Capstone Project-3

Credit Card Default Prediction

Individual Contributor

Nargis Nasreen

Content

- Introduction
- Problem Statement
- Data Summary
- EDA on features
- Feature Engineering
- Feature Selection
- Models used for training
- Observation
- Conclusion

Introduction

- We are all aware what is credit card. It is type of payment payment card in which charges are made against a line of credit instead of the account holder's cash deposits. When someone uses a credit card to make a purchase, that person's account accrues a balance that must be paid off each month.

- Credit card default happens when you have become severely delinquent on your credit card payments.Missing credit card payments once or twice does not count as a default. A payment default occurs when you fail to pay the Minimum Amount Due on the credit card for a few consecutive months.

Problem Statement

The main objective of our project is to predict which customer might default in upcoming months.

Data Summary

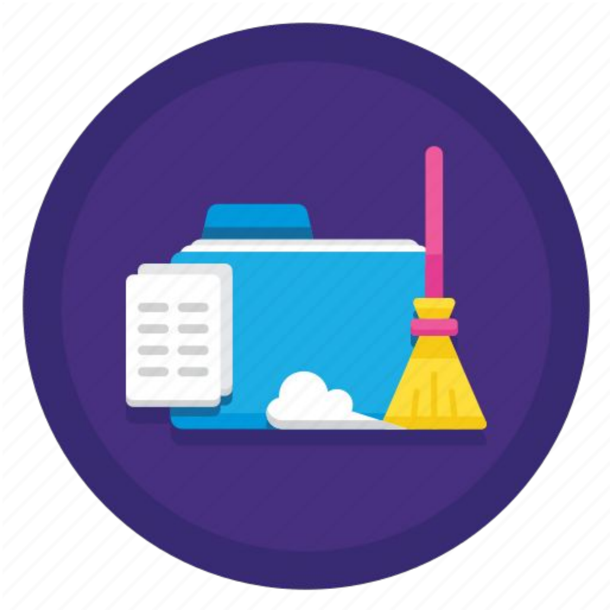
Dataset name: **default of credit card clients.csv**

Shape: **Rows: 30000, Columns: 25**

Features: **ID:** ID of each client, **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit), **SEX**: Gender (1 = male, 2 = female), **EDUCATION**: (1 = graduate school, 2 = university, 3 = high school, 0,4,5,6 = others), **MARRIAGE**: Marital status (0 = others, 1 = married, 2 = single, 3 = others), **AGE**: Age in years, **PAY_X :** History of past payment from April to September, **BILL_AMT_X:** Amount of bill statement from April to September, 2005 (NT dollar), **PAY_AMT_X:** Amount of previous payment from April to September, 2005 (NT dollar)

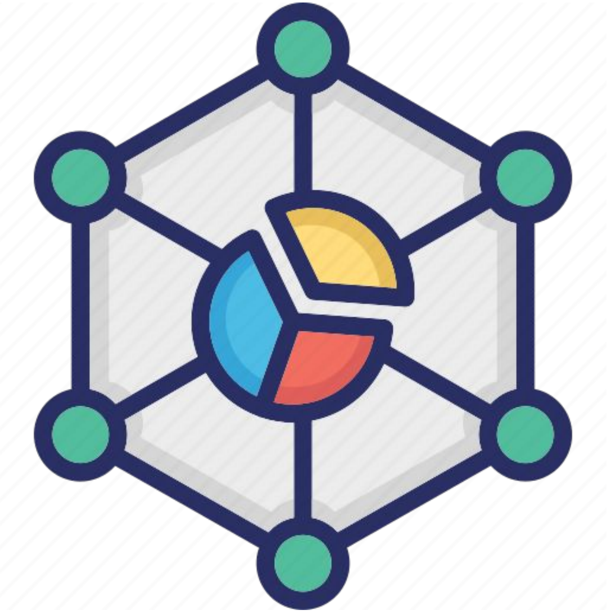
Target Variable: **default payment next month**

Approach Overview



Data Cleaning

Data Exploration



Data Modeling

Data Cleaning

- Dataset has no null values
- There is no duplicates

Dataset is ready for exploratory analysis as we don’t have to do basic cleaning.

