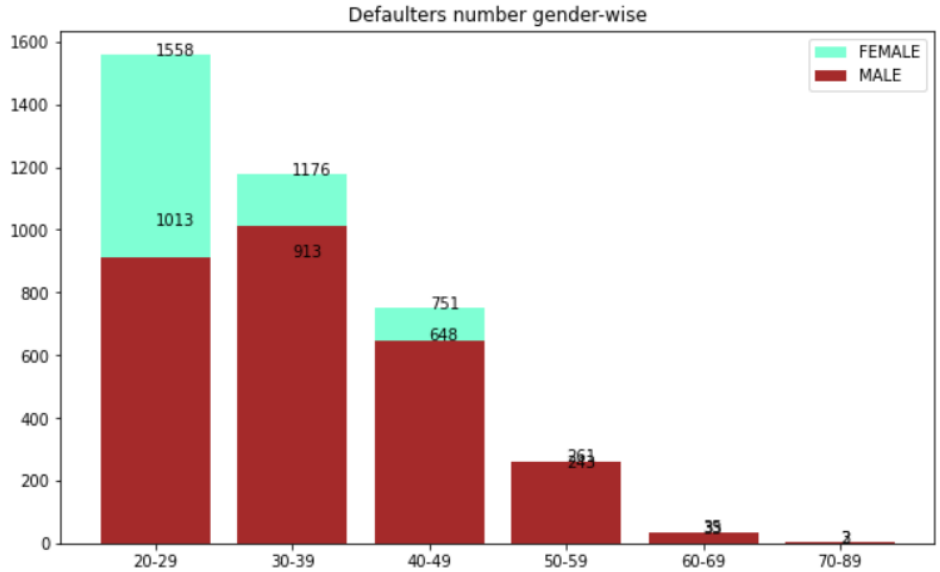
- The number of female customers is more. Since we have more female customers, the female customers who defaulted on the payment are also more.
- There are few customers whose education status is 'others'. People with higher education defaulted more.
- There is no significant relationship between marital status and defaulted payment. The proportion of defaulted unmarried people is slightly higher than defaulted married people.

Number of defaulted customers category-wise



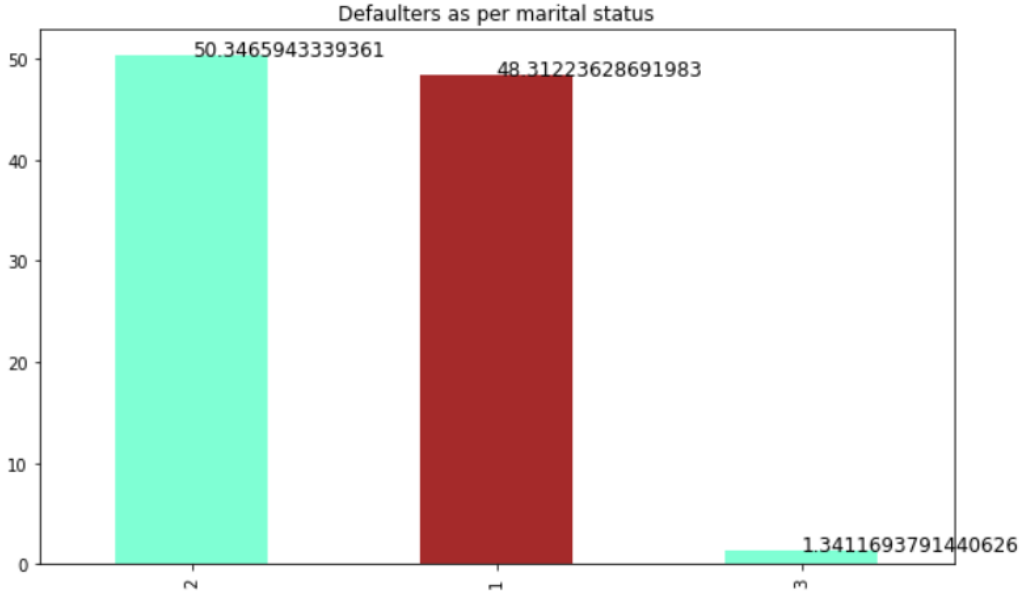
We can observe that almost 50 percent of the defaulted customers have university-level education. As observed earlier, customers with a low level of education do not tend to default on payment.

