# Supervised ML-Classification

## Credit Card Default Prediction

## Presentation by

Behara Sai Keerthi

# Points for Discussion

- Problem Statement
- Introduction
- Data Inspection
- Exploratory Data Analysis
- Data Preprocessing
- Data Preparation using SMOTE
- Classification Models Building
- Conclusion

# Problem Statement

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.

The main objective is to build a predictive model, which could help them in predicting the customers who might default in upcoming months.

# Introduction

The given dataset consist of 30000 rows and 25 columns, and there were no null values. The columns description is as follows:

1. ID: ID of each client, categorical variable
2. LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
3. SEX: Gender, categorical variable (1=male, 2=female)
4. EDUCATION: level of education, categorical variable (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
5. MARRIAGE: Marital status, categorical variable (1=married, 2=single, 3=others)
6. AGE: Age in years, numerical variable

# Introduction

7-12 - PAY_0: Repayment status in September 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above) , PAY_2:August, PAY_3: July, PAY_4: June 2005, PAY_5:  May, PAY_6: April

13-17- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar), BILL_AMT2: August, BILL_AMT3: July, BILL_AMT4: June, BILL_AMT5: May, BILL_AMT6: April

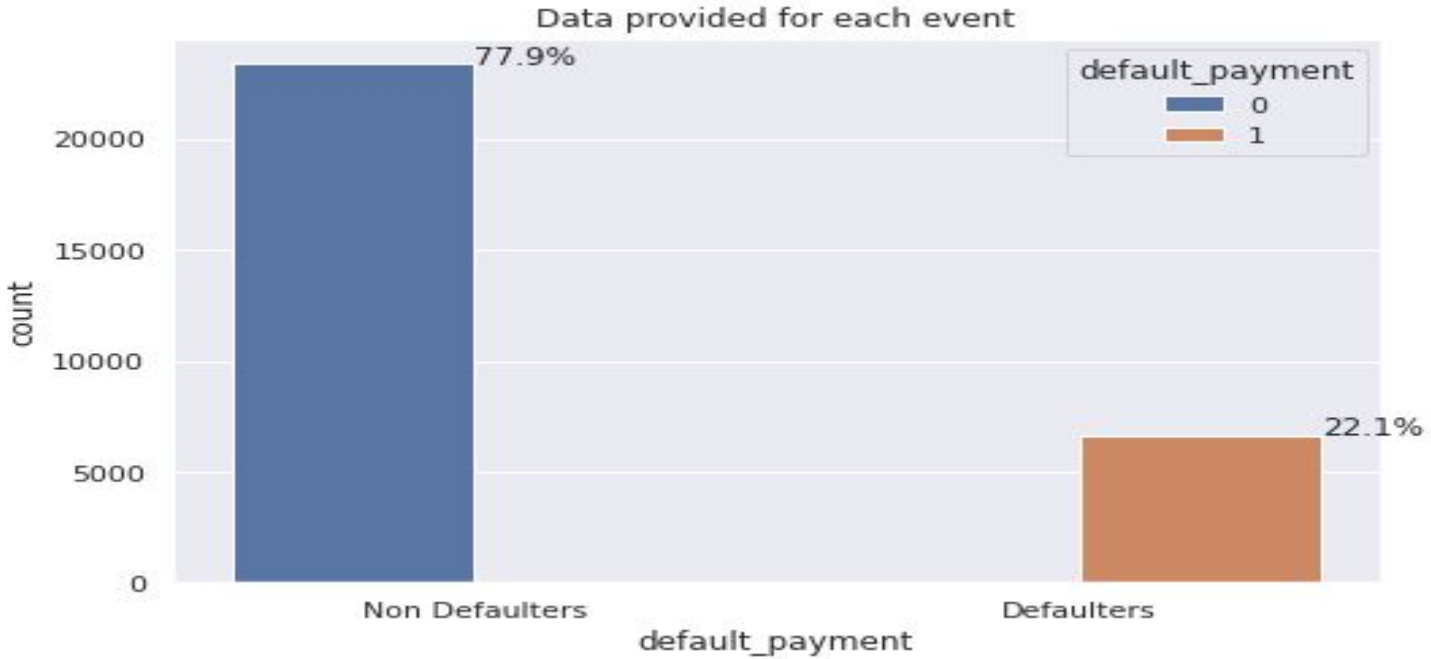
18-24- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar), PAY_AMT2: August, PAY_AMT3: July, PAY_AMT4:June, PAY_AMT5: May, PAY_AMT6:April