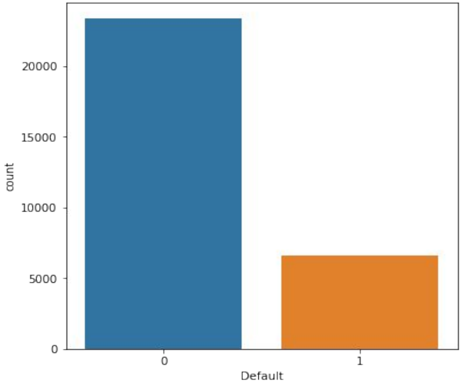
EDA on features

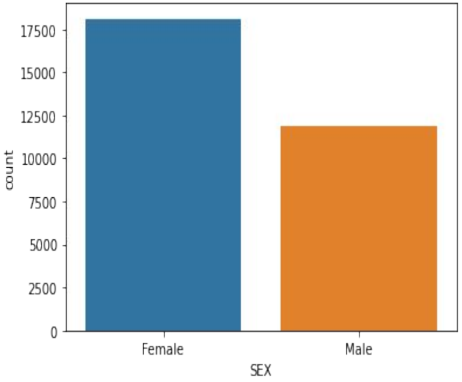
- The ratio of non-defaulters and Defaulters is very high.

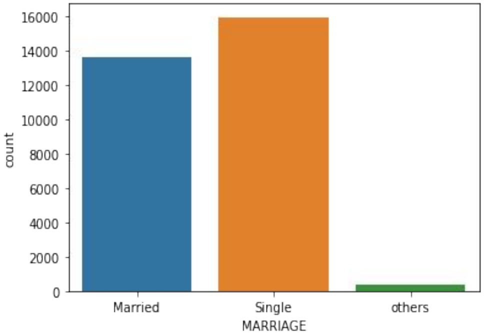
- The number of female users is Higher than male users.

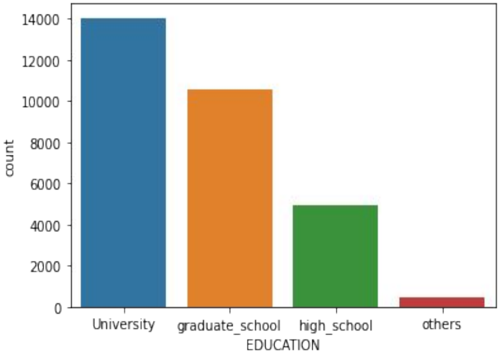
- Most of the users are single.

- Most of the users are still in university. So, higher is the education level lower is the risk.