Number of defaulted customers category-wise



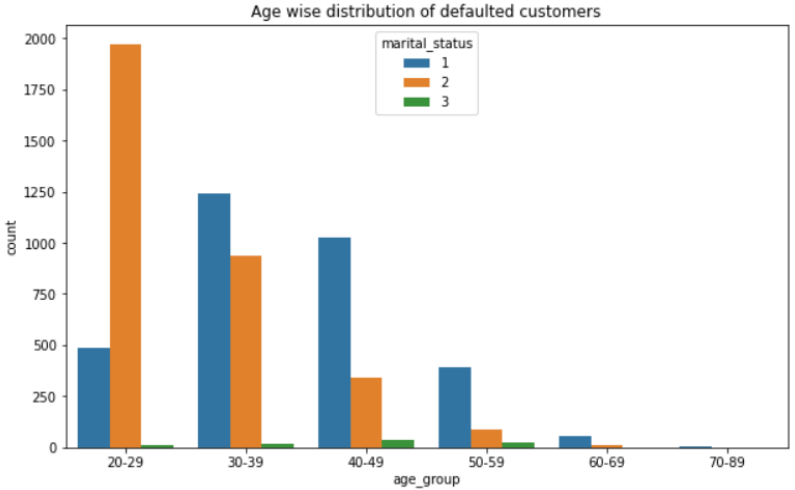
In the age group of 20-29, we have the highest number of female customers who have defaulted. The male customers who defaulted are more in the 30-39 age group.

Number of defaulted customers category-wise



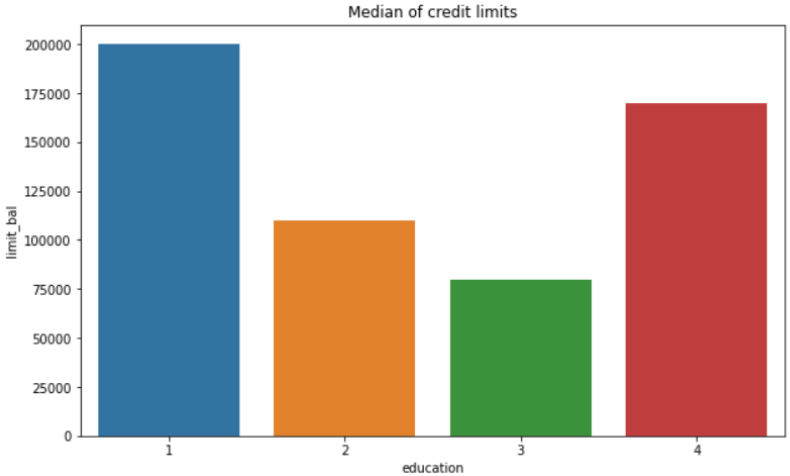
There is no significant relationship between marital state and default payments.

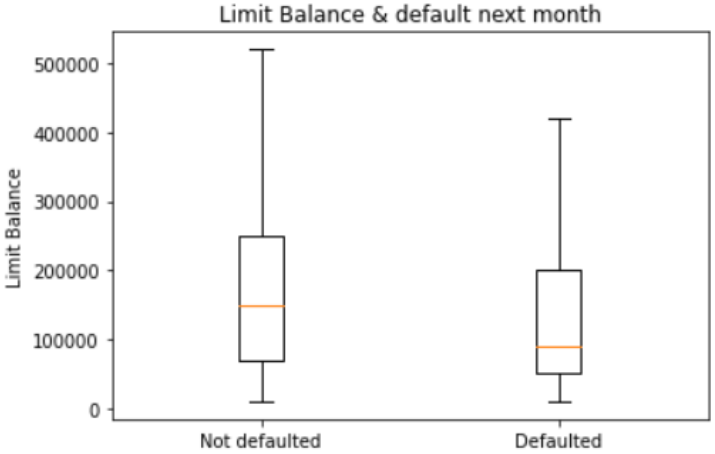
Number of defaulted customers category-wise



- It can be observed that as the age increases, the number of defaulted customers decreases.
- In each age group, we have a more or less similar distribution of married, single, and other people. So we can conclude that age is more significant in this case, not marital status. As the age increases, customer default decreases.

