# Capstone Project Classification

## Predicting whether a customer will default on his/her credit card

# TEAM MEMBERS

Abhishek Mishra

Kurva Mallesh

Arunesh Mishra

# WHAT IS customers default payments

This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

# Objective of The Project

The main objective of this project is to find the business insider from the data. Also create the model which can predict customers default payments .

In this project we need to predict weather the customer is default or not because the company need to create a model to check for new customer default predication behalf of old customers data.

# Business Problems

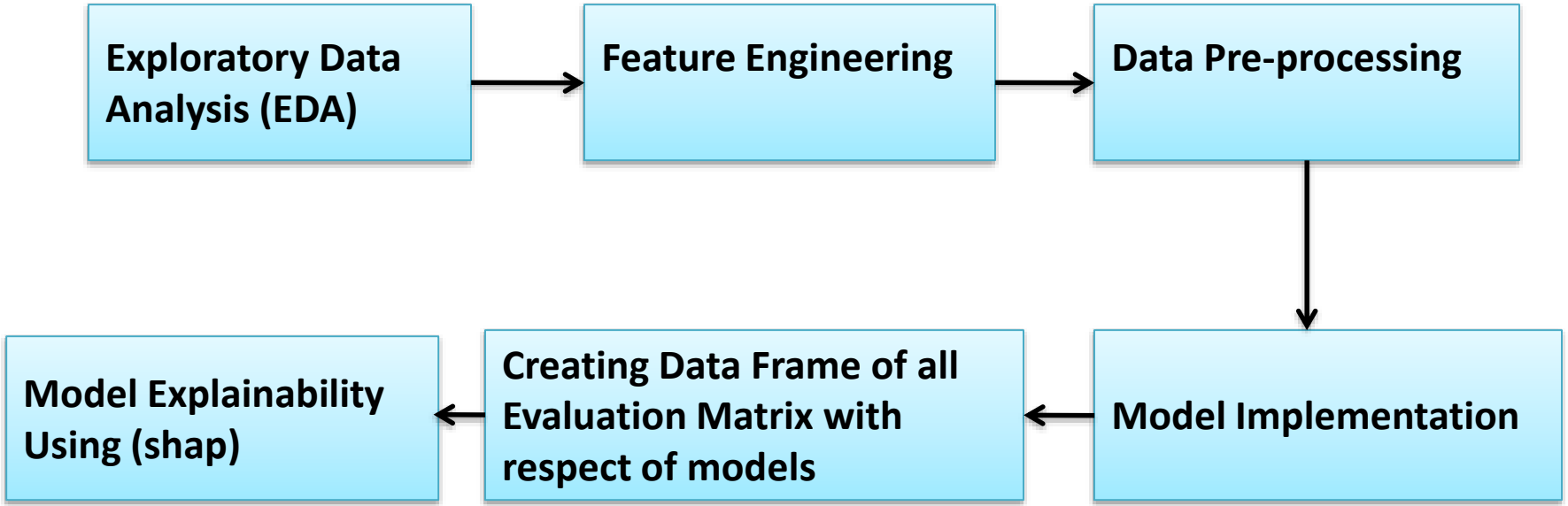
We can see that the data type of all columns are object data types So we need to convert it?

Fist row is the actual column names so we need to replace it?

What can we learn about "Education" Column ?

What can we learn about MARRIGE column ?

# Road map of the project



# Exploratory Data Analysis (EDA)

- ✓ Notice code 0 and -2 are in the PAY columns but are not included in the data description.
- ✓ We found there are 95919 values are listed as 0. Using some google search we found that 0 meaning the payment wasn't due, which makes sense that most customers were using the revolving credit.
- ✓ Also we found 24415 values as -2 which means No consumption.
- ✓ There are no duplicate IDs or rows.
- ✓ 30% male have default payment while 26% female have defaul payment, the difference is not significant.
- ✓ 'EDUCATION' column: notice 5 and 6 are both recorded as 'unknown' and there is 0 which isn't explained in the dataset description.

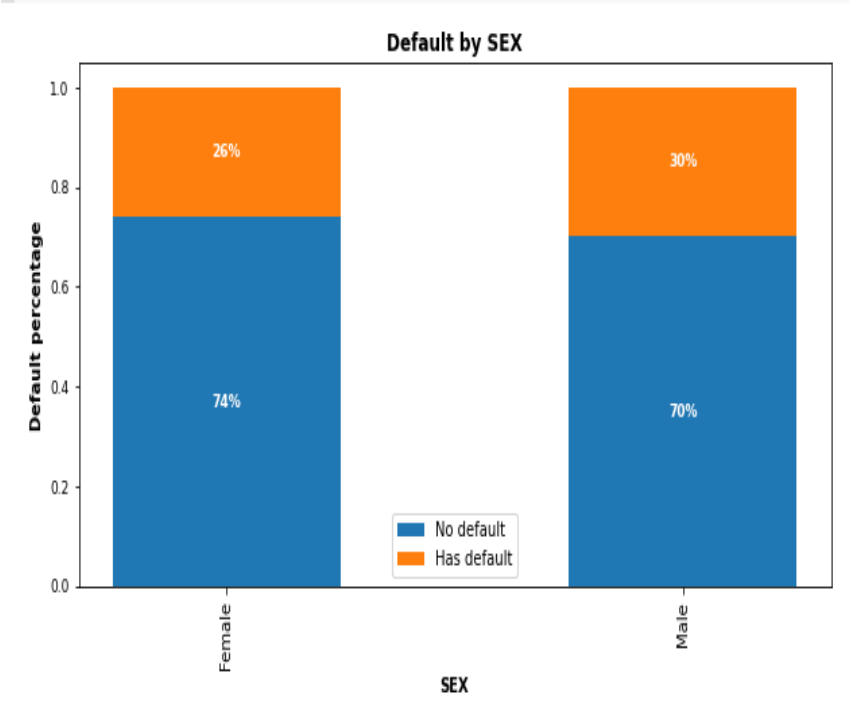
# Feature Engineering

- ✓ This dataset is also imbalanced, with 78% non-default vs 22% default.
- ✓ We use SMOTE for imbalanced data.
- ✓ Do some Feature Selection techniques on the data.
- ✓ Used correlation plot and Pair plot for months and Pair plot for payment for features selections.

# Model Implementation

- ✓ Used Logistic Regression with (hyperparameter tuning).
- ✓ Used RANDOM FOREST with (hyper parameter tuning).
- ✓ Also used some Tree Base Models like Decision Tree with (hyper parameter tuning).
- ✓ We used some Boosting Regression like XGBoost Regression.