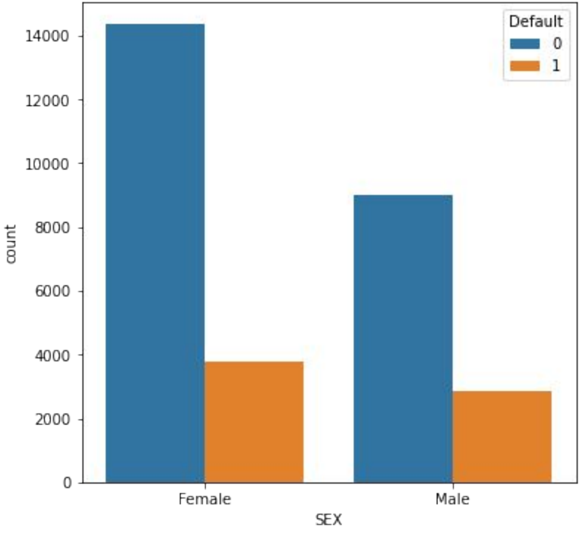


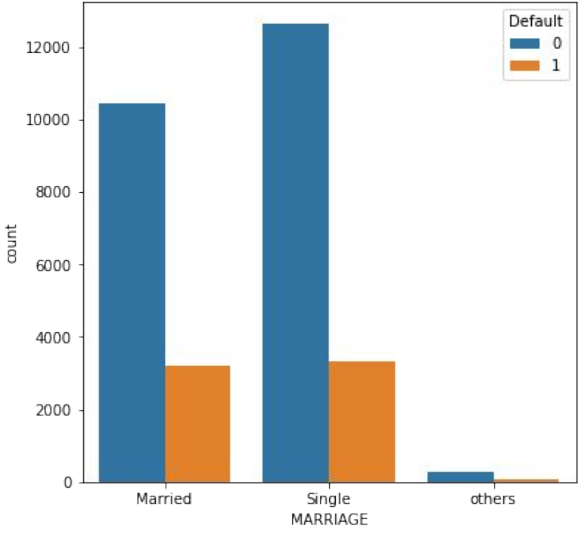


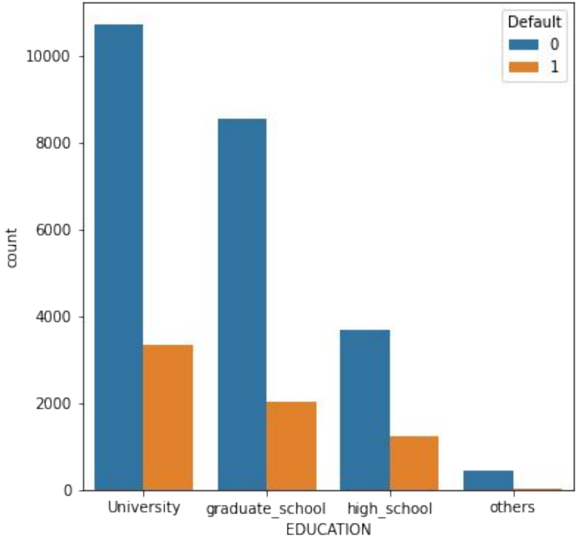




EDA on features

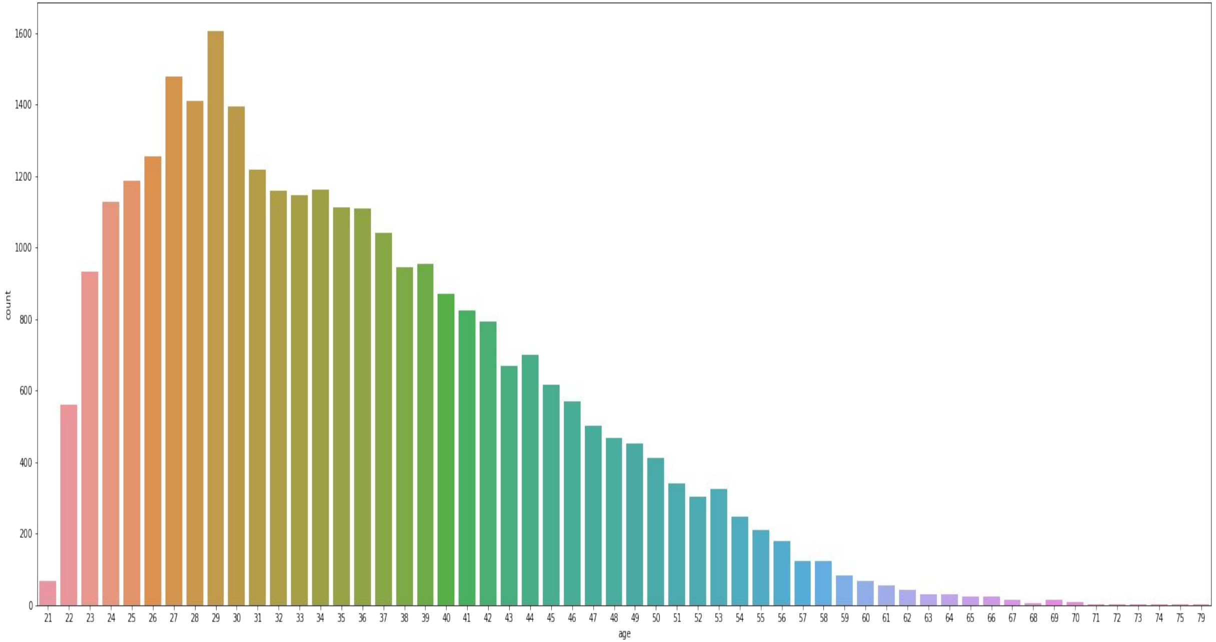