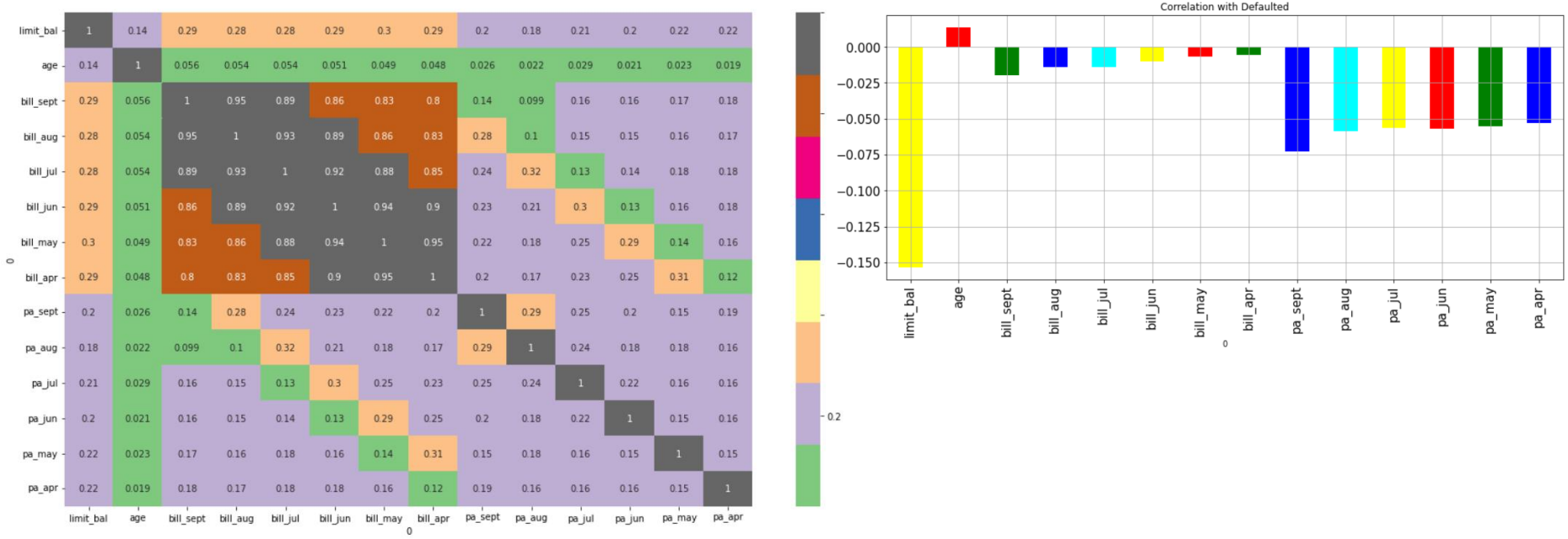
Credit limit for different customers





- From the first graph, we can see that people with higher education got higher credit limits.
- We can note that the credit limit for defaulted customers was lower than non- defaulted customers.

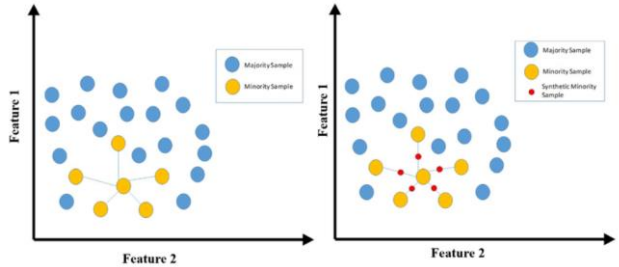
Correlation Heatmap



- **The limit balance is the most negatively correlated feature with the defaulted payment.**

Train- Test Split and Modelling

- As we have observed, the dataset is imbalanced , I.e. , only a small proportion of customers have defaulted. Some of the ML models will not perform better with this issue.
- To overcome the same, we have implemented SMOTE (Synthetic Minority Oversampling Technique). By this method, we generated a number of minority data points, which are synthesized from the existing datapoints.
- We have used Random Forest, Logistic Regression, K-Nearest Neighbors, and Decision Tree Algorithms. Each model was trained with and without implementing SMOTE so that we can compare the performance difference.