25. Default_payment: indicate whether the credit card holders are defaulters or non-defaulters (1=yes, 0=no)

# Data Inspection

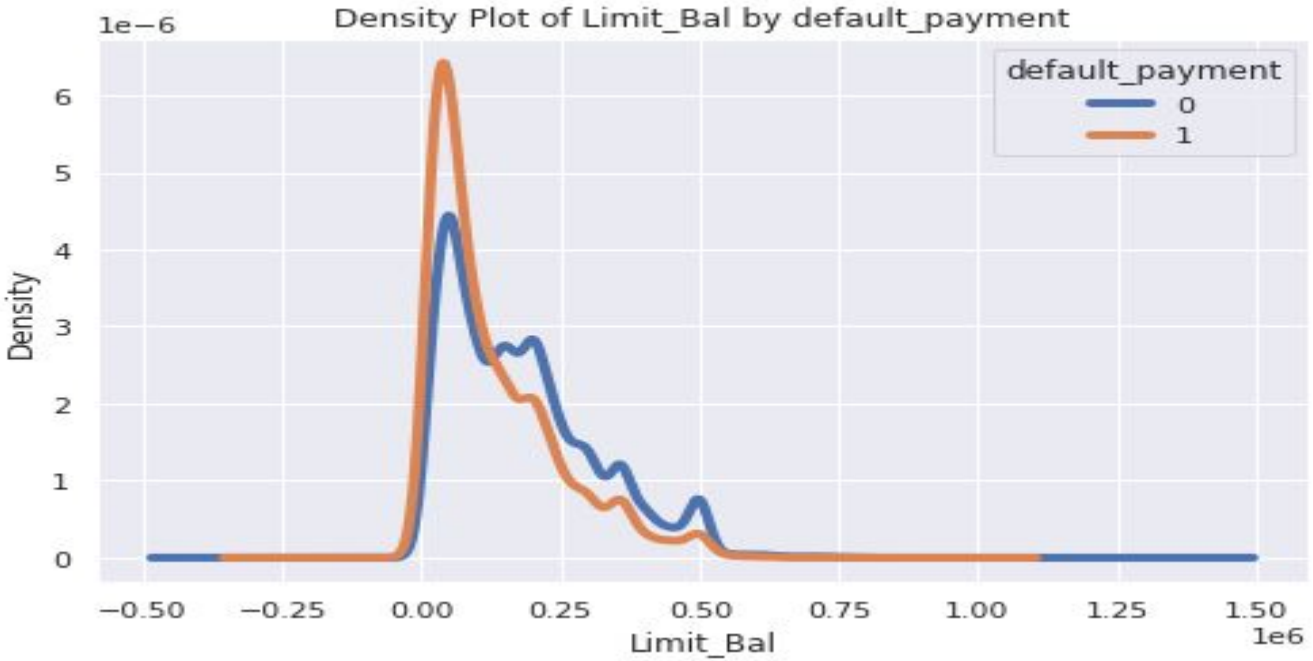
1. Here we have checked the head, tail, description,columns, information and shape of the data
2. Column names were changed
3. Datatype was object which was changed to integer
4. Checked for null values

# Exploratory Data Analysis



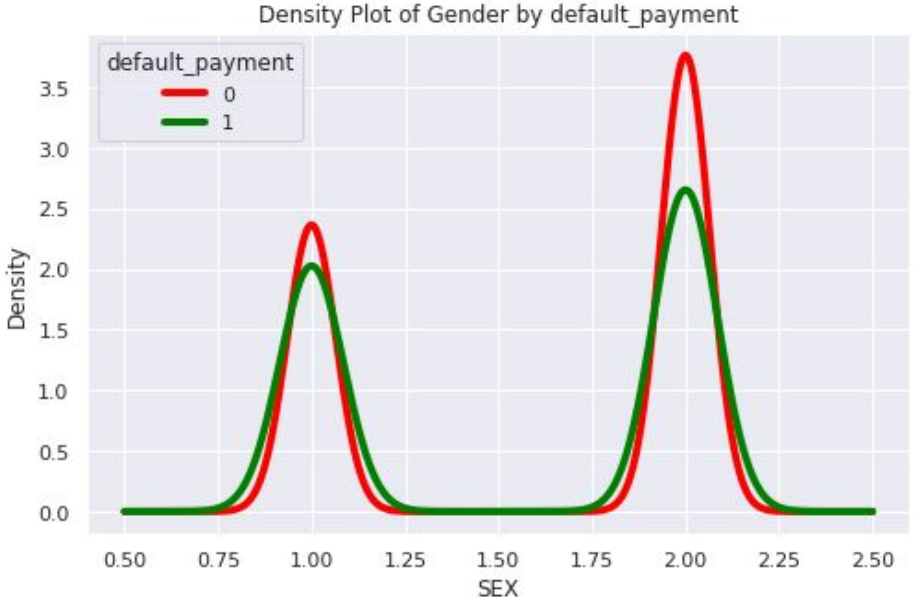

Since our target variable is default payment we have checked the number of defaulters and non-defaulters. There is clearly a class imbalance detected which was fixed using sampling technique.