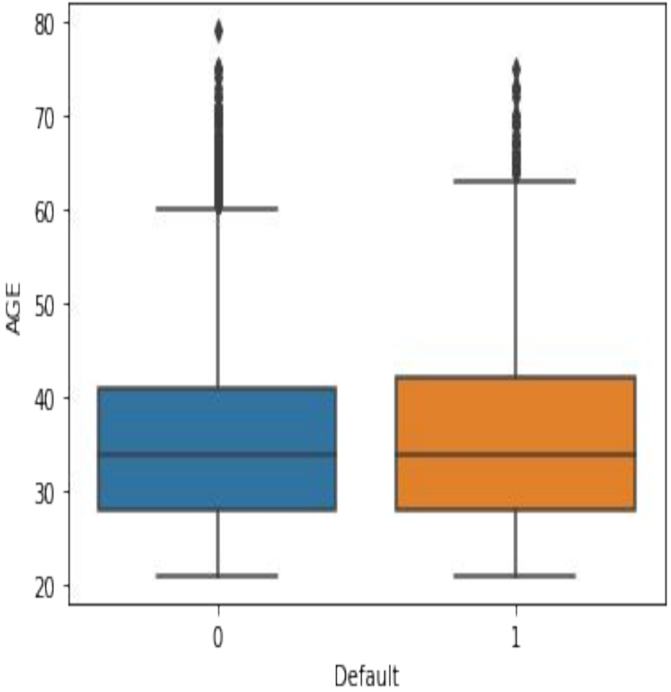




- From this comparison we can conclude that users who are female, single and still in university are more likely to make their payments on time.