Performance Parameters

1. Accuracy

Ratio of number of correct predictions to the total number of predictions.

2. Precision and Recall

Precision is the ratio of number of correctly classified positive data points to the total number of positive predictions.

Recall is the ratio of number of correctly classified positive datapoints to the total number of actual positive datapoints. It is given by TP/TP+FN. In this problem, Recall is an important performance parameter. The banks can’t afford any False Negatives, i.e. we can not classify any defaulted customer to be non- defaulted.

Performance Parameters

3. Confusion Matrix

It indicates the extent to which the ML algorithm can classify the data points into specific classes.

4. Area under ROC Curve

ROC curve consists of two axes, namely, True positive rate and False Positive Rate. The higher the area under the curve more is the model performance.

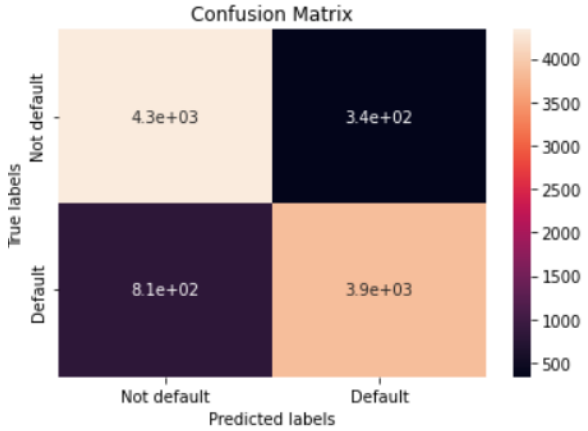
5. F1 Score

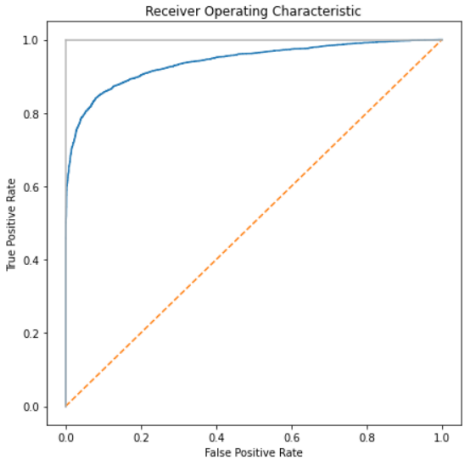
It is the harmonic mean of the precision and recall.

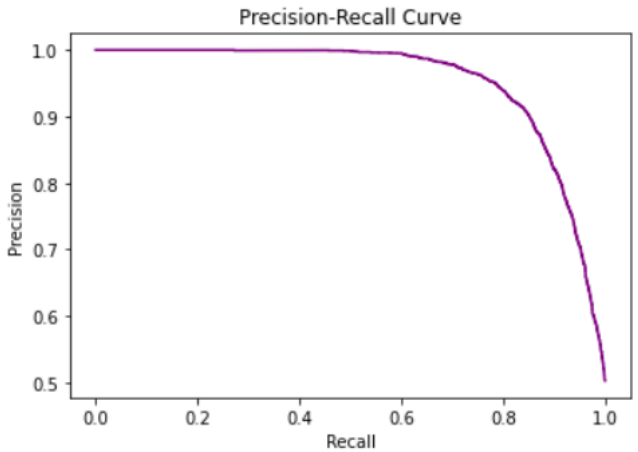
Machine Learning Models

1. Random Forest

Random Forest is a classifier algorithm that trains several decision trees on various subsets of the parent dataset and produces the final output by taking the average of the outputs of all trees.