# Exploratory Data Analysis

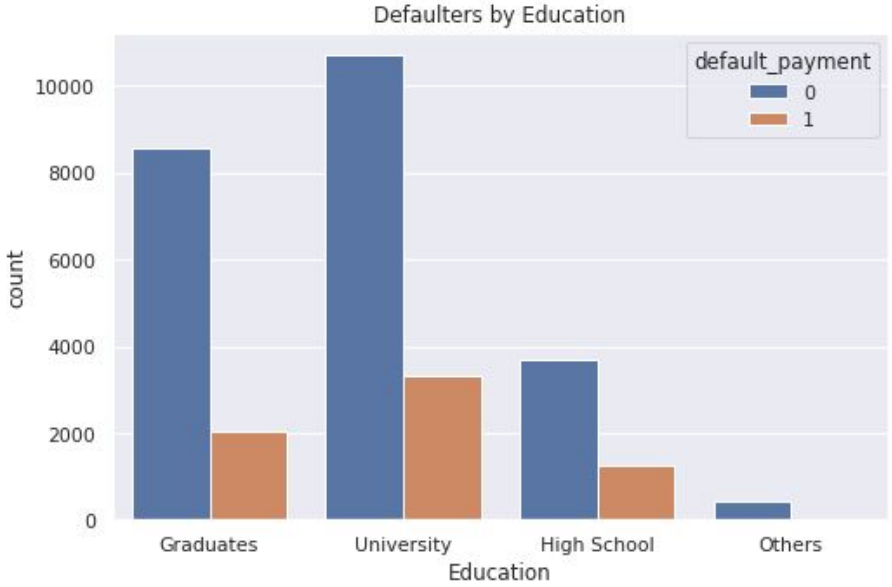



We can observe that the defaulters are more when limit balance is low.

# Exploratory Data Analysis

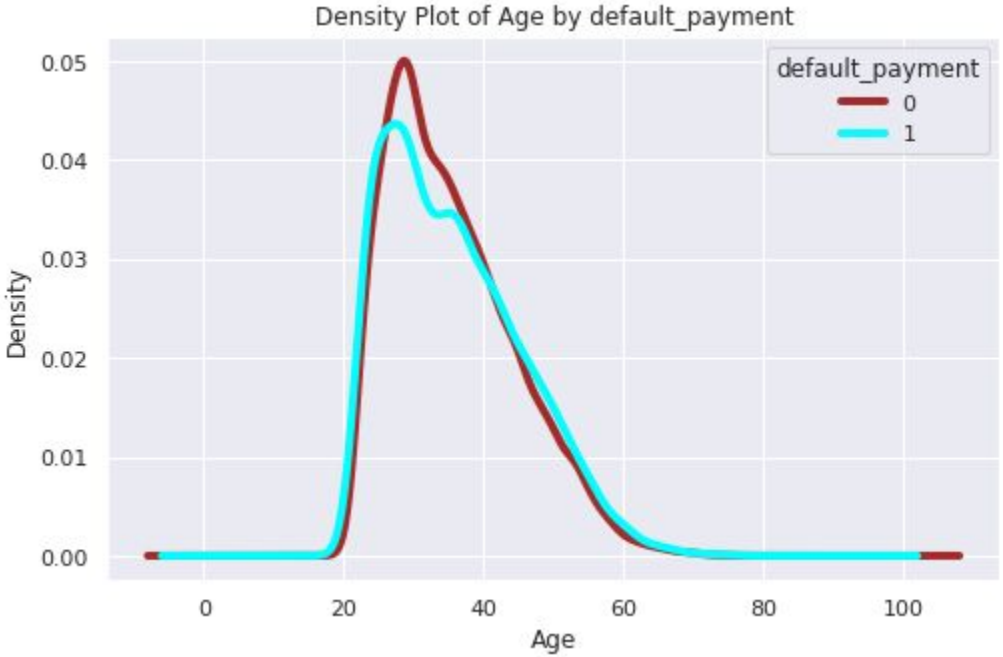



There is no much difference in defaulters between male and female but female customers are more likely to be a defaulter.

