Capstone Project-3

Credit Card Default Prediction

Individual Member

LOVA KUMAR POLUPARTI

content

- Introduction
- Problem Statement
- Data Summary
- Exploratory Data Analysis
- Model overview
- Feature Importances
- Challenges
- Conclusion

Introduction

In today's world credit cards have become a lifeline to a lot of people so banks provide us with credit cards.

A Credit Card is a type of payment card in which charges are made against a line of credit instead of the account holder's cash deposit. When someone uses a credit card to make purchase, that person's account accrues a balance that must be paid off each month.

Now we know the most common issue there is in providing these kind of deals are people not being able to pay the bills. These people are what we call "defaulters".

Problem Statement

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

Points for Discussion

Data visualization of defaulters

Data visualization of credit balance

Data visualization of Gender analysis

Data visualization of Education analysis

Data visualization of Marital analysis

Data visualization of Age distribution analysis

Data visualization of Bill amount analysis

Data visualization payment distribution analysis

continue……

Data visualization pair plot payment distribution analysis

Data visualization payment distribution analysis

Data visualization pair plot payment distribution analysis

Data visualization Gender vs Defaulters

Data visualization Education vs Defaulters

Data visualization Marital wise Defaulters

Applying SMOTE technique for Unbalanced data set

Applying Logistic Regression and Random forest classifier and KNN classifier and XG boost classifier

Data Summary

X1 - Amount of credit(includes individual as well as family credit)
X2 - Gender
X3 - Education
X4 - Marital Status
X5 - Age
X6 to X11 - History of past payments from April to September
X12 to X17 - Amount of bill statement from April to September
X18 to X23 - Amount of previous payment from April to September

Data Analysis Steps

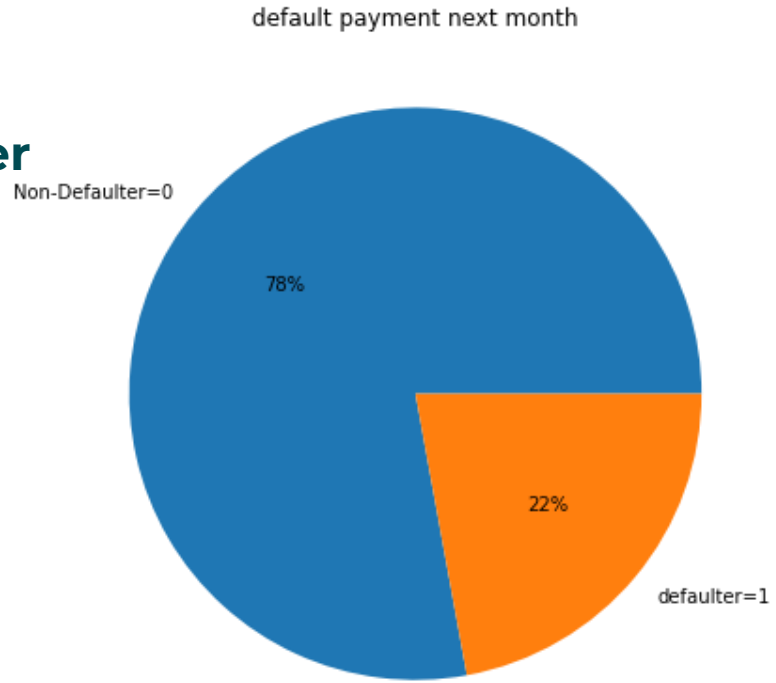
- **Import Libraries**: In this part, we had imported require libraries like numpy,pandas, matplotlib,seaborn, to perform Exploratory Data Analysis and For prediction we import sklearn library for credit card fraud detection
- **Descriptive Statistics**: In this part, we start by looking at descriptive statistic parameters for the dataset. We will use describe() this tell about unique values and standard deviation,mean, median, mode etc
- **Missing Value Imputation**: We will now check for missing values in our dataset. NOT EXISTED MISSING VALUES, In case there are any missing entries, we will impute them with appropriate values
- **Graphical Representation**: We will start with Univariate Analysis. and end with bivariate analysis during this i draw pie chart, bar chart, count plot etc

Exploratory Data Analysis

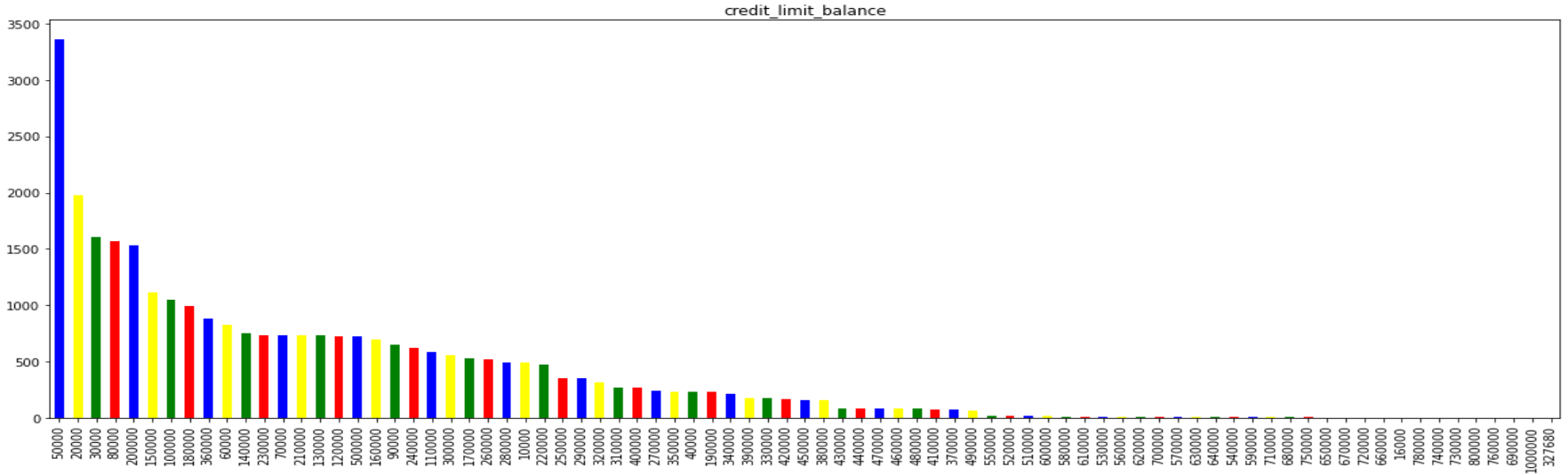
Defaulters:

- Non-defaulter were 23364, Defaulter
- were 6636

- The above pie charts said
- Non-defaulters 78% ,and defaulter
- were 22%



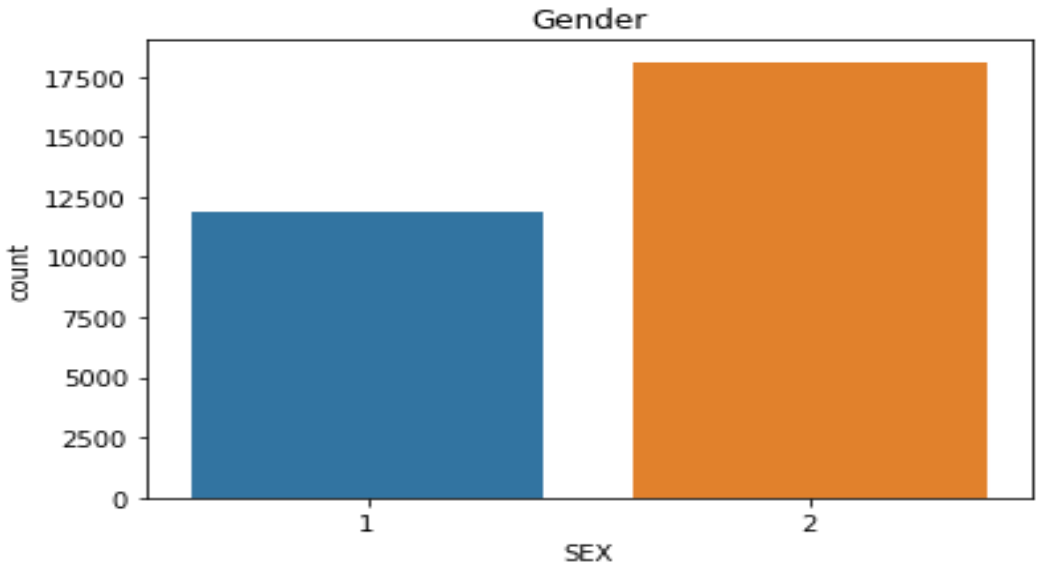
Credit Limit Balance Analysis



- Most of people available credit balance is 50,000 to 3365 members.
- Maximum available credit balance limit is 1,00,000
- Minimum available credit balance limit is 10,000