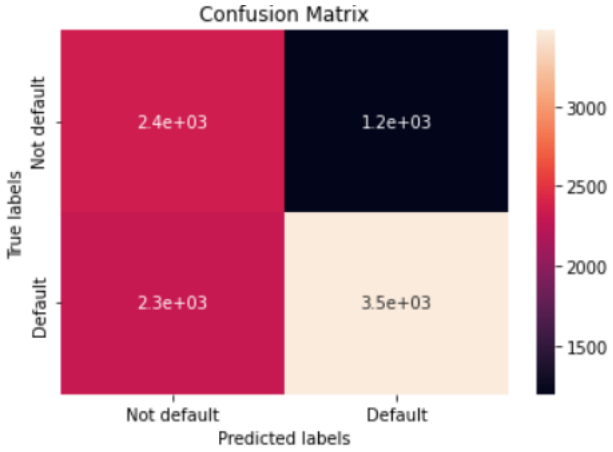


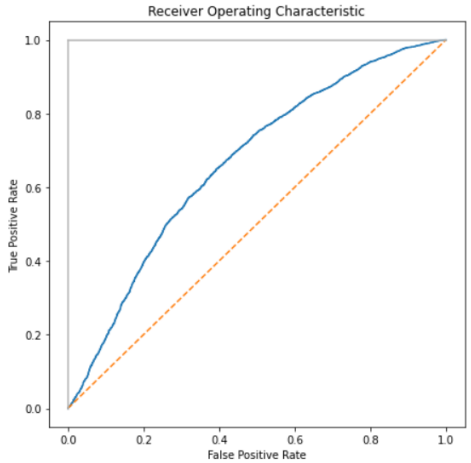


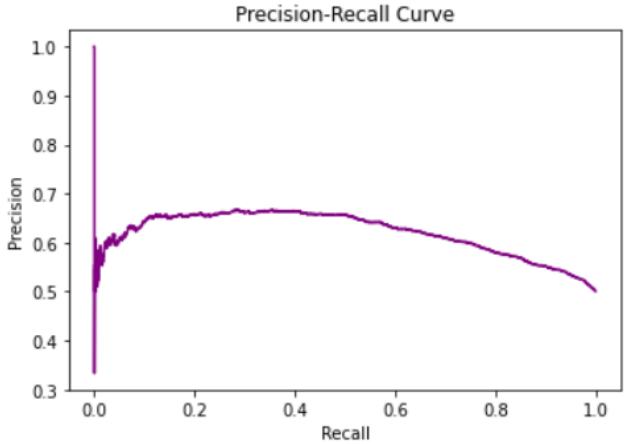
Machine Learning Models

2. Logistic Regression

Logistic Regression is used to build ML models where the dependent variable is discreet and dichotomous. It fits an S-shaped sigmoid curve into the dataset and makes predictions depending on that.