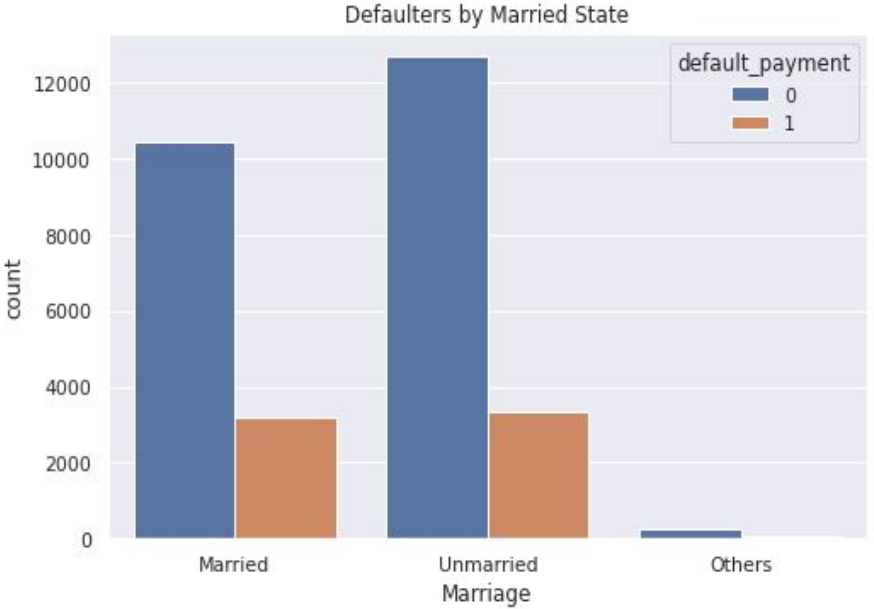



The defaulters are university level people followed by graduates.This could be because there are more number of university going customers.