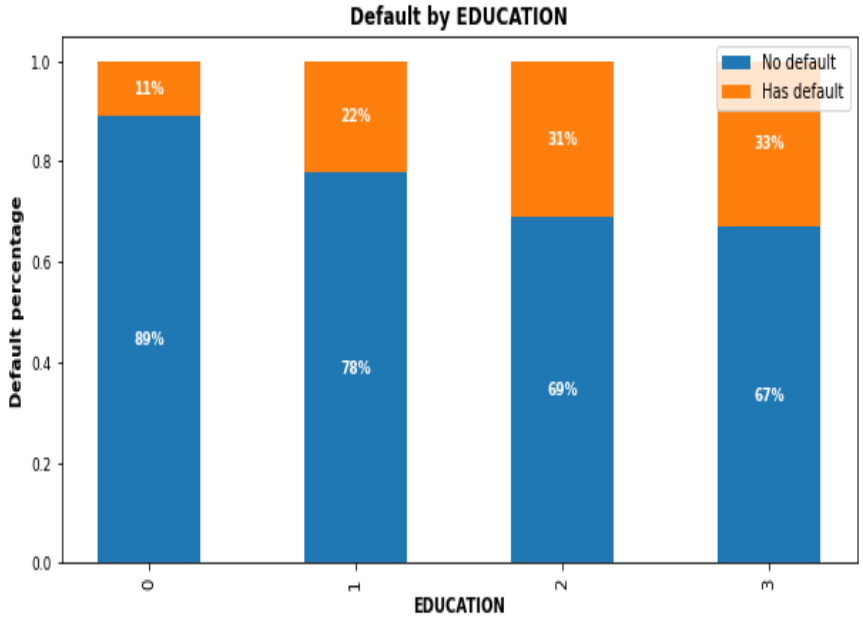
# Learn about "SEX" column




✓30% male have default payment while 26% fema le have default payment, the difference between the m is not significant.

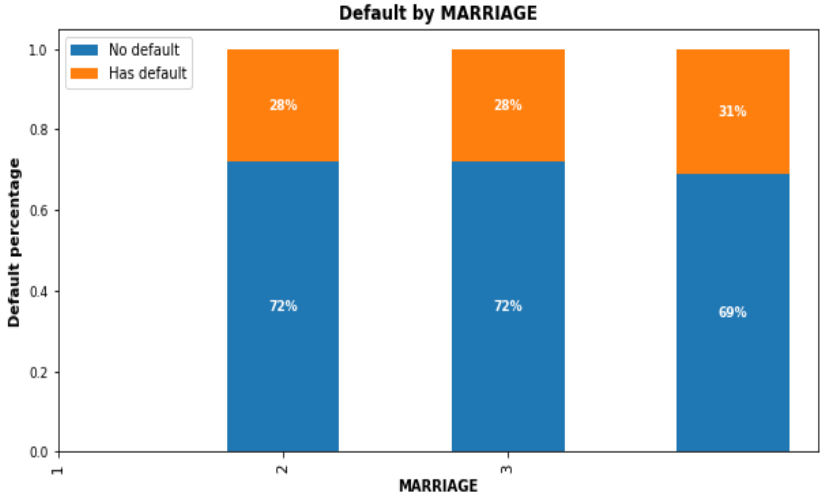
✓Also we can observe that female have more coun t than male

# Learn about "Education" Column



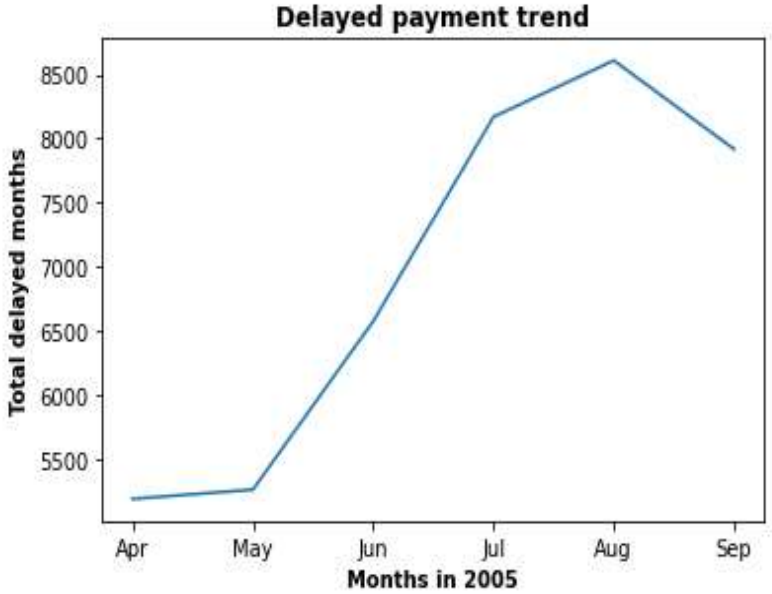

- ✓ The data indicates customers with lower education levels default more.
- ✓ Customers with high school and university educational level had higher default percentages than customers with grad school education did.

# Learn about MARRIGE column



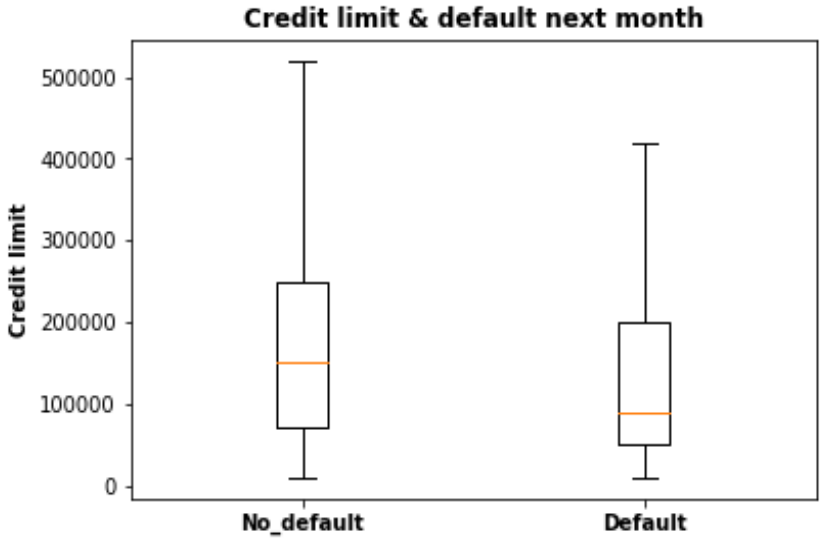

- ✓ 'MARRIAGE' column: what does 0 mean in 'MARRIAGE'? Since there are only 54 observations of 0.
- ✓ we will combine 0 and 3 in one value as 'others'.

# check the change status of the payment in months



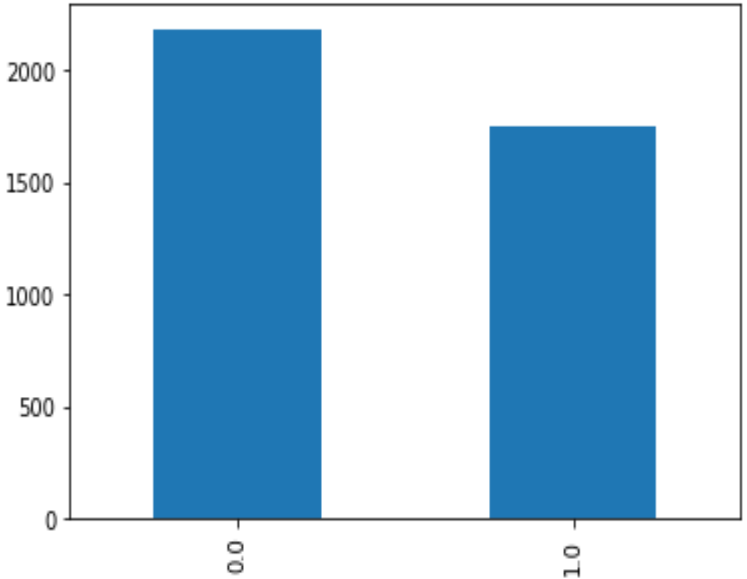

- ✓ There was a huge jump from May 2005 to July 2005
- ✓ When delayed payment increased significantly, then it peaked at August 2005.
- ✓ Things started to getting better in September 2005 .