Gender Analysis

- Male are 11,888
- represented
- as 1 and female are 18,112
- represented as 2



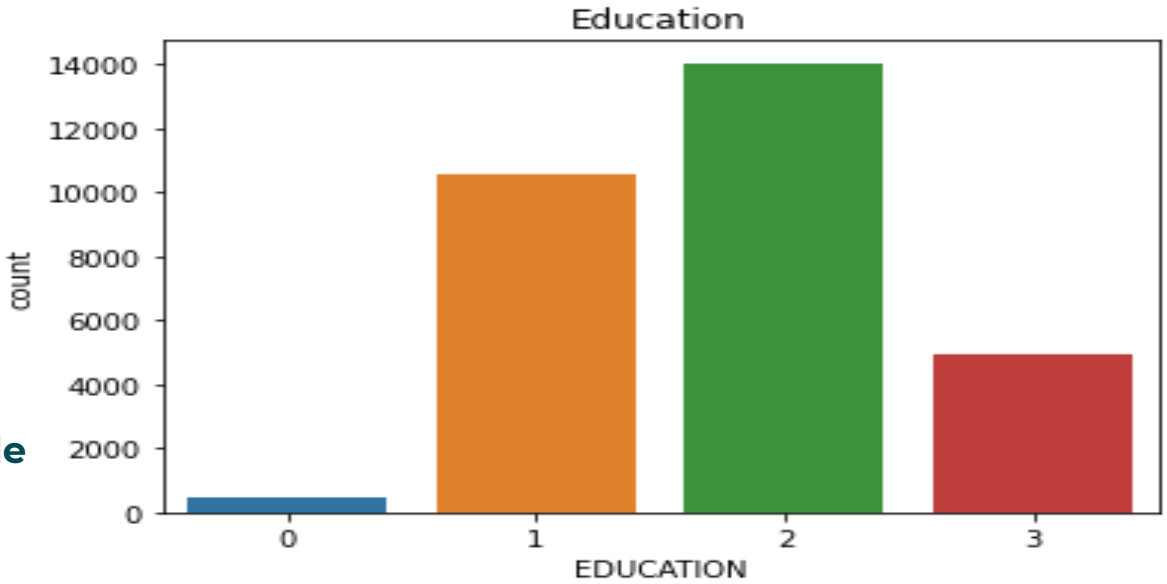
Education Analysis

Graduate school by 10,585 people

University by 14,030 people

High school by 4917 people

Other studying by 468 people

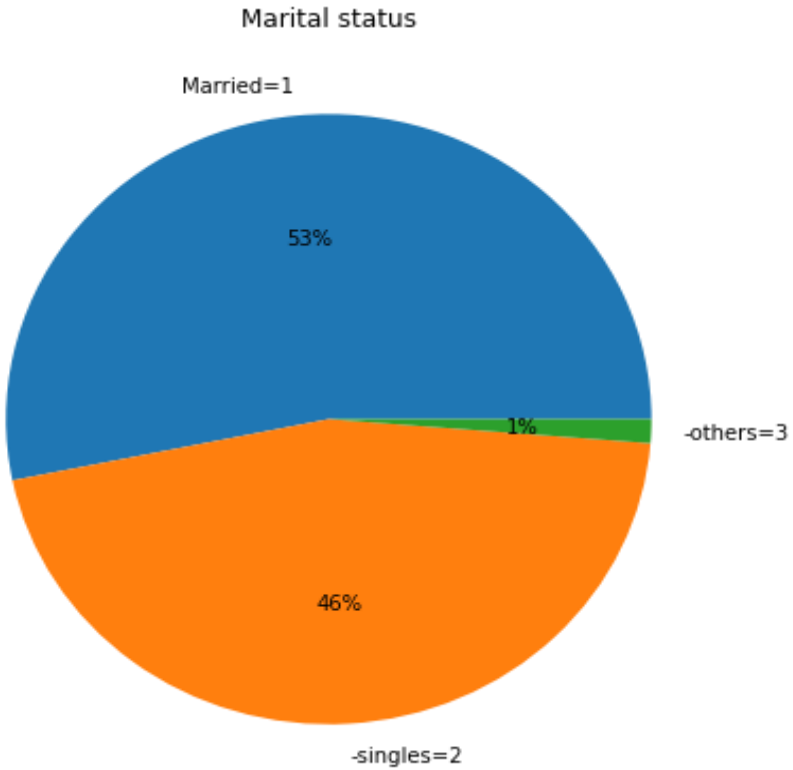


Marital Status Analysis

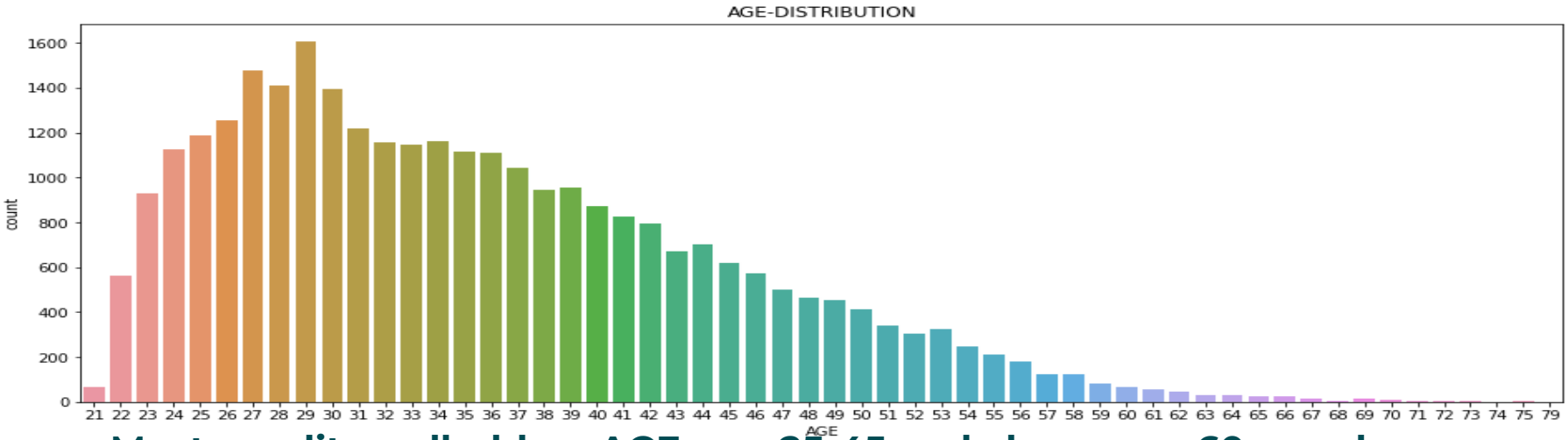
Married persons are 13,659 represented as 1