# Exploratory Data Analysis

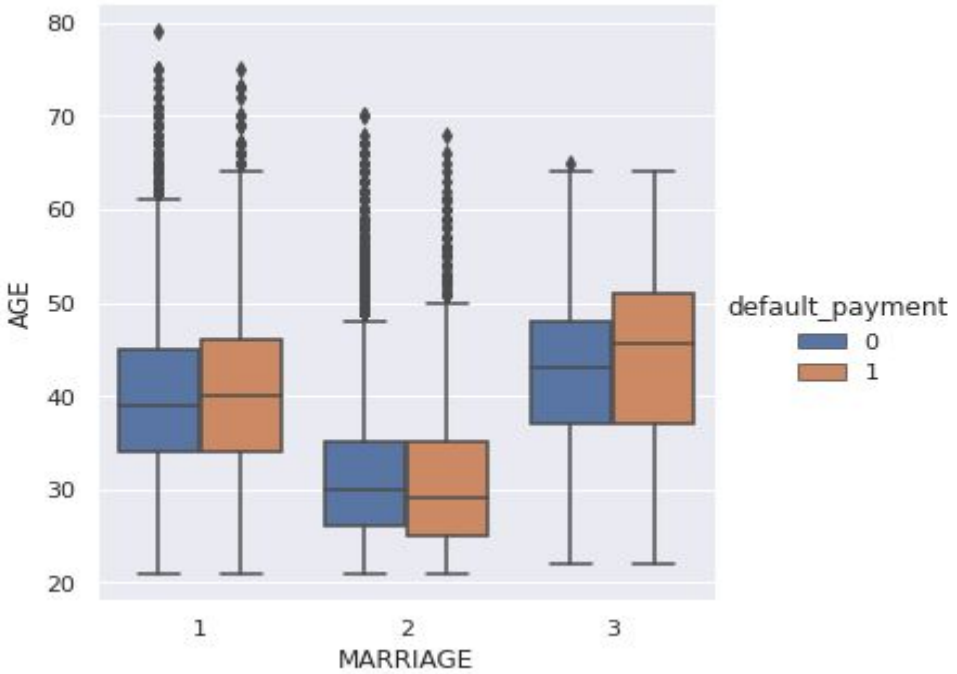



Most defaulters are in the range 30-40 years

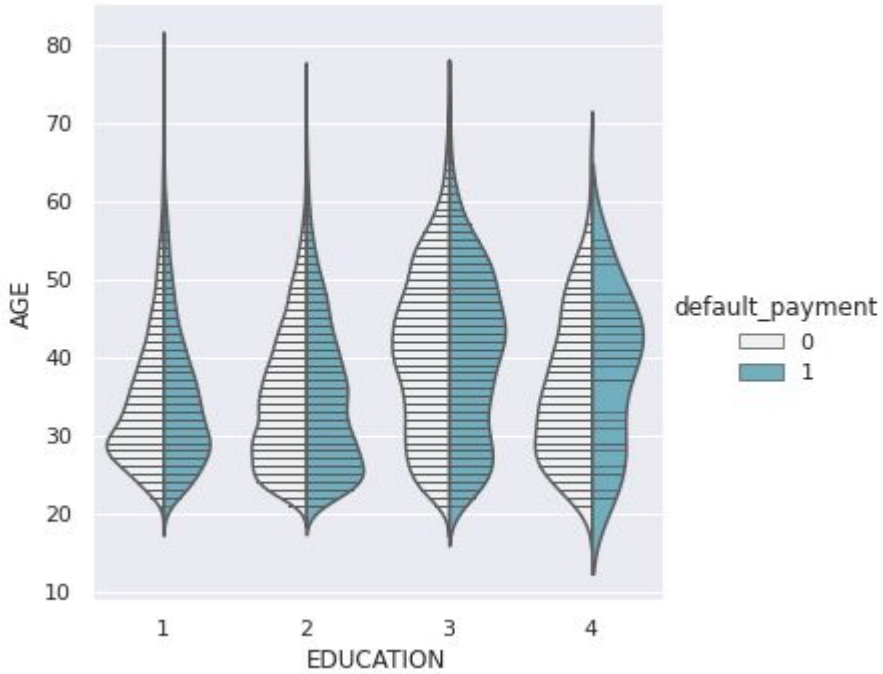



Both married and unmarried customers are equally defaulters, so clearly marriage status does not affect.

# Exploratory Data Analysis

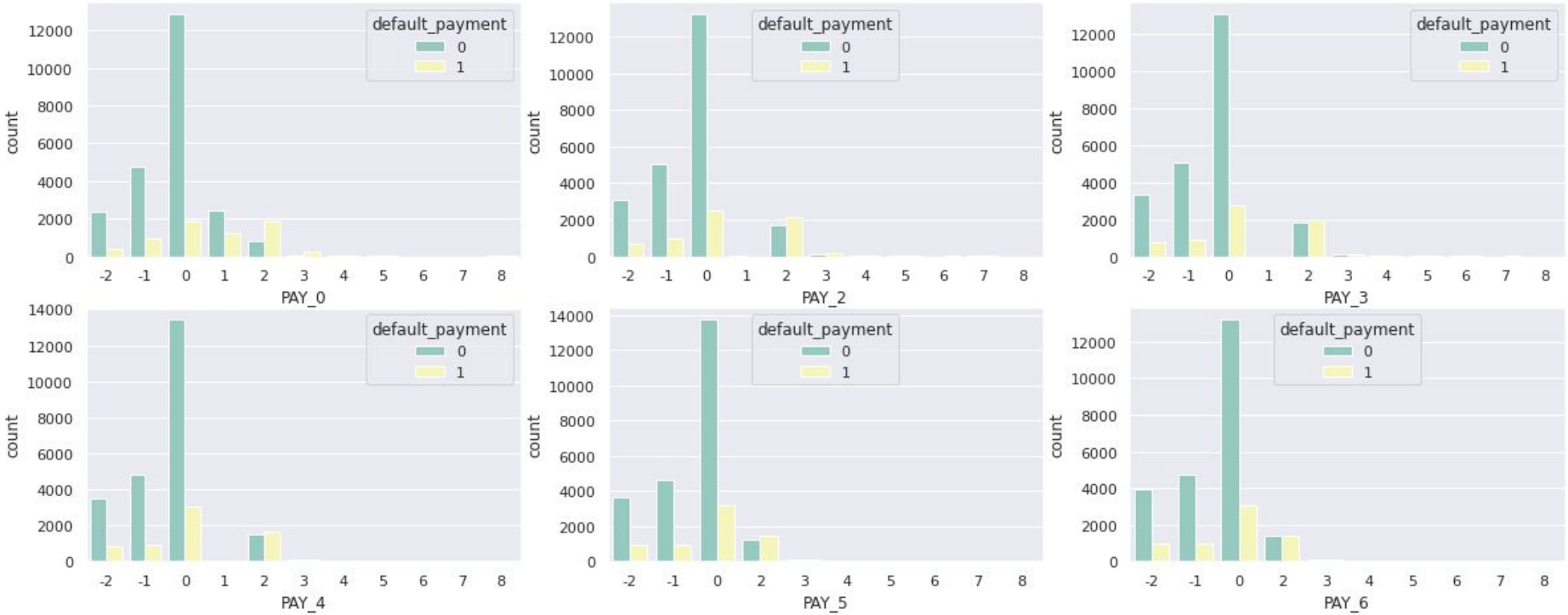



Married people with 40 years of age and unmarried people with 30 years of age and others above 40 are defaulters




All groups in have defaulters from the age 25 to 40 years.