# Relation between credit limit and the default payment next month



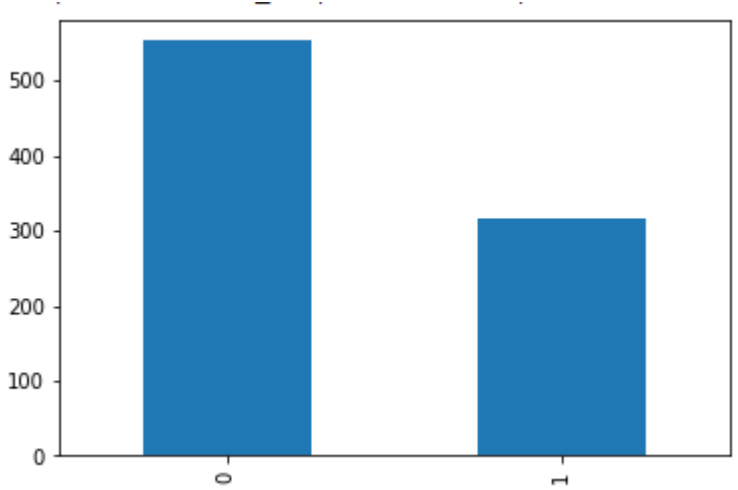

- ✓ Unsurprisingly, customers who had higher credit limits had lower delayed payment rates.

# some bill statement amounts greater than credit limit




- ✓ The common sense is that the bill statement amount shouldn't exceed credit limit.
- ✓ however, there are 3931 customers whose bill amounts are greater than credit limit.
- ✓ Could the difference be late payment interest assuming these customers had delayed payment?

# customers who had no consumption in 6 months then default in the next month




- ✓ There are 870 customers whose bill amount was 0 in 6 months.
- ✓ 317 customers had default payment next month which is against common sense.
- ✓ We will investigate this in the data analysis process.

# EDA Conclusion

✓ Notice code 0 and -2 are in the PAY columns but are not included in the data description.

✓ We found there are 95919 values are listed as 0. Using some google search we found that 0 meaning the payment wasn't due, which makes sense that most customers were using the revolving credit.

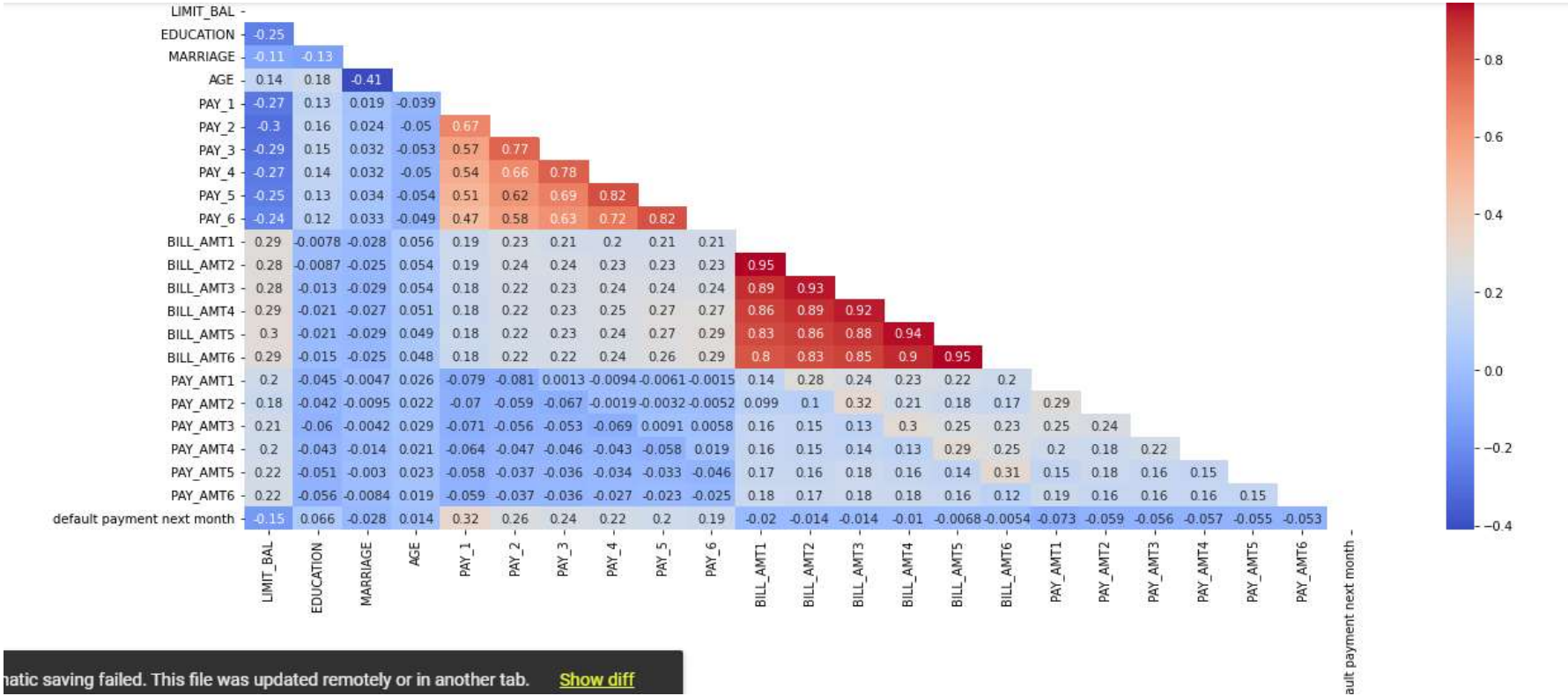
✓Also we found 24415 values as -2 which means No consumption.
✓There are no duplicate IDs or rows.

✓30% male have default payment while 26% female have defaul payment, the difference is not significant.
✓Also we can see Female have more count than male
✓'EDUCATION' column: notice 5 and 6 are both recorded as 'unknown' and there is 0 which isn't explained in the dataset description.

# Feature Engineering

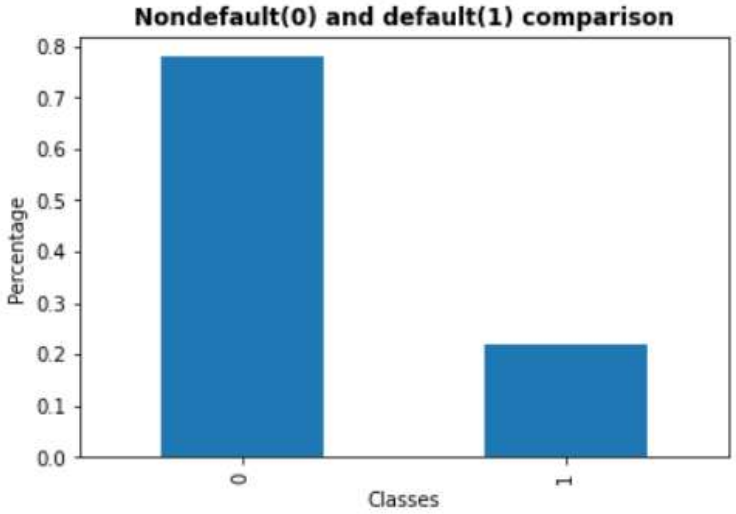
- Used Histogram plot, Pair plot to understand the data.
- Done some feature engineering on the data set.