Single persons are 15,964 represented as 2

Others are 377 represented as 3

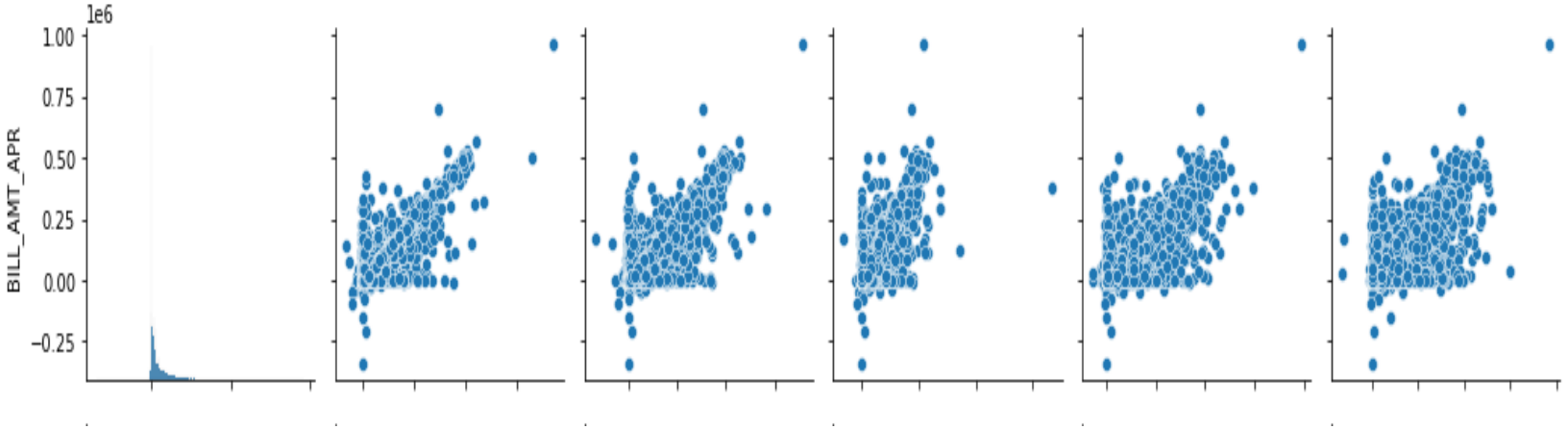


Age Distribution



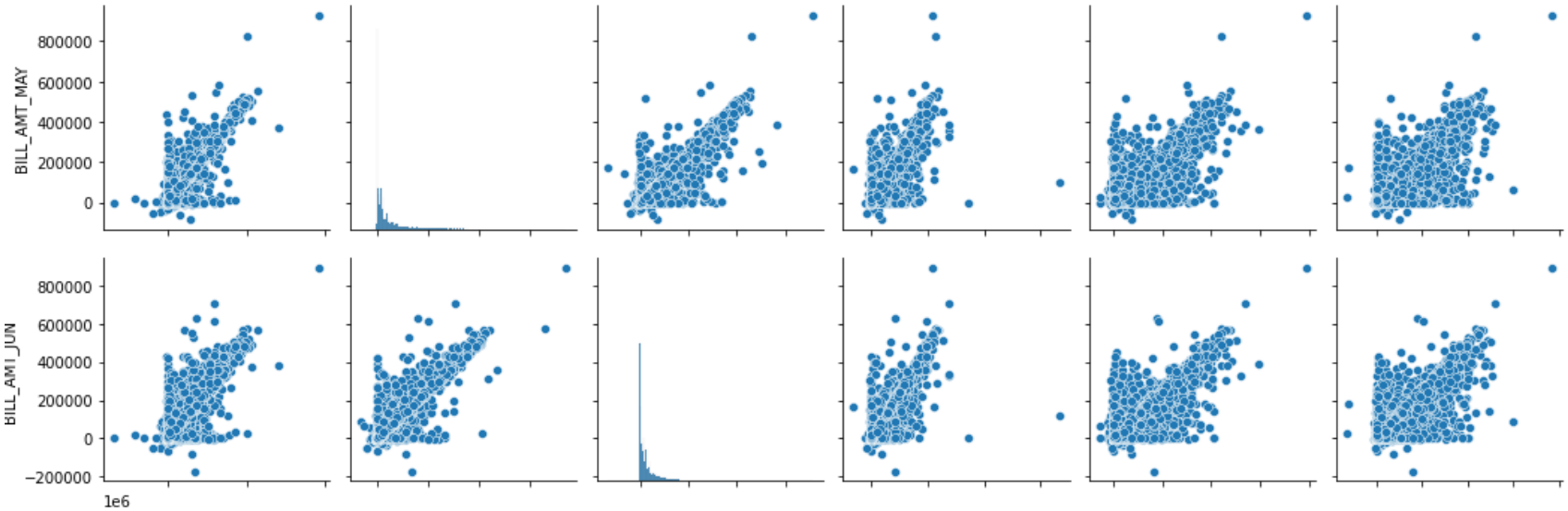
- Most credit cardholders AGE was 25-45 and above age 60 people are used rarely credit cards
- Most of 29th age people used huge credit cards that number is 1605 and second place was 27th age people it's number 1477

Bill Amount Distribution



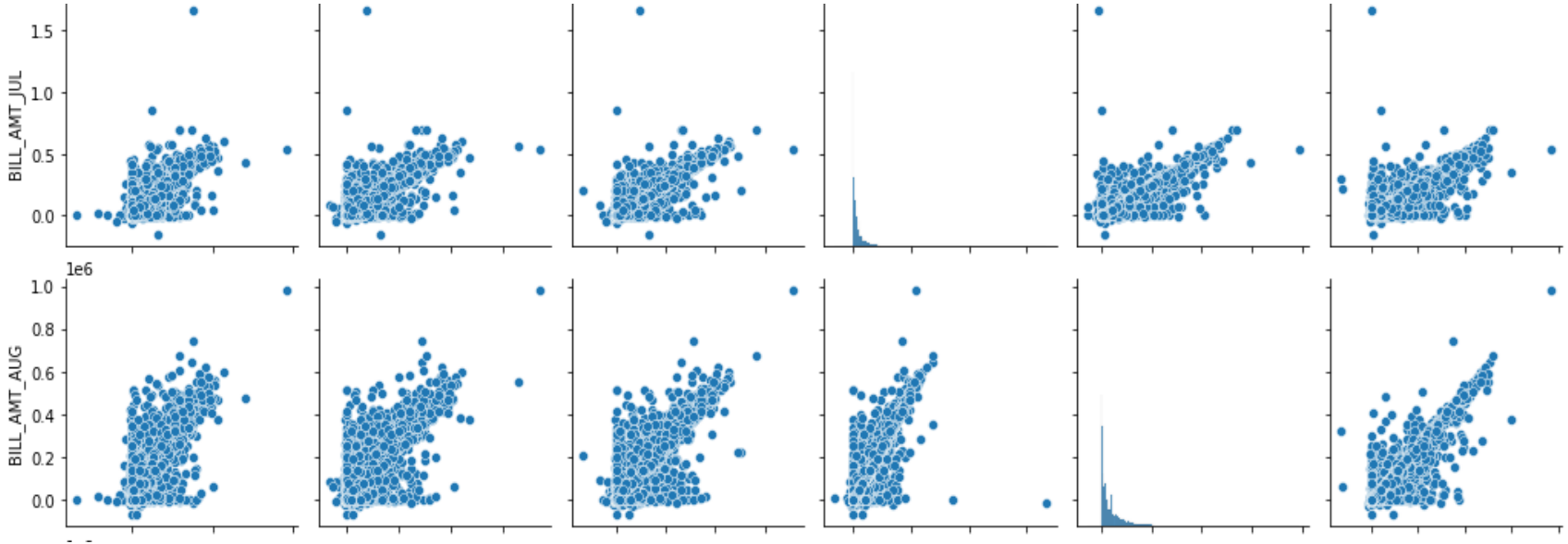
This pairplot shows that distribution of bill amount statements for April ,it states for defaulters and non-defaulters