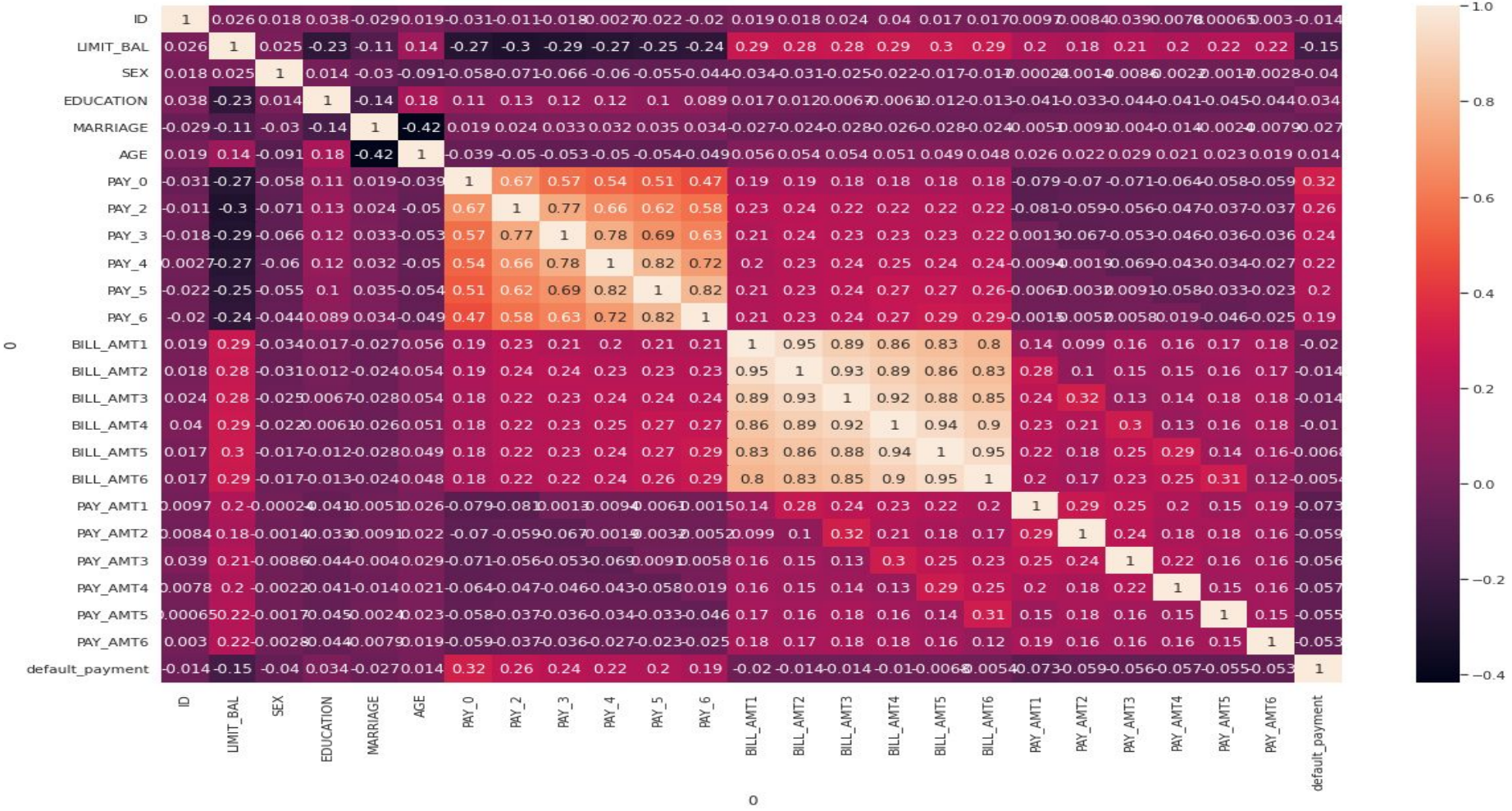
# Exploratory Data Analysis




Non Defaulters have a much higher proportion of zero or negative PAY_X variables (this means that being current or ahead of payments is associated with not defaulting in the following month)