# Features Selections using Correlation Matrix




matic saving failed. This file was updated remotely or in another tab. Show diff

# Check Class Imbalance

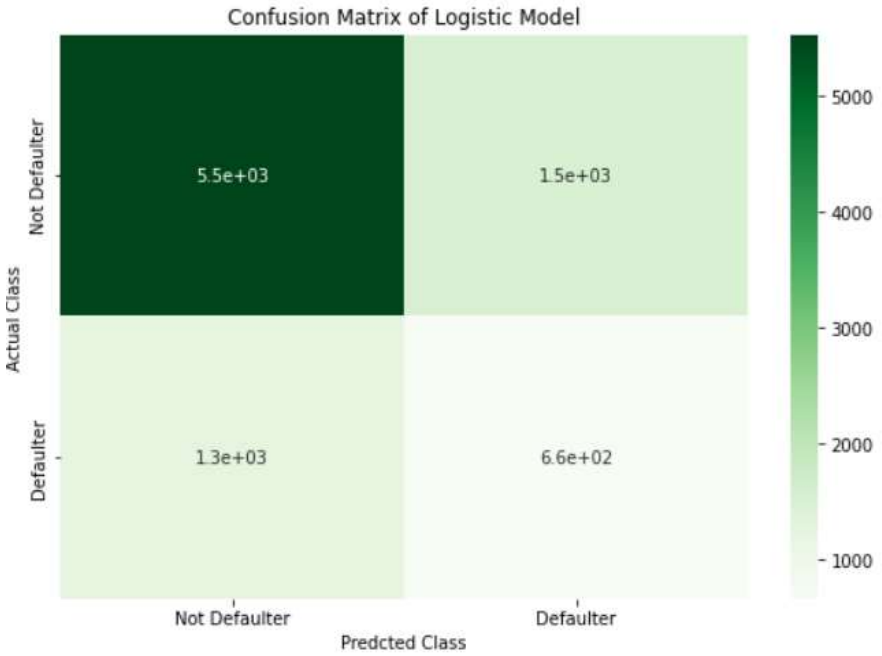



✓With typical default classification problems, we expect imbalanced classes as we know most people will not default.

✓This dataset is also imbalanced, with 78% non-default  vs 22% default.

# Logistic Regression

## (hyper parameter tuning)




✓Training accuracy: 0.5806
✓Testing accuracy: 0.6885

✓Precision score of logistic model: 0.3026

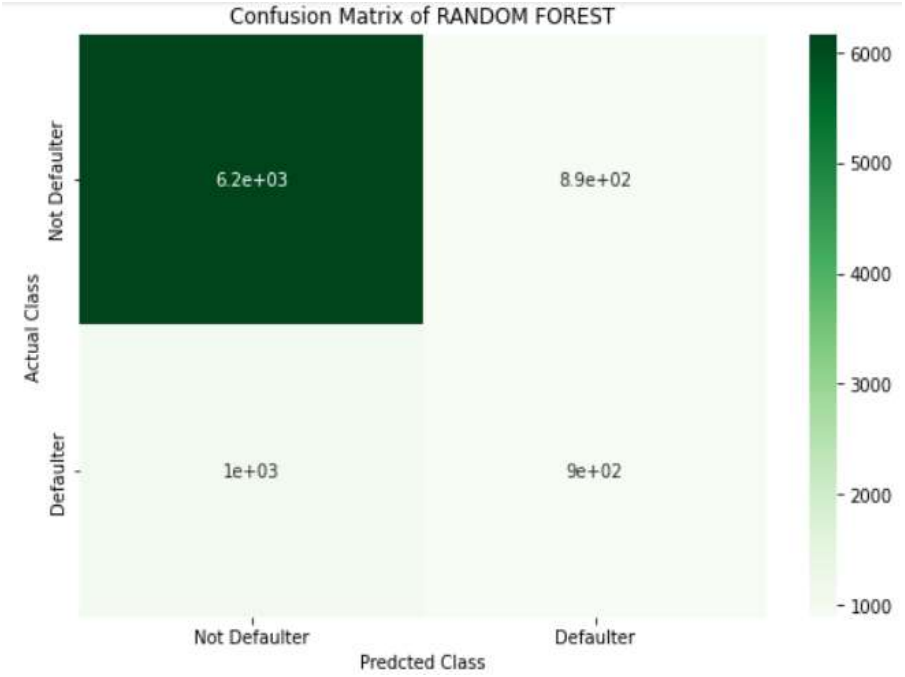
✓Recall score of logistic model: 0.3412

✓F1 score of logistic model: 0.3208

✓ROC AUC score of logistic model: 0.5626

✓Confusion matrix [[5535 1525]
[1278 662]]

# RANDOM FOREST




✓Training Accuracy of Random Forest: 0.9981
✓Testing Accuracy of Random Forest: 0.7853

✓Precision score of logistic model: 0.5022

✓Recall score of logistic model: 0.4628

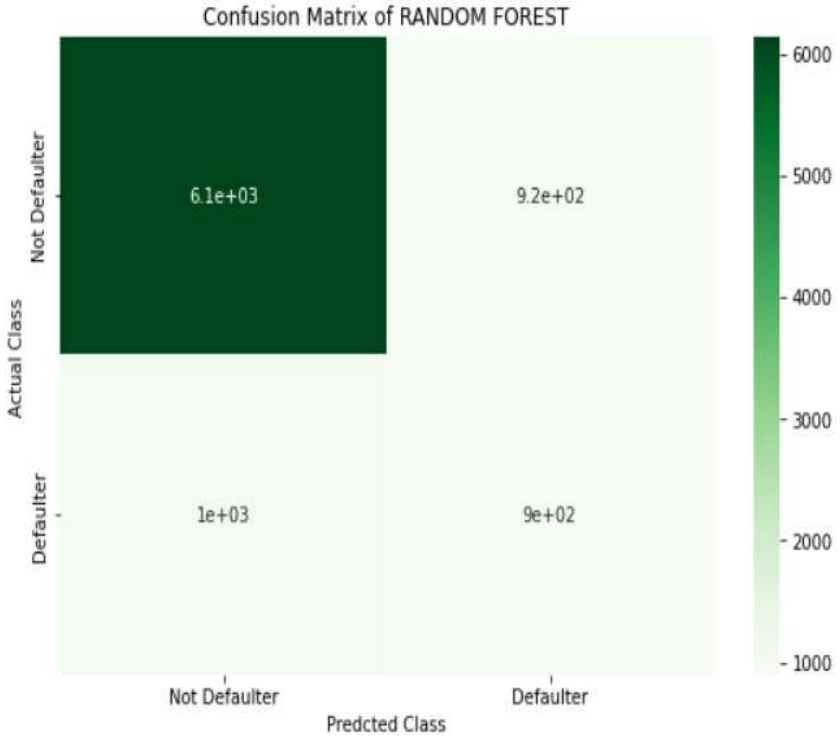
✓F1 score of logistic model: 0.4817

✓ROC AUC score of logistic model: 0.6684

✓Confusion matrix of Random forest
: [[6170 890]
[1042 898]]