EDA on features



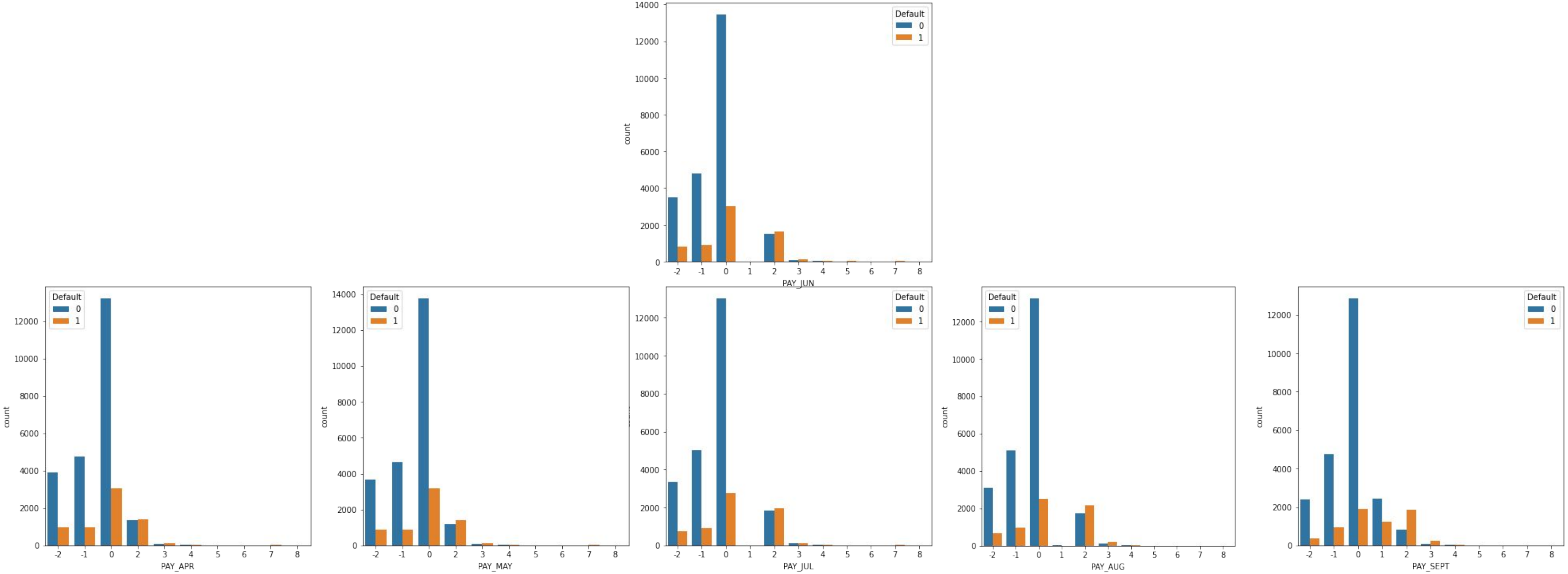


- Most of the users are in the age group of 25-40.

EDA on features

History payment status of users in different months and how many of them are defaulters.

Most of the users have no pending payment in their list.



Modeling Overview

Data Preprocessing:

- Feature selection
- Feature engineering
- SMOTE

Data fitting and tuning:

- Hyper parameter tuning

Model Evaluation:

- Model testing
- precision/recall score
- Compare with other models

Data Preprocessing

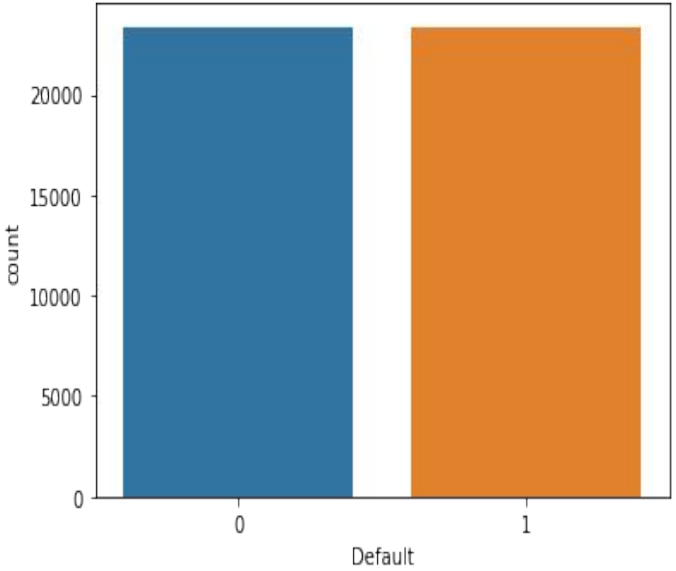
As we have seen earlier that we have imbalanced dataset.

So to remediate Imbalance we are using SMOTE.

Count of 0 and 1 is same now in 'Default' column.

Applied one hot encoding on 'EDUCATION', 'MARRIAGE', 'PAY_SEPT', 'PAY_AUG', 'PAY_JUL', 'PAY_JUN', 'PAY_MAY', 'PAY_APR' columns.

After feature engineering new columns are added, that's why number of features increased to 83.