Bill Amount Distribution



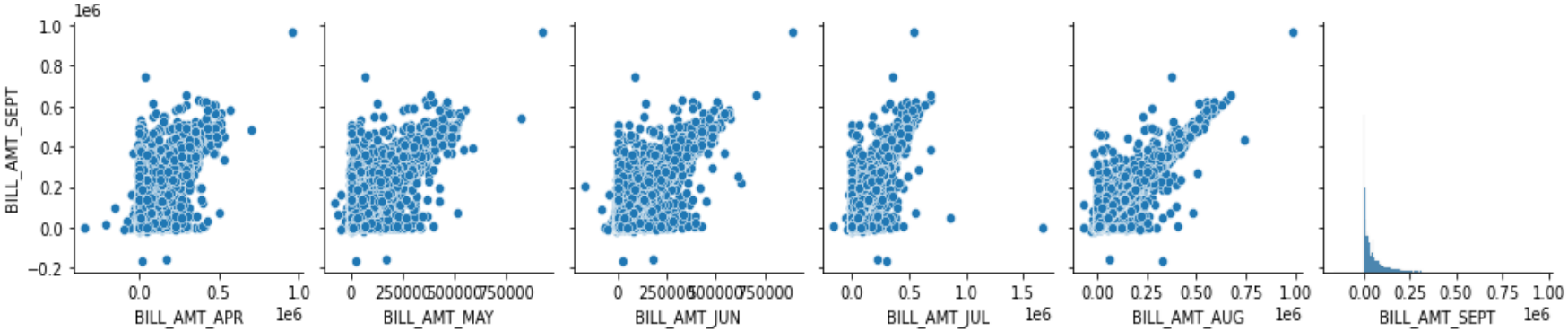
This pairplot shows that distribution of bill amount statements for May and June month it states for defaulters and non-defaulters.

Bill Amount Distribution



This pairplot shows that distribution of bill amount statements for July and August month ,it states for defaulters and non-defaulters.

Bill Amount Distribution

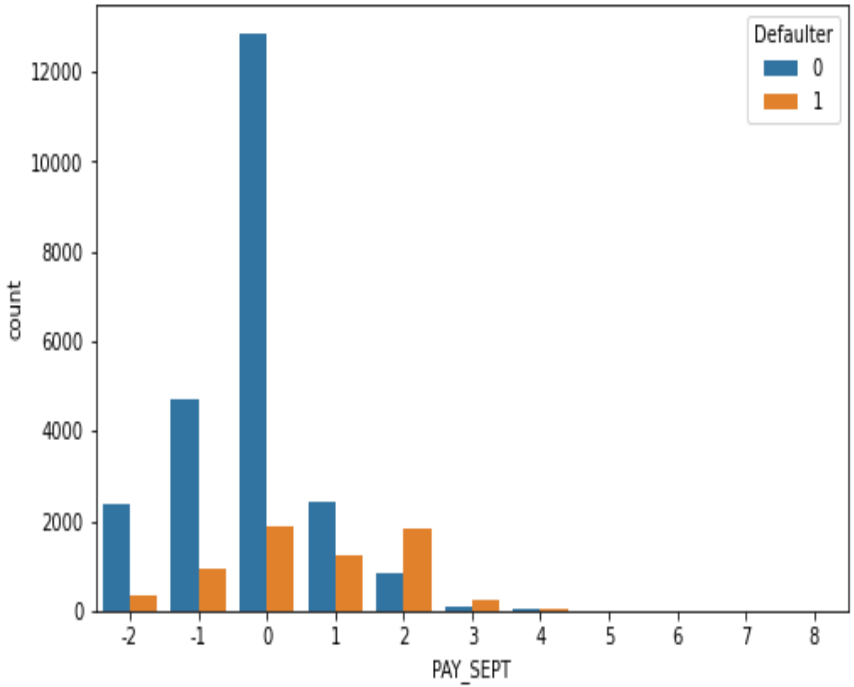


- This pairplot shows that distribution of bill amount statements for each month it states for defaulters and non-defaulters.

Previous Payment Distribution
(September Month)

- In the payment of september month
- single rupee not receive from
- Non-defaulters were
- 2394 persons, and from defaulters were
- 365 persons

- Full amount received from
- Non-defaulters
- were 4732 persons and defaulters
- were 954 persons
- Minimum amount received from Non-
- defaulters were 12,849 persons and Defaulters
- Were 1888 persons.

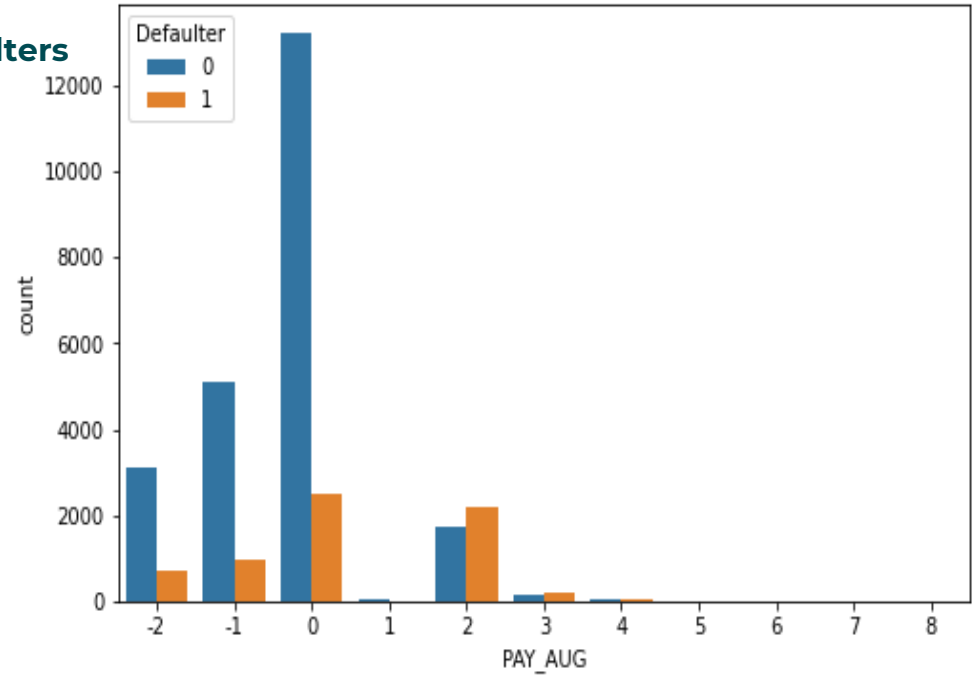


August Payment Distribution

- In the payment of august month
- single rupee not receive from Non-defaulters
- were 3,091 persons, and from
- defaulters were 691 persons

- Full amount received from
- Non-defaulters were 5,084 persons and
- defaulters were 966 persons

- Minimum amount received rom
- Non- defaulters were 13,227 persons and
- Defaulters were 2,503 persons