# RANDOM FOREST

## (hyper parameter tuning)




The accuracy on train data is 0.9980
The accuracy on test data is 0.7824

Precision score of RANDOM FOREST: 0.4950

Recall score of RANDOM FOREST: 0.4634

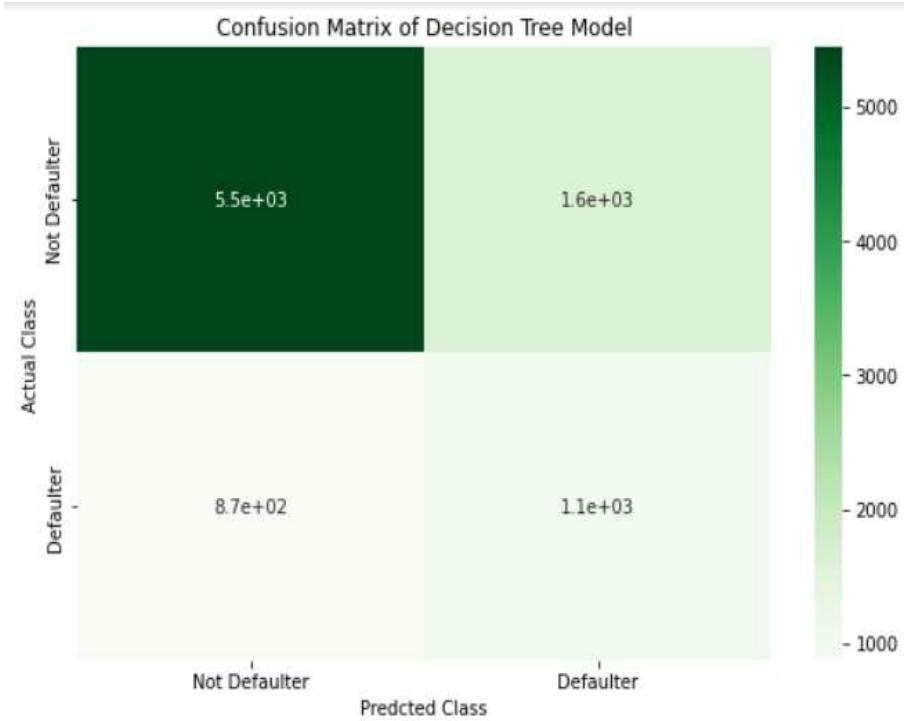
F1 score of RANDOM FOREST: 0.4787

ROC AUC score of RANDOM FOREST: 0.6667

Confusion matrix of RANDOM FOREST model :
[[6143 917]
[1041 899]]

# Decision Tree Classifier

## (hyper parameter tuning)




Training accuracy of decision tree : 0.7702
Testing accuracy of decision tree : 0.7242

Precision score of Decision Tree: 0.3987

Recall score of Decision : 0.55

F1 score of Decision Tree model: 0.4623

ROC AUC score of Decision Tree: 0.6610

Confusion matrix of Decision Tree model :
[[5451 1609]
[ 873 1067]]

# XGBoost

## (hyper parameter tuning)


✓Training Accuracy of XGB Classifier: 0.7917
✓Testing Accuracy of XGB Classifier: 0.763

✓Precision score of XGBoost model: 0.4569

✓Recall score of XGBoost model: 0.5283

✓F1 score of XGBoost model: 0.4900

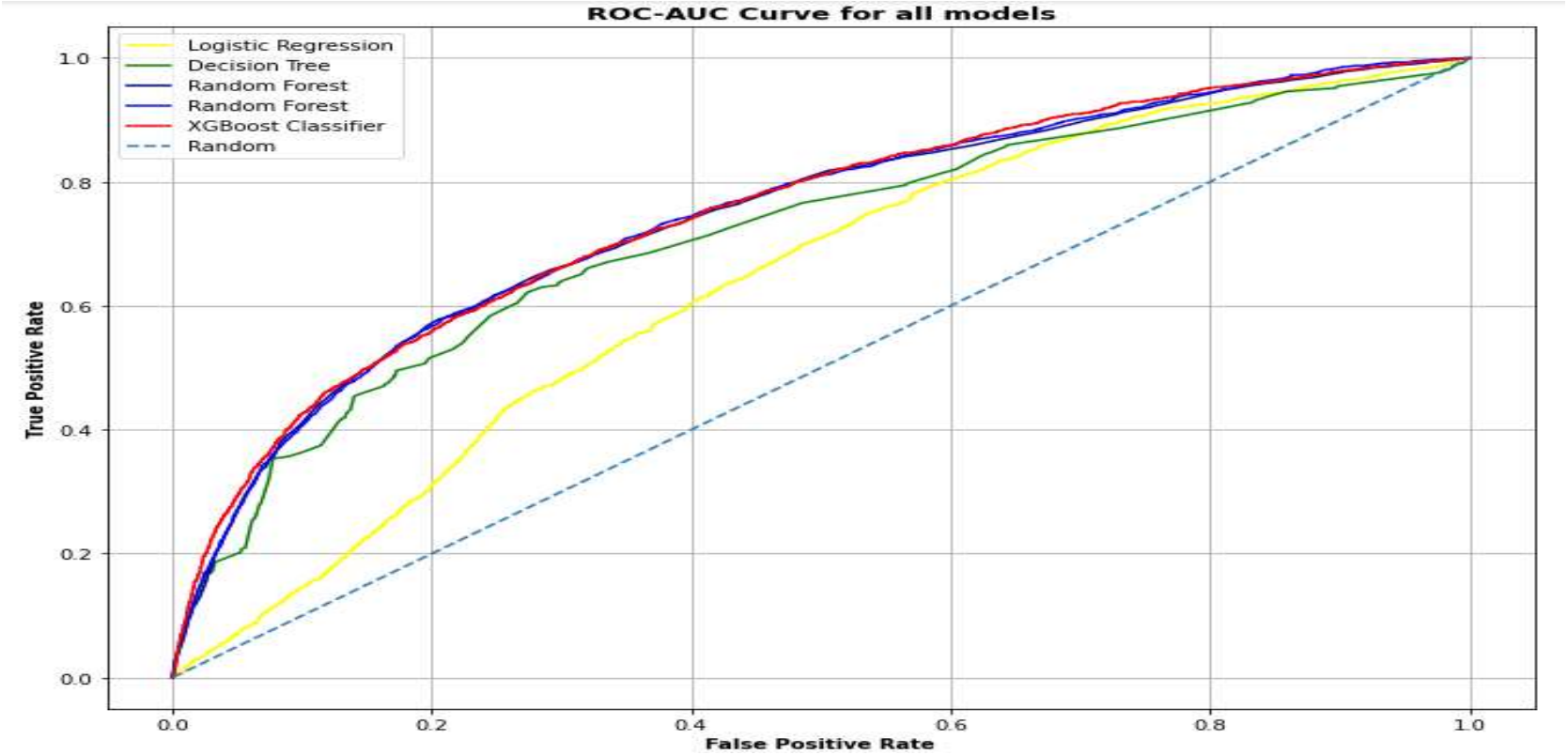
✓ROC AUC score of XGBoost model: 0.6779

✓Confusion matrix of XGBoost model
: [[5842 1218]
[ 915 1025]]