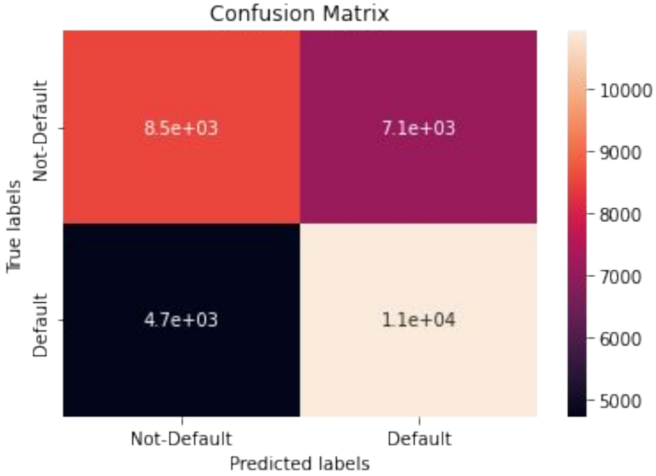


Logistic Regression

Parameters: Penalty= L2, C=0.001

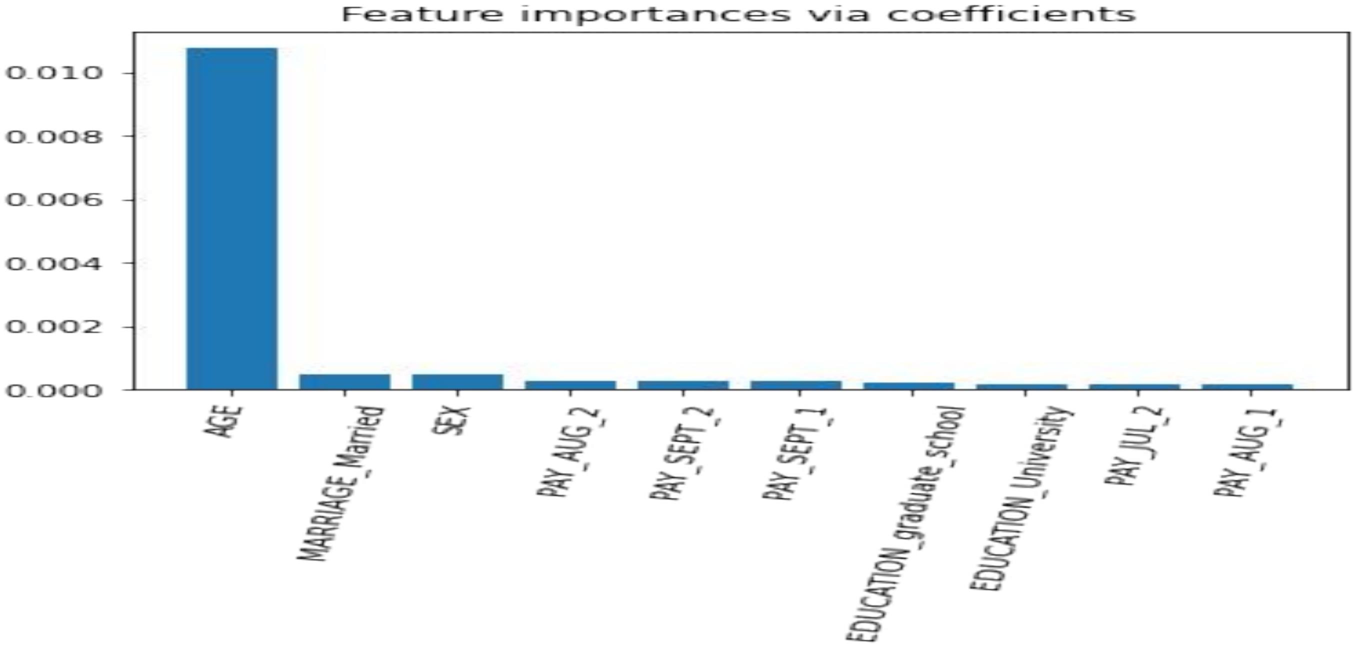
The accuracy on train data is 0.621330692816303
The accuracy on test data is 0.6157836716166267

The precision score on test data is 0.6907911802853437
The recall score on test data is **0.6006541107477162**
The f1 score on test data is 0.6425770646075889
The roc score on test data is 0.6184533904364153



True Negative and False Positive rate is quite high because according to recall score it is only predicting 60 correct out of 100.

Logistic Regression



- For logistic regression 'Age' feature is of highest importance.
- Overall performance of logistic regression is not that good.