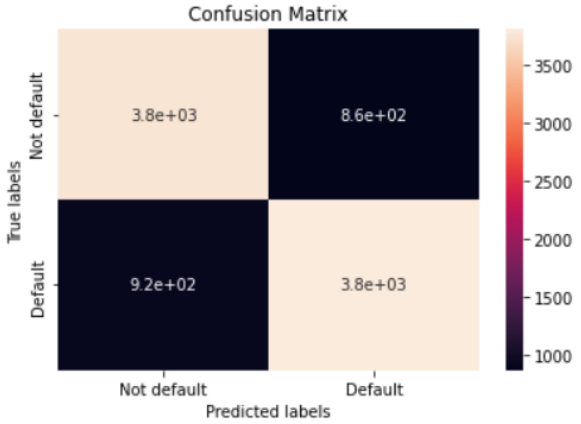


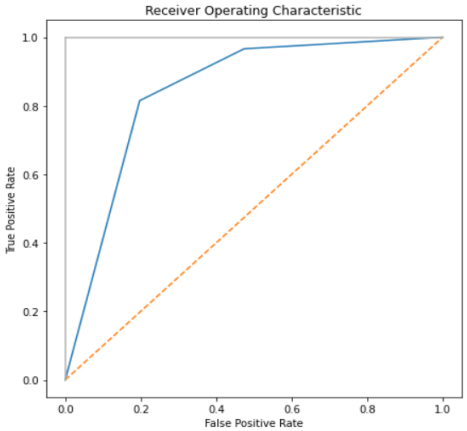


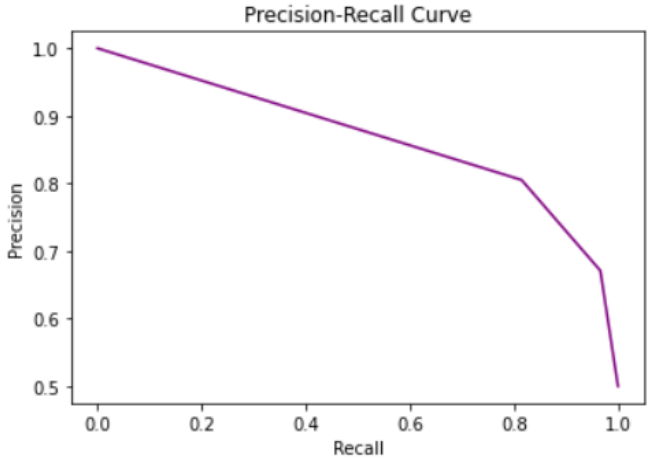
Machine Learning Models

3. K- Nearest Neighbors (KNN)

KNN uses the similarity of the features to predict the value of new data points. The new data point is assigned a value based on how closely it matches its neighboring points.



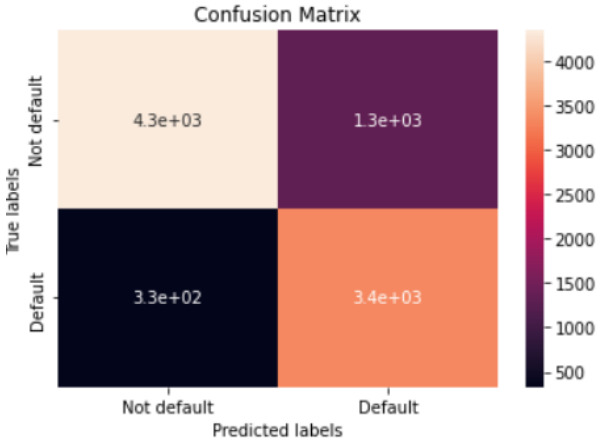


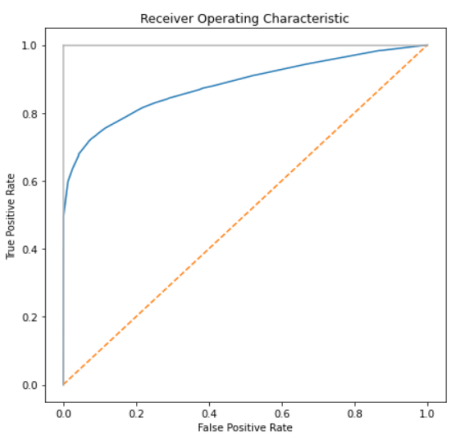


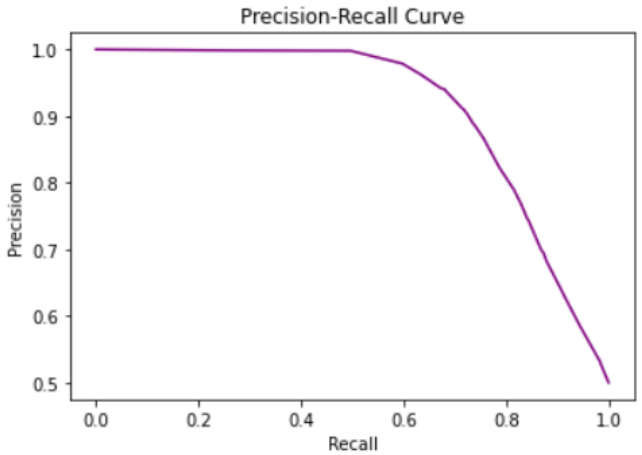
Machine Learning Models

4. Decision Tree

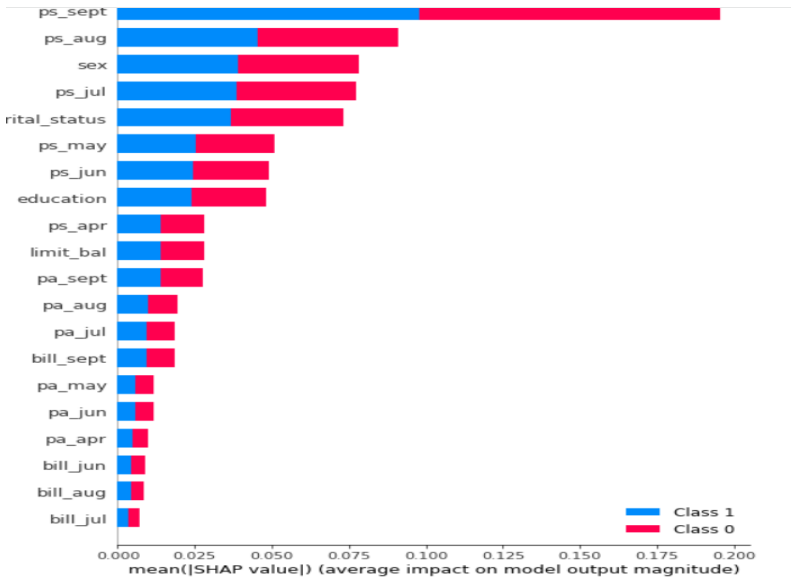
A Decision Tree is a supervised ML model where data is continuously split based on certain parameters. It contains decision nodes and leaves. The data is split into decision nodes and leaf nodes contain the final output.