# Data Frame of all Evaluation Matrix with respect to each models

| | Classification Models | Training Accuracy | Testing Accuracy | Precision Score | Recall Score | F1 Score | ROC-AUC Score |
|---|---|---|---|---|---|---|---|
| 0 | Logistic Regression | 0.578754 | 0.687667 | 0.303208 | 0.345876 | 0.323140 | 0.563731 |
| 1 | Random Forest | 0.998129 | 0.785333 | 0.502237 | 0.462887 | 0.481760 | 0.668412 |
| 2 | Random Forest tuning | 0.998099 | 0.782444 | 0.495044 | 0.463402 | 0.478701 | 0.666758 |
| 3 | Decision Tree Classifier | 0.770210 | 0.724222 | 0.398729 | 0.550000 | 0.462305 | 0.661048 |
| 4 | XGBoost Classifier | 0.791738 | 0.763000 | 0.456977 | 0.528351 | 0.490079 | 0.677915 |

# ROC-AUC Curve



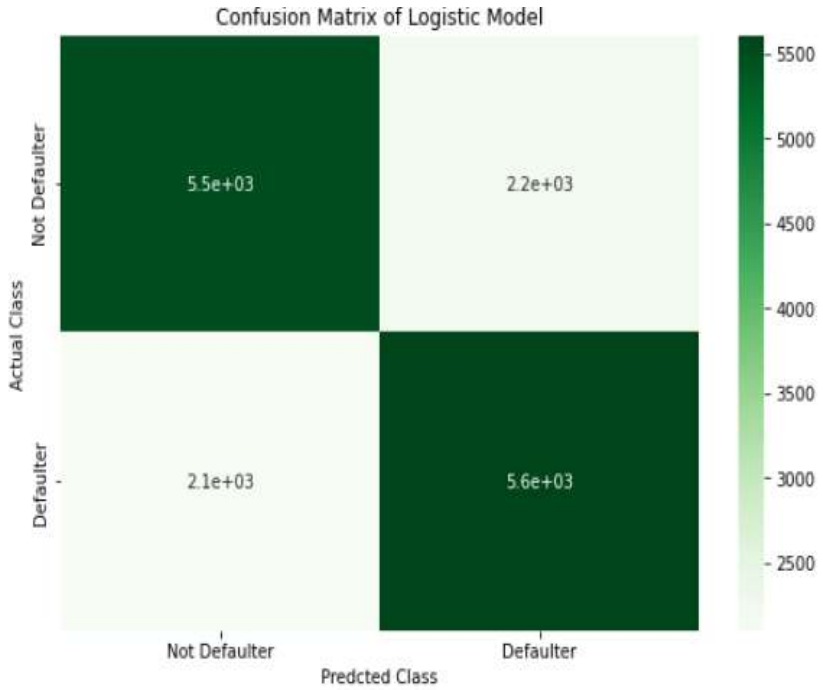
# Try to create model using another way

**Steps for preparing the data for another model:-**

- ✓ Rename default default payment next month to is defaulter
- ✓ Use smote because the data is imbalanced
- ✓ Create another feather named Payement_Value after adding all pay columns
- ✓ Create another feater Dues.
- ✓ Replace 5,6,0 to 4 in education column
- ✓ Replace 0 with 3 in marriage column
- ✓ Using one hot encoding on Education, marriage column
- ✓ Using level encoding on Sex column.

# Logistic Regression

## (hyper parameter tuning)




✓Training accuracy: 0.72
✓Testing accuracy: 0.72

✓Precision score of logistic model: 0.72

✓Recall score of logistic model: 0.71

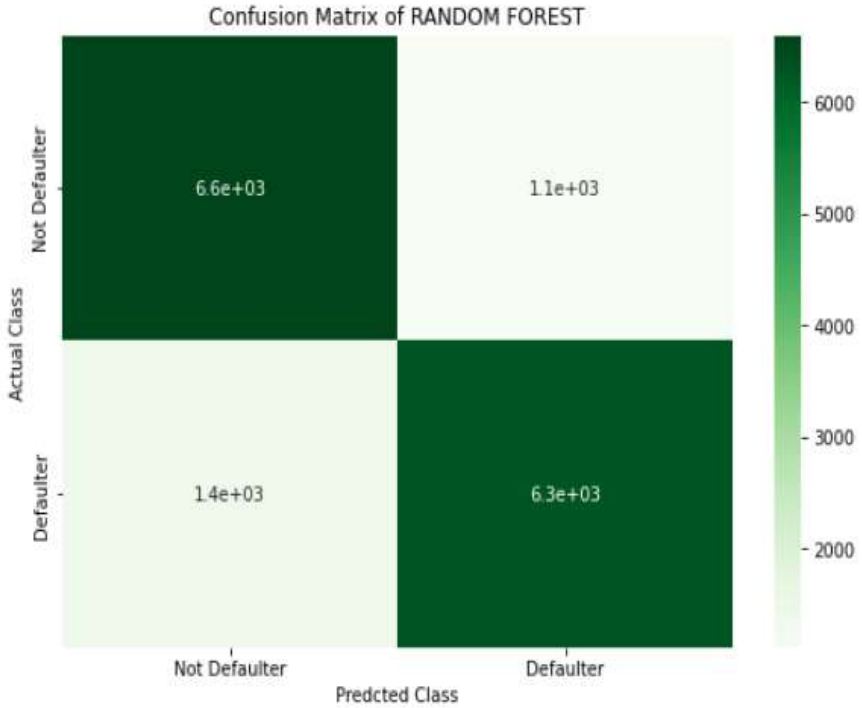
✓F1 score of logistic model: 0.72

✓ROC AUC score of logistic model: 0.72

✓Confusion matrix [[5510 2201]
[2101 5609]]

# RANDOM FOREST

## (hyper parameter tuning)




✓Training accuracy: 0.99
✓Testing accuracy: 0.83

✓Precision score of logistic model: 0.83

✓Recall score of logistic model: 0.84

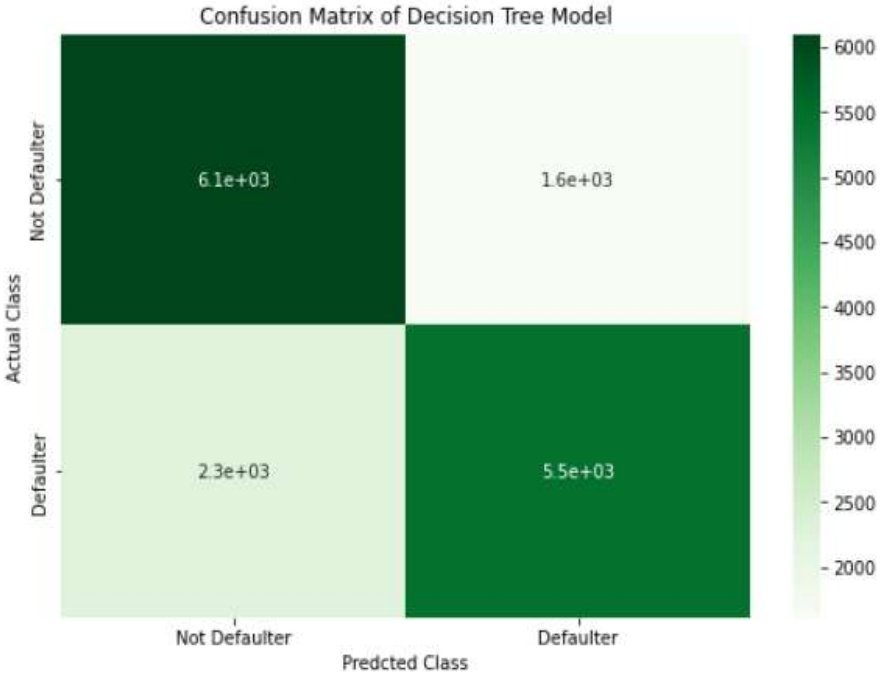
✓F1 score of logistic model: 0.81

✓ROC AUC score of logistic model: 0.83

✓Confusion matrix [[6600 1111]
[1428 6282]]