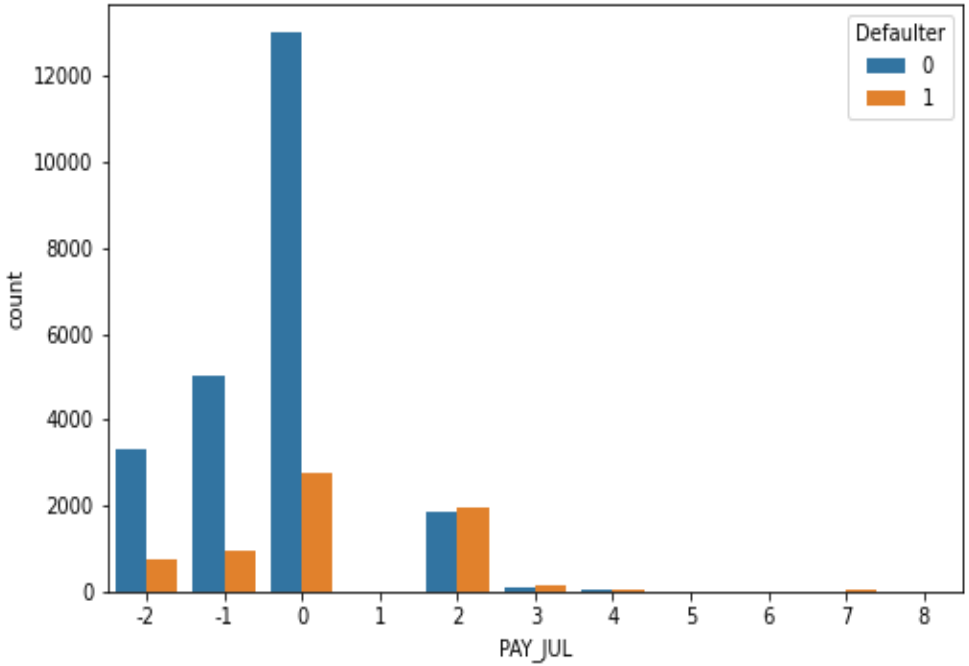


July Payment Distribution

In the payment of July month
single rupee not receive from
Non defaulters were 3328 persons,
and from defaulters were 757 persons

Full amount received from
Non-defaulters were 5012 persons and
defaulters were 926 persons

Minimum amount received from
Non- defaulters were 13,013 persons and
Defaulters were 2,751 persons

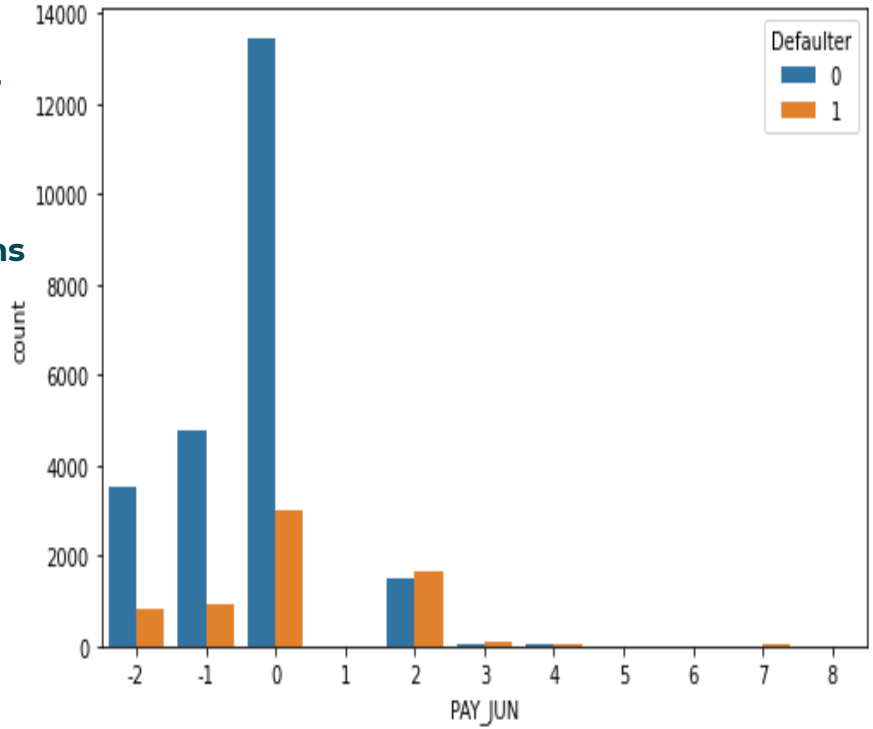


June Payment Distribution

- In the payment of June month single rupee
- not receive from Non-defaulters were 3,511 persons,
- and from defaulters were 837 persons

- Full amount received from Non-defaulters
- were 4,783 persons and defaulters were 904 persons

- Minimum amount received from
- Non- defaulters were 13,439 persons
- and Defaulters were 3,016 persons

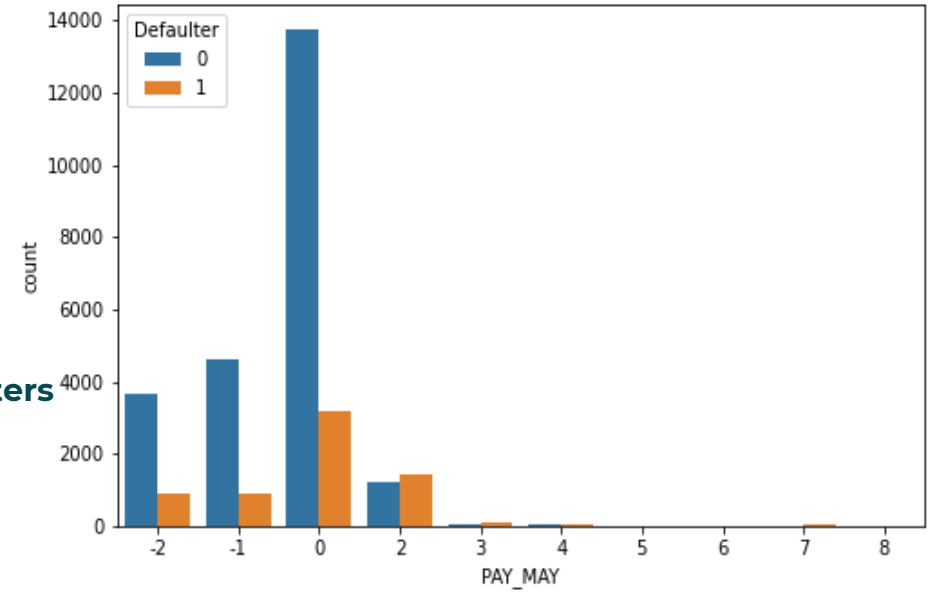


May Payment Distribution

- In the payment of May month
- single rupee not receivefrom
- Non-defaulters were 3,651 persons, and
- from defaulters were 895 persons

- Full amount received from Non-defaulters
- were 4,642 persons and
- defaulters were 897 persons

- Minimum amount received from Non- defaulters
- were 13,752 persons and
- Defaulters were 3,195 persons