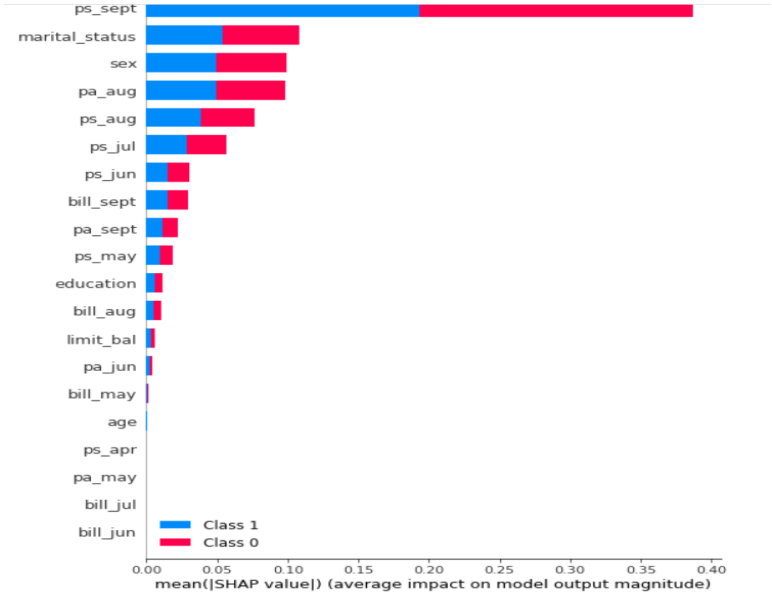






Feature importance





These plots represent the most important features according to Random Forest and Decision Tree Algorithm.

Summary

ALGORITHM	RECALL SCORE	F1 SCORE	ROC SCORE
Random Forest	0.66(Without SMOTE) 0.83 (With SMOTE)	0.46(Without SMOTE) 0.87 (With SMOTE)	0.75(Without SMOTE) 0.88 (With SMOTE)
Logistic Regression	0.25(Without SMOTE) 0.66 (With SMOTE)	0.01(Without SMOTE) 0.66 (With SMOTE)	0.51(Without SMOTE) 0.63(With SMOTE)
KNN	0.46(Without SMOTE) 0.81 (With SMOTE)	0.18(Without SMOTE) 0.81 (With SMOTE)	0.62(Without SMOTE) 0.81 (With SMOTE)
Decision Tree	0.66(Without SMOTE) 0.91(With SMOTE)	0.46(Without SMOTE) 0.80(With SMOTE)	0.75(Without SMOTE) 0.83(With SMOTE)

Conclusion

- Accuracy is not a good performance parameter in this problem where the data is imbalanced. We can not classify a defaulter as a non-defaulter by mistake, i.e., we need to keep an eye on the number of false negatives. Hence, recall is a better performance parameter.
- We have used Random Forest, Logistic Regression, KNN, and Decision Tree algorithms. Amongst these models, the performance of Logistic Regression was quite poor. It had a heavy imbalance in the precision and recall scores, which was balanced after implementing SMOTE.
- Decision Tree and Random Forest algorithms performed better with and without SMOTE. Random Forest had a recall score of 93% (for class 0), and 83% (for class 1), and Decision Tree had a recall score of 77% (for class 0), and 91%(for class 1).

Conclusion

- The performance of each model got improved with SMOTE, especially KNN and Logistic Regression.
- Repayment Status in September, Repayment Status in August and Gender were the most important features.

Thank you