# Decision Tree Classifier

## (hyper parameter tuning)




✓Training accuracy: 0.77
✓Testing accuracy: 0.74

✓Precision score of logistic model: 0.77

✓Recall score of logistic model: 0.70

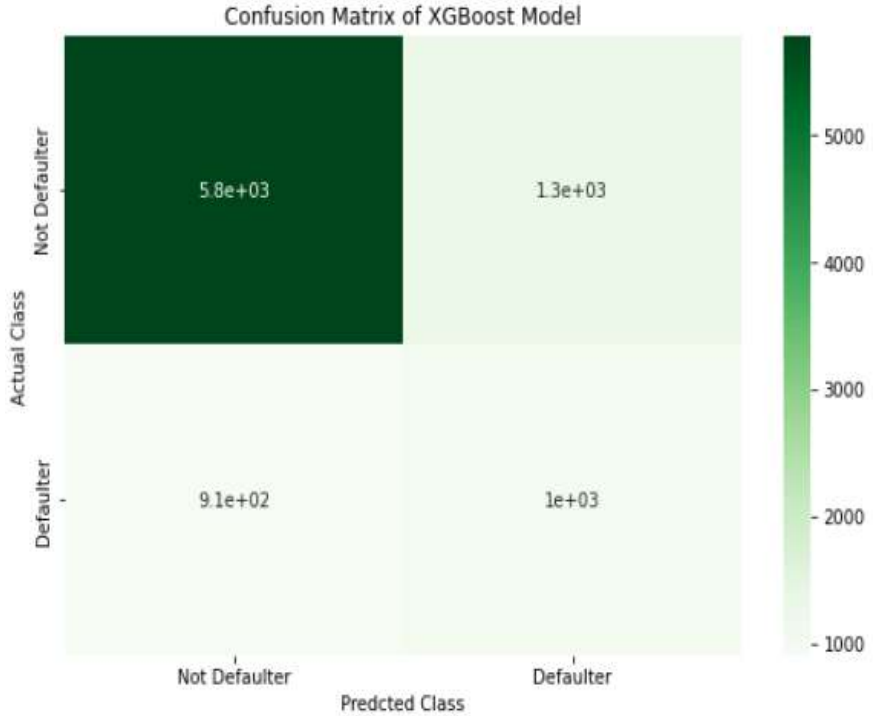
✓F1 score of logistic model: 0.73

✓ROC AUC score of logistic model: 0.74

✓Confusion matrix [[6099 1612]
[2253 5457]]

# XGBoost

## (hyper parameter tuning)




✓Training accuracy: 0.79
✓Testing accuracy: 0.78

✓Precision score of logistic model: 0.80

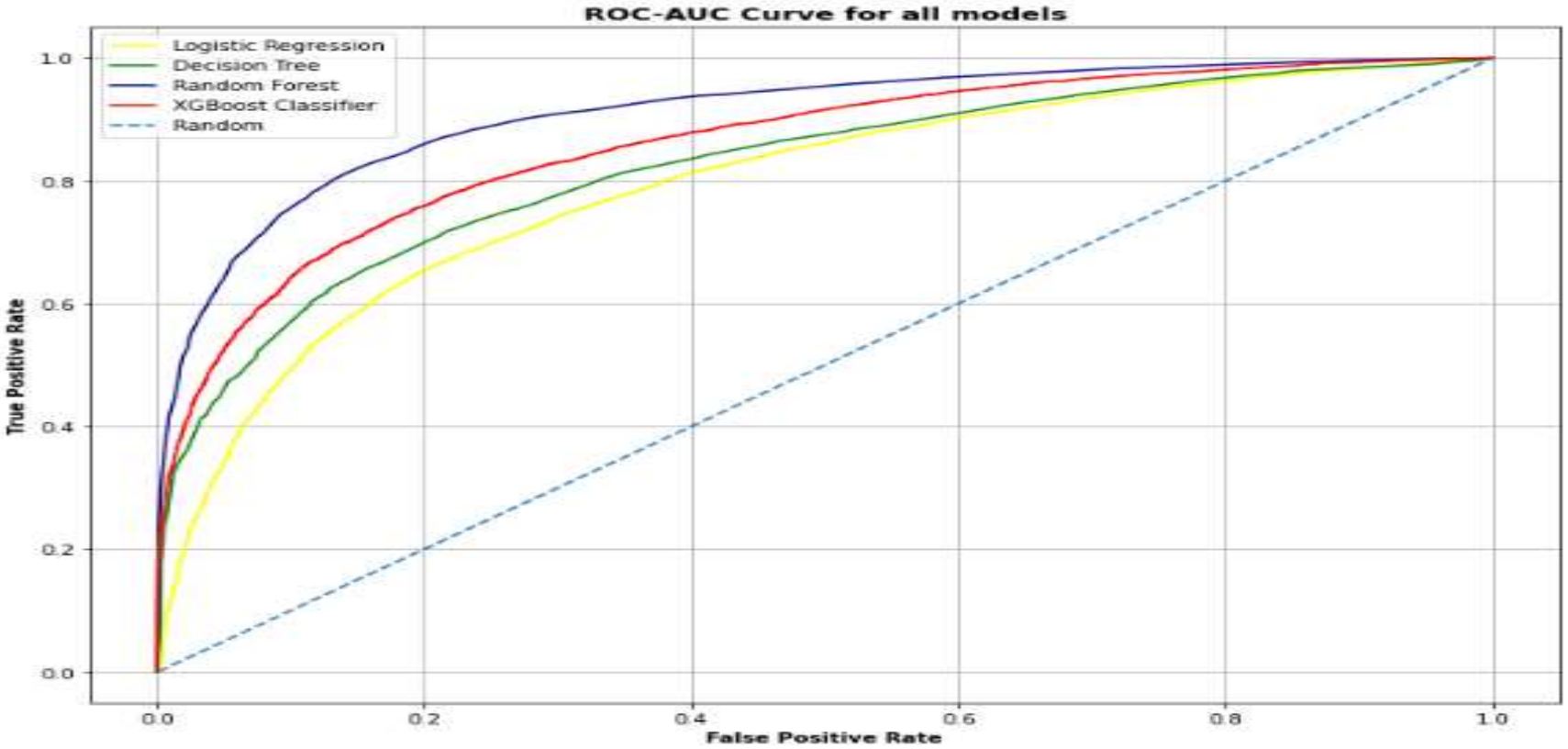
✓Recall score of logistic model: 0.73

✓F1 score of logistic model: 0.77

✓ROC AUC score of logistic model: 0.78

✓Confusion matrix [[6349 1362]
[2015 5695]]