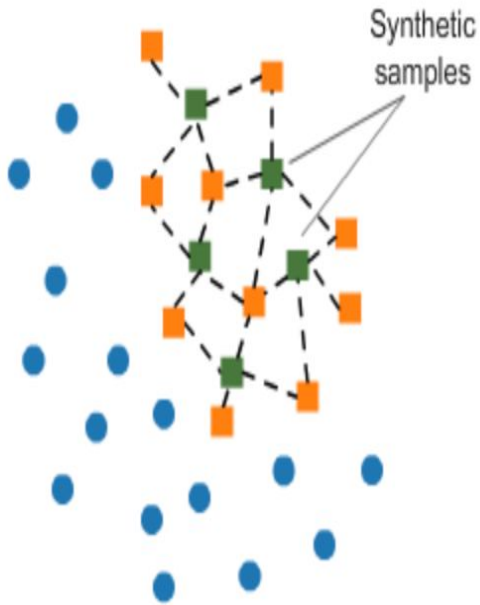
# Exploratory Data Analysis- Pair plot

# Exploratory Data Analysis- Heatmap



# Data Preprocessing and Preparation

- One hot encoding has been used for columns sex, education and marriage
- Since there was imbalance detected in the data **SMOTE** has been used to solve this. **SMOTE (Synthetic Minority Oversampling Technique)** works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.
- Test_train_split function was used to split the data into 80% training and 20% testing data.
- MinMaxScaler was used to standardize the features



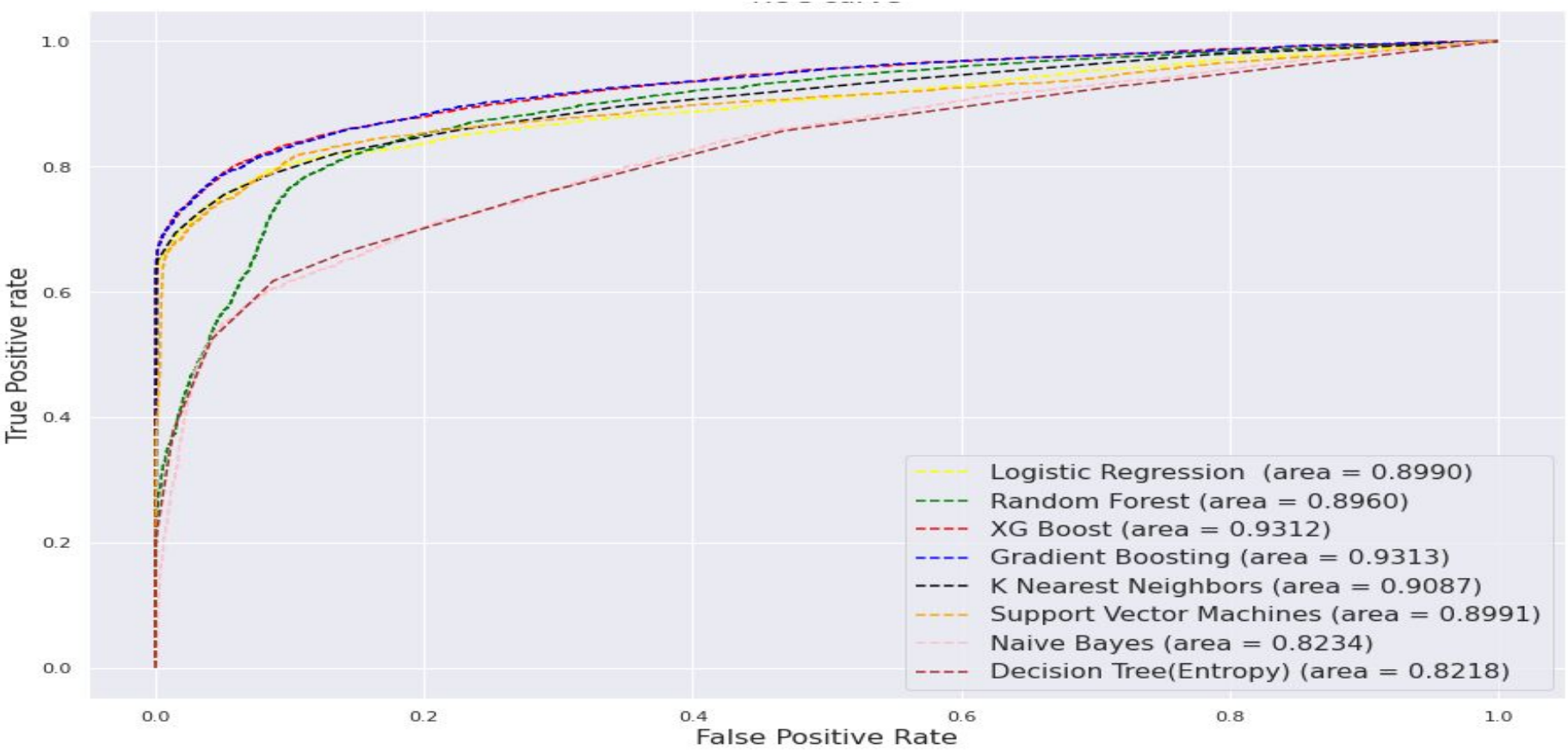
# Classification Models

The following models were used:

1. Logistic Regression
2. Random Forest Classifier
3. XGBoost Classifier
4. Gradient Boost Classifier
5. K Nearest Neighbors
6. Support Vector Machines
7. Naïve Bayes Classifier
8. Decision Tree Classifier

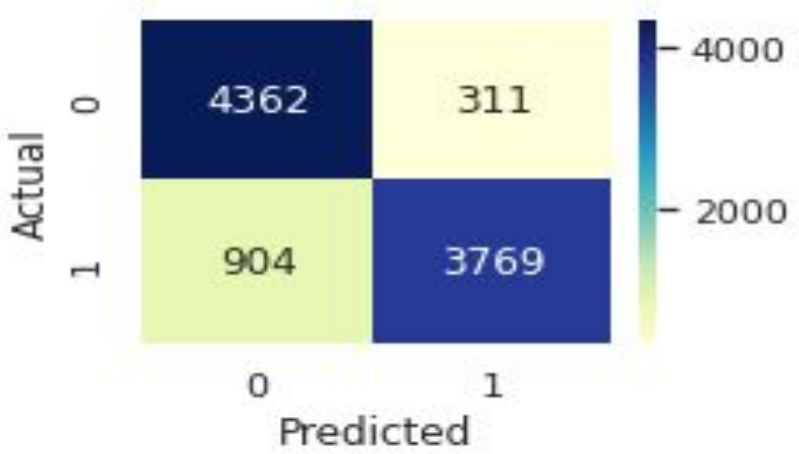
| Model | ROC_AUC Score |
|---|---|
| Logistic Regression | 0.850 |
| Random forest | 0.832 |
| XGBoost | 0.867 |
| Gradient Boost | 0.865 |
| KNN | 0.852 |
| SVM | 0.846 |
| Naive Bayes | 0.743 |
| Decision Trees | 0.759 |

# AUC Curve for all the models

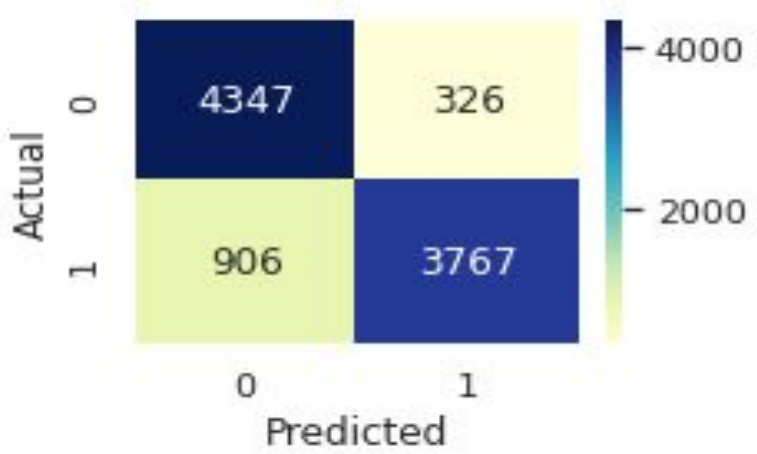


# Best Model Results

From ROC_AUC score it is clear that XGBoost and Gradient Boosting Classifiers have performed the best. Confusion Matrix for both models




In XGBoost, 4326 were class 0 and were predicted 0, 311 were class 0 but predicted 1, 904 were class 1 but predicted as 0, 3796 were class 1 and predicted as 1




In Gradient Boost, 4347 were class 0 and were predicted 0, 326 were class 0 but predicted 1, 906 were class 1 but predicted as 0, 3767 were class 1 and predicted as 1

# Conclusion

1. There are around 77% non-defaulters and 22% of defaulters. When the limit balance is low, it is likely that there will be more defaulters.

2. There are a greater number of female credit card holders and so there are majority female defaulters.

3. The majority of defaulters are university graduates and most then are in the age group 30-40 years.

4. Both Married and unmarried customers are equal defaulters although unmarried defaulters are slightly higher.

5.Multiple classification models were built and hyperparameter tuning was performed. All models performed well with ROC_AUC score above 74%.

6. XGBoost is the best of all the models with score of 86.7%.

# THANK YOU