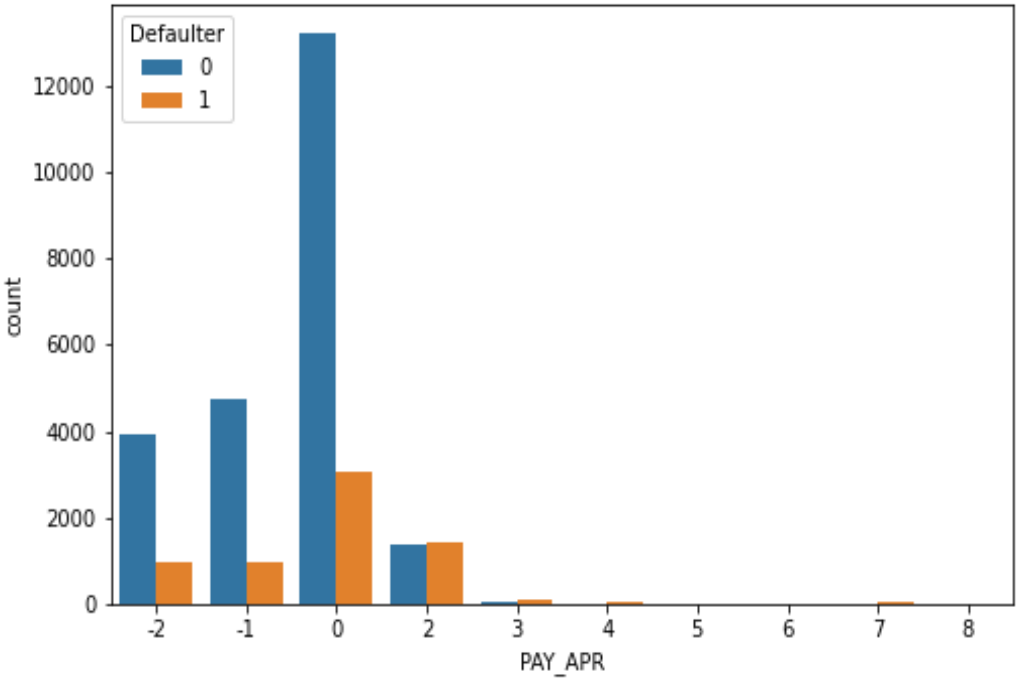


April Payment Distribution

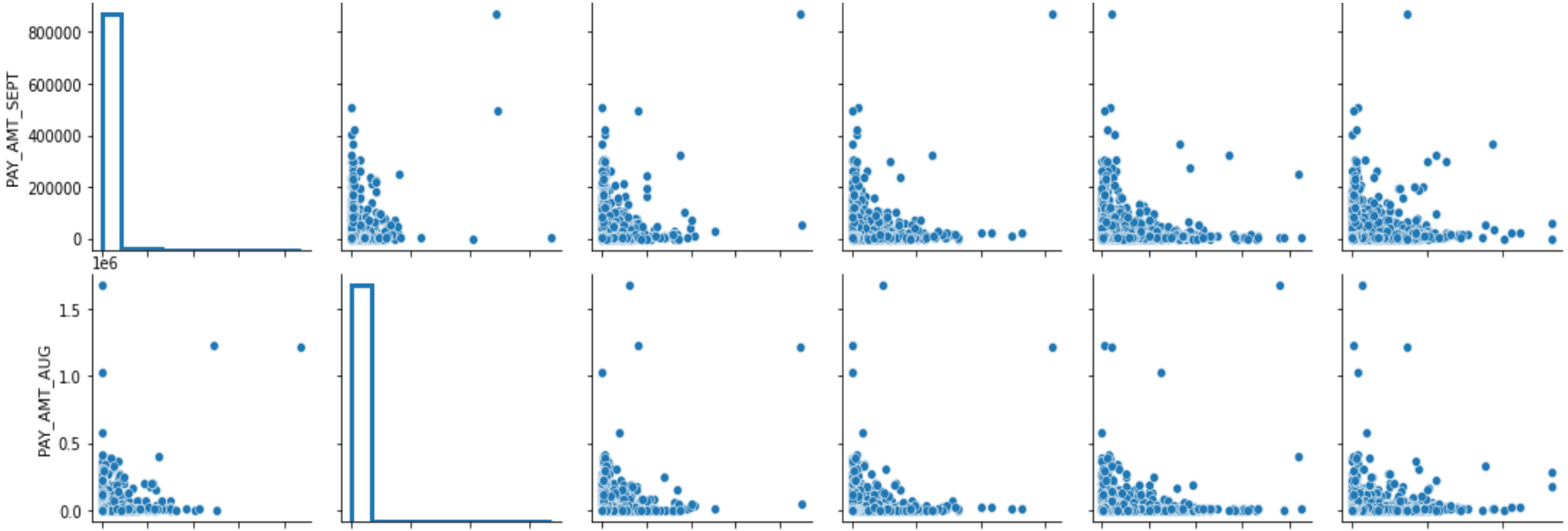
In the payment of April month single rupee not receive from Non-defaulters were 3,914 persons, and from defaulters were 981 persons

Full amount received from Non-defaulters were 4,765 persons and defaulters were 975 persons

Minimum amount received from Non- defaulters were 13,217 persons and Defaulters were 3,069 persons

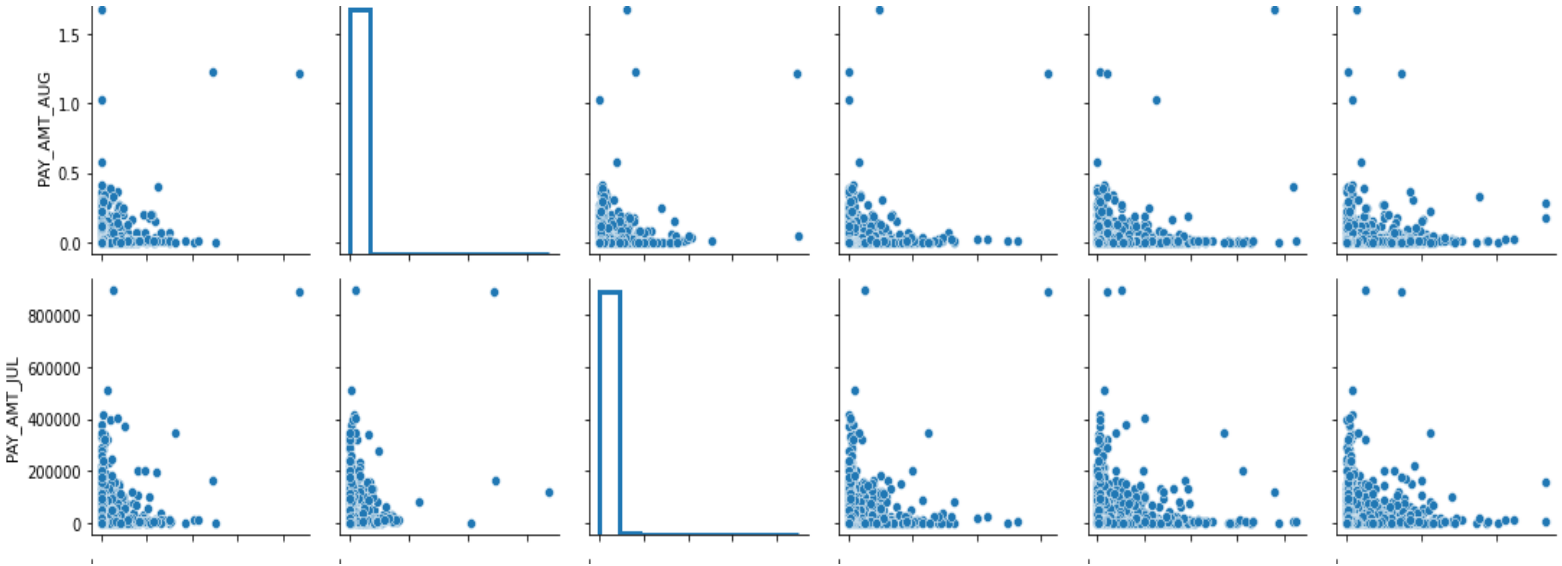


Distribution Of Previous Payment



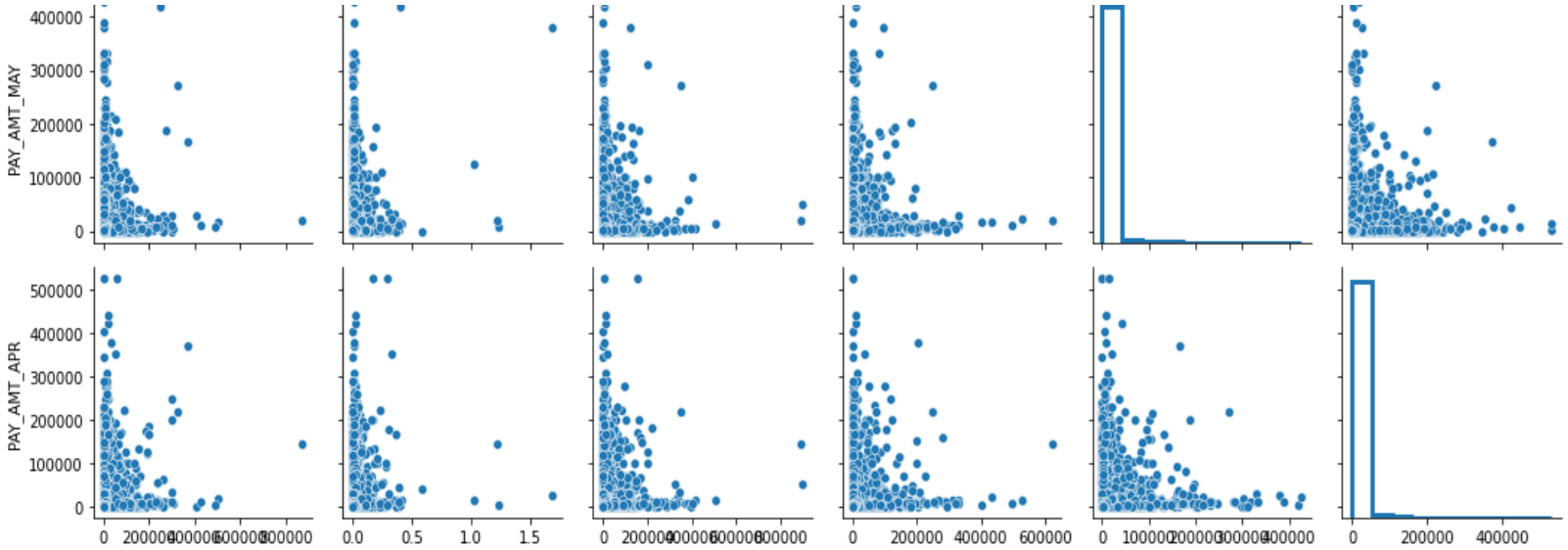
- **Pairplot said that Sep and Aug month of the payment ,it states Defaulters and Non-defaulters.**