# ROC-AUC Curve

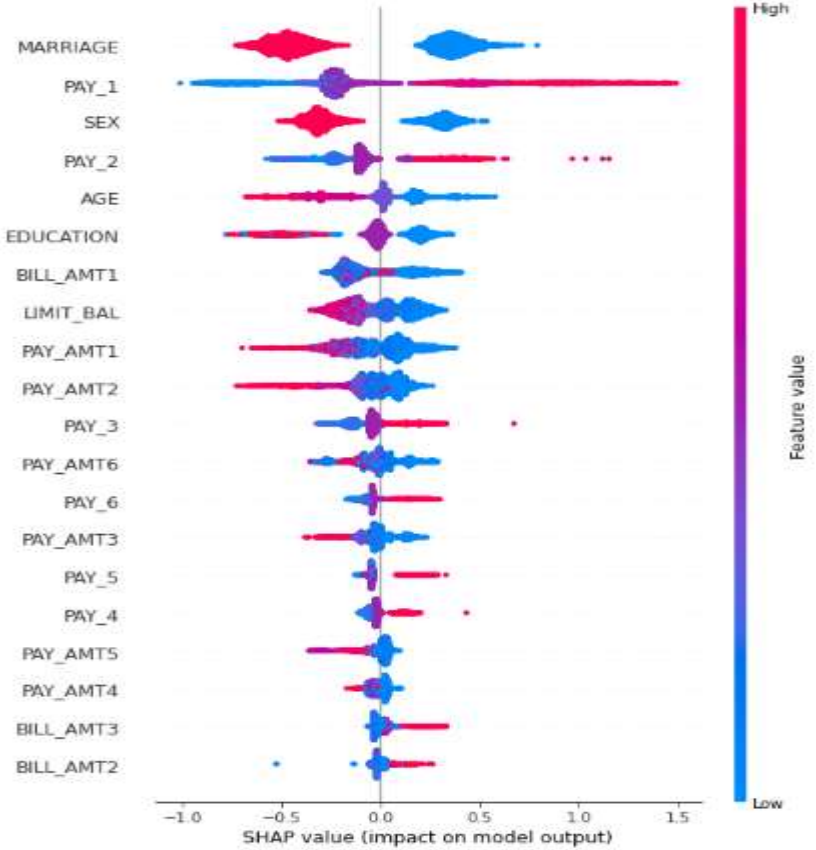


# Data Frame of all Evaluation Matrix with respect to each models

I was surprised to get these result after changing few things in the data pre-processing.

| | Classification Models | Training Accuracy | Testing Accuracy | Precision Score | Recall Score | F1 Score | ROC-AUC Score |
|---|---|---|---|---|---|---|---|
| 0 | Logistic Regression | 0.725205 | 0.721030 | 0.718182 | 0.727497 | 0.722809 | 0.721030 |
| 1 | Random Forest | 0.999393 | 0.835354 | 0.849723 | 0.814786 | 0.467470 | 0.835353 |
| 2 | Decision Tree Classifier | 0.775290 | 0.749368 | 0.771962 | 0.707782 | 0.738480 | 0.749365 |
| 3 | XGBoost Classifier | 0.790302 | 0.781013 | 0.807000 | 0.738651 | 0.771314 | 0.781010 |

# MODEL EXPLAINABILITY

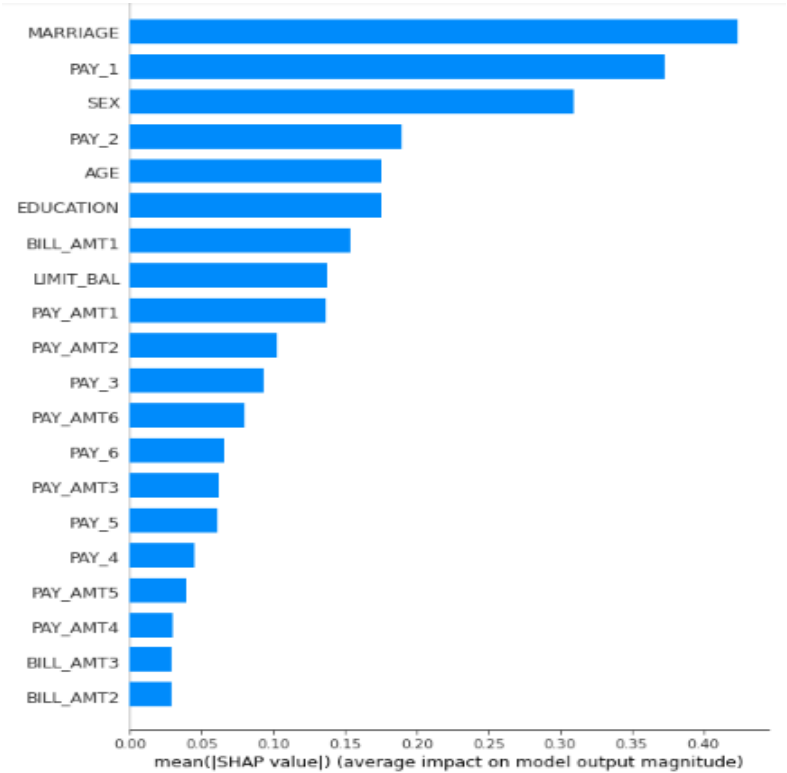



**From the Shap Summary Plot we can explain our complex model:-**

- ✓ Red column shows the high features values
- ✓ Blue column shows low feature values.
- ✓ On X axis there are shape values, positive will tell you about defaulter and negative values will tell customers will not default in next month.

- ✓ On y-axis, features are ordered in decreasing order in sense of importance for the XGBoost model to predict the default.

# Model Explanation




- ✓ 'MARRIGE' is the most important features and pay_1 and SEX are also important.
- ✓ pay_amt4 is the least important feathers

# Conclusion from Model-1

**About models -1:-**

- ✓ Using a Logistic Regression classifier, we can predict with 68.37% accuracy, whether a customer is likely to default next month.
- ✓ Using Decision Tree classifier, we can predict with 73.83% accuracy whether a customer is likely to default next month or not.
- ✓ Using Random Forest, we can predict with 78.38% accuracy whether a customer will be defaulter in next month or not.
- ✓ Using Random Forest (hyper parameter tuning), we can predict with 79.23% accuracy whether a customer will be defaulter in next month or not.
- ✓ By applying XGBoost Classifier with recall 75%, we can predict with 81.60% accuracy whether a customer is likely to default next month.

# Conclusion from Model-2

**About models -2:-**

- ✓ Using a Logistic Regression classifier, we can predict with 72% accuracy, whether a customer is likely to default next month.
- ✓ Using Decision Tree classifier, we can predict with 83% accuracy whether a customer is likely to default next month or not.
- ✓ Using Random Forest, we can predict with 74% accuracy whether a customer will be defaulter in next month or not.
- ✓ By applying XGBoost Classifier with recall 78%, we can predict with 81.60% accuracy whether a customer is likely to default next month.

# Thank You