Decision Tree Classifier

Parameters: max_depth = 20, min_samples_split = 0.1

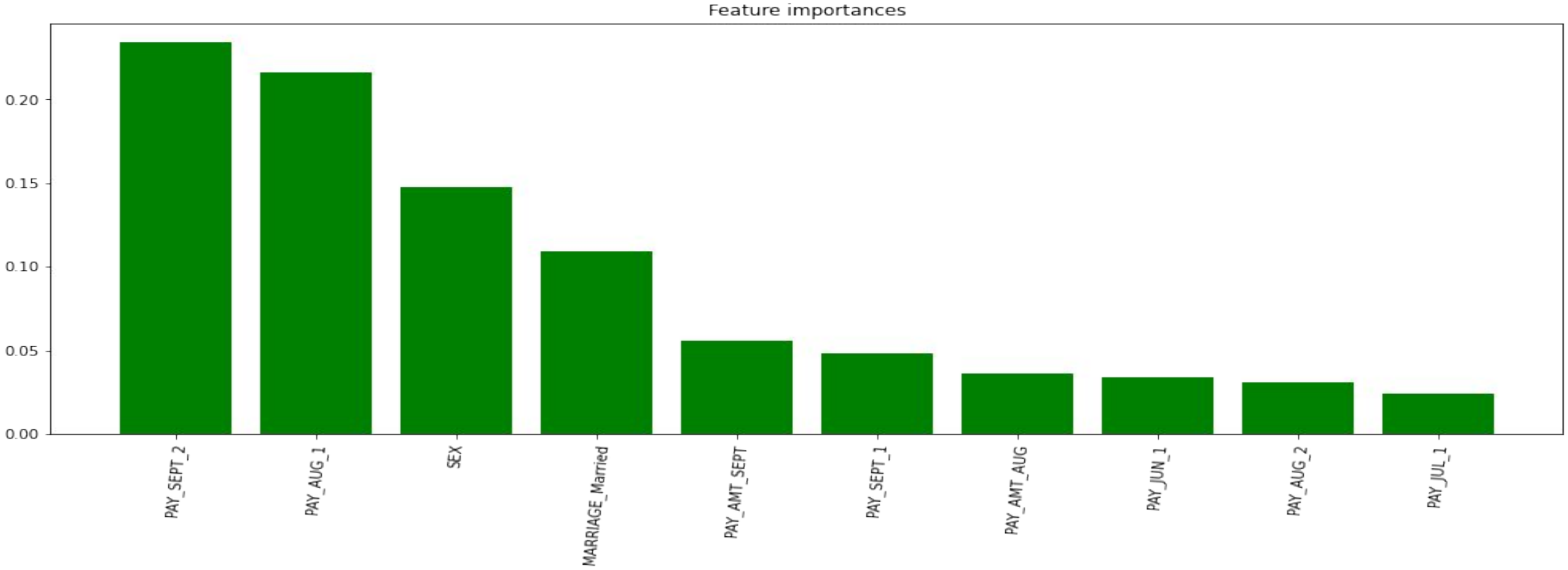
The accuracy on train data is 0.7121410547161977
The accuracy on test data is 0.7109136891252189

The precision score on test data is 0.6016861219195849
The recall score on test data is **0.7698307334882177**
The f1 score on test data is 0.6754513686662784
The roc score on test data is 0.7214773678085047

Decision tree classifier giving better results than logistic regression and no overfitting because we tuned the data.

Decision tree classifier performing better as out of 100 it is predicting 76 correct according to recall score.

Decision Tree Classifier



- **PAY_SEPT_2 is the most important feature followed by PAY_AUG_1 and SEX.**
- **Still performance of decision tree classifier is not satisfactory.**