Distribution Of Previous Payment



- Pairplot said that Aug and July month of the payment ,it states Defaulters and Non-defaulters.

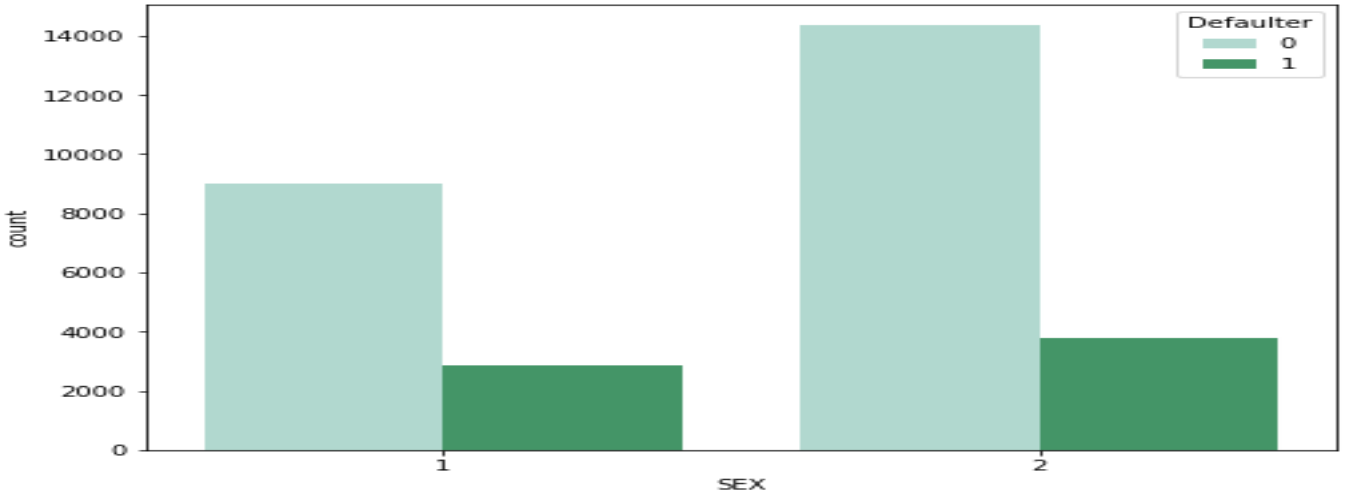
Distribution Of Previous Payment



- Pairplot said that May and Apr month of the payment ,it states Defaulters and Non-defaulters.

Bivariate Analysis

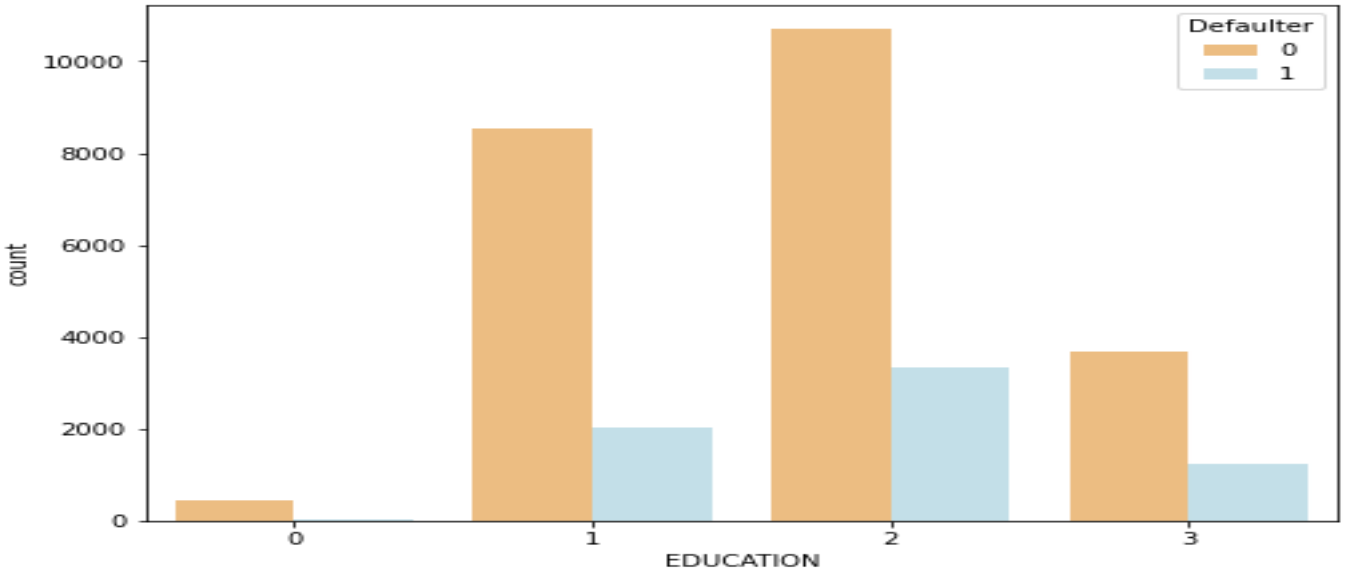
Gender wise Defaulters



Highest Defaulters were females, their number was 3763 and

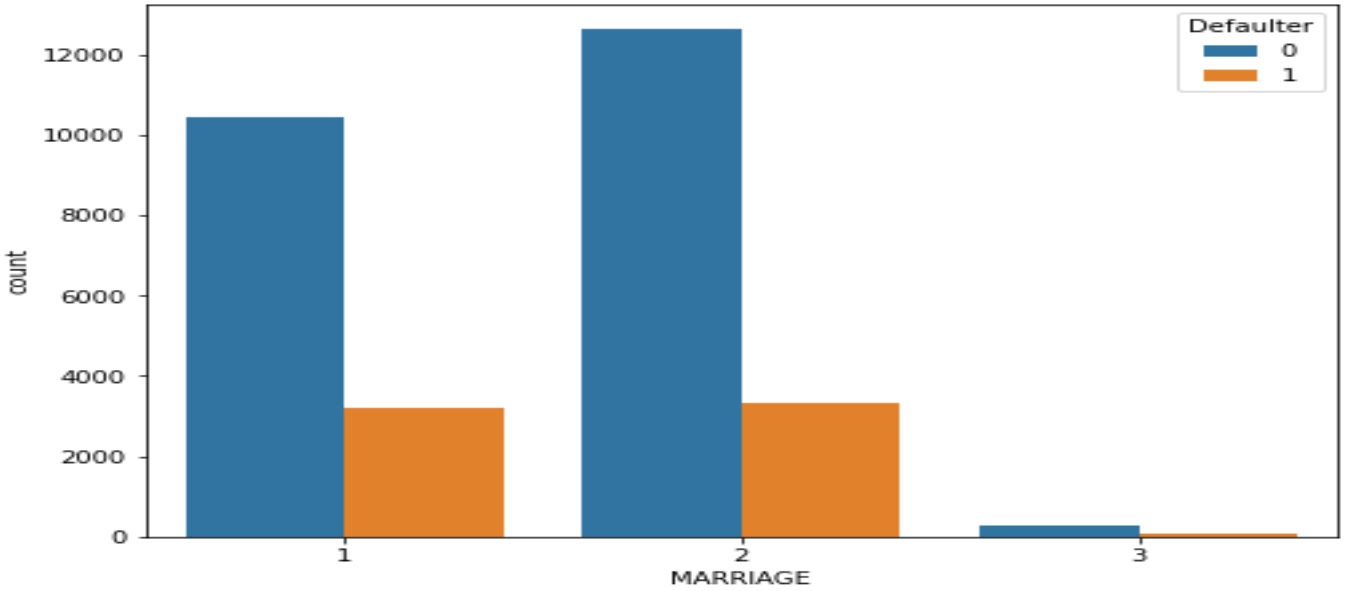
Highest Non- defaulters also famales it's number 14,349