Random Forest Classifier

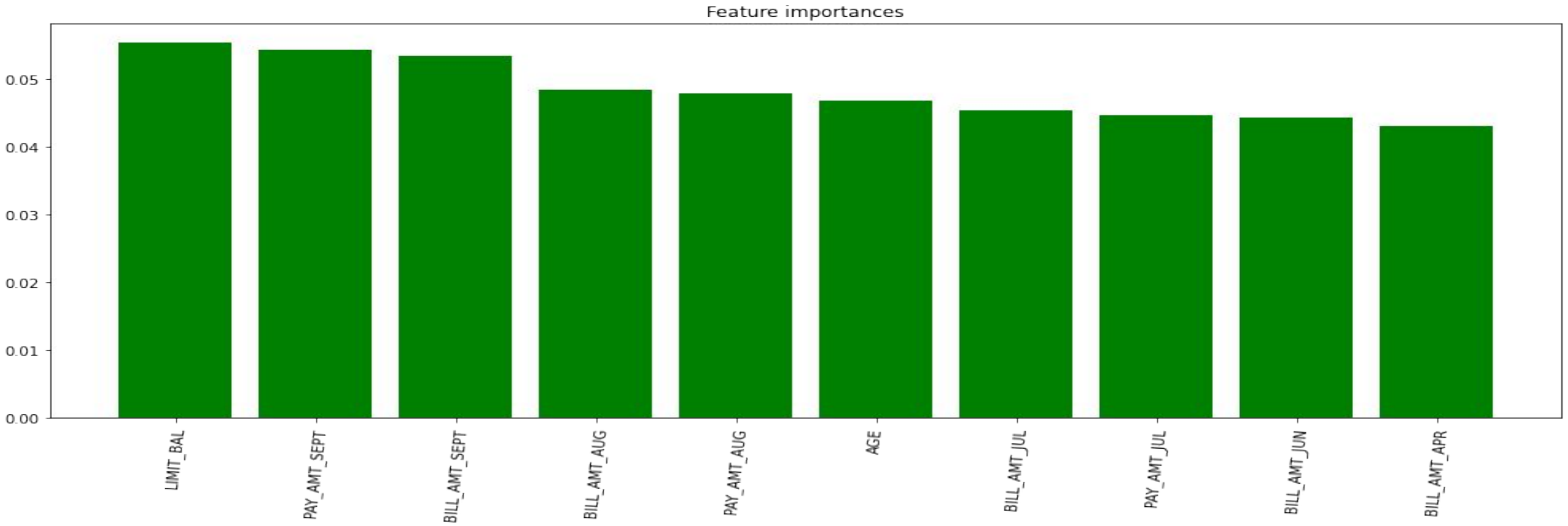
Parameters: max_depth = 30, n_estimators = 200

```
The accuracy on train data is  0.9983070878717 22
The accuracy on test data is  0.8354192335127424

The precision score on test data is  0.8029831387808042
The recall score on test data is  0.8586685159500693
The f1 score on test data is  0.829892761394102
The roc score on test data is  0.8368363892623322
```

Random forest classifier has performed the best so far as it is predicting 85 correct out of 100.

Random Forest Classifier



- **LIMIT_BAL is the most important feature followed by PAY_AMT_SEPT and BILL_AMT_SEPT**

Models Comparison

Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
Logistic Regression	0.621331	0.615784	0.690791	0.600654	0.642577
Random Forest CLf	0.998307	0.835419	0.802983	0.858669	0.829893
Decision Tree CLF	0.712141	0.710914	0.601686	0.769831	0.675451

Observation

- Using a Logistic Regression classifier, we can predict with ~60% accuracy, whether a customer is likely to default next month.

- Using a Decision Tree classifier, we can predict with ~76% accuracy, whether a customer is likely to default next month.

- Using a Random Forest classifier, we can predict with ~85% accuracy, whether a customer is likely to default next month.

- It mean out of 100 defaulters 85 will be correctly caught by Random Forest Classifier.

- Random Forest outperforms Logistic Regression and Decision Tree if measured on their F1 scores.

- But it's better to go with decision tree as random forest is overfitting the model or we can try different modeling techniques.

Conclusion

- The strongest predictors of default are the PAY_AMTX, the LIMIT_BAL & the BILL_AMT_X, PAY_X, AGE and SEX on the basis of models used.

- Demographics: We see that being Female, More educated, Single and between 30-40 years old means a customer is more likely to make payments on time.

Thank you