Education Wise Defaulters



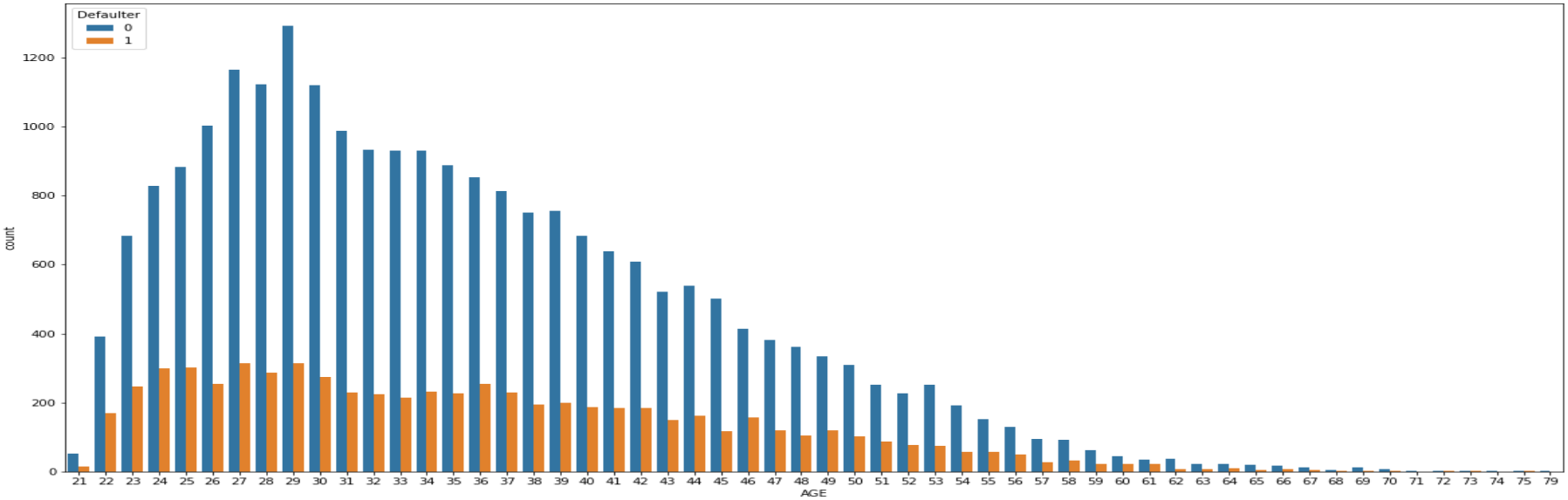
- Defaulters who did study in University people its number was 3,330 and after that graduate school people are defaulter their number was 2036.
- 1 represented as graduate school,2 represented as university,3 represented as high school,0 represented as others.

Marital Wise Defaulters



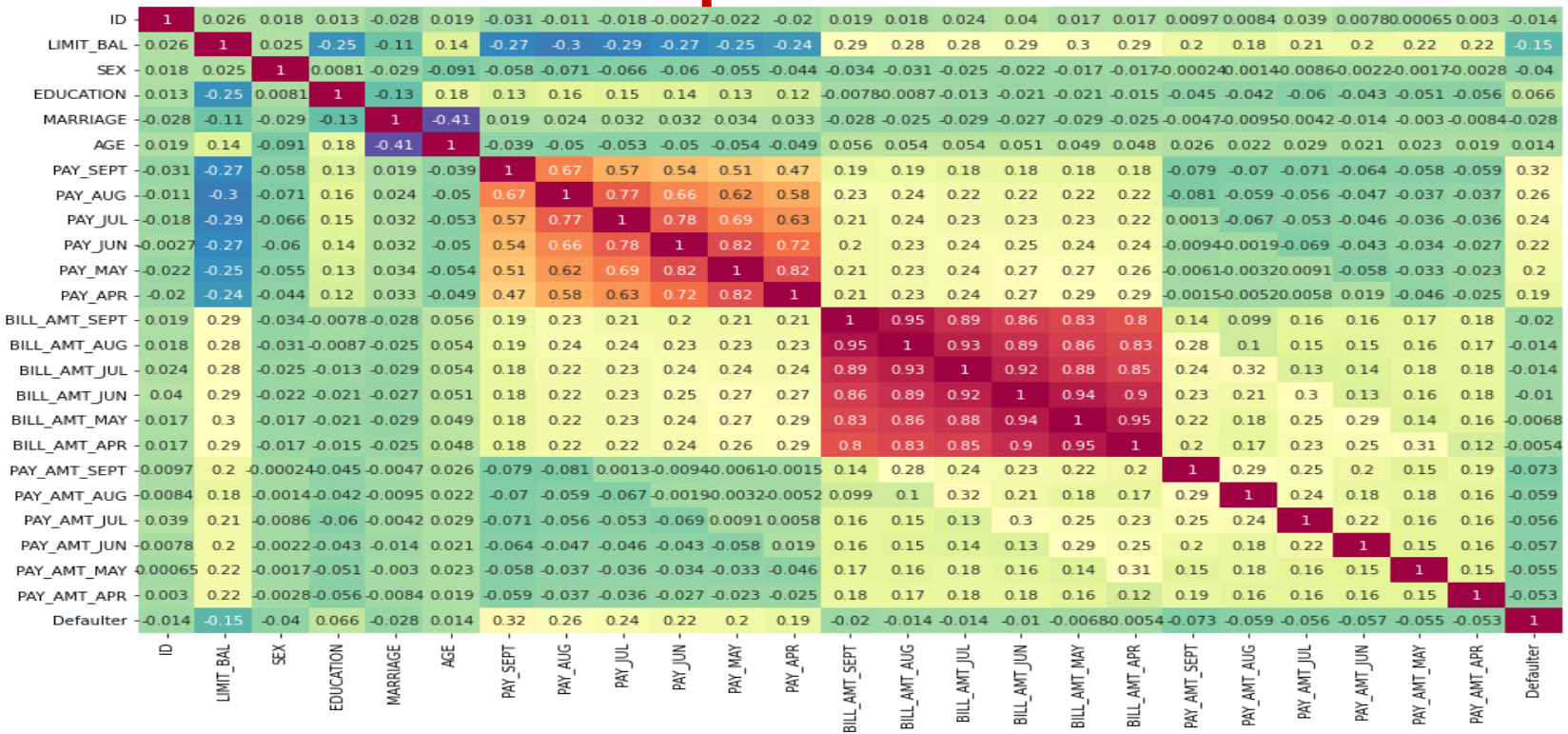
- Most of the defaulters were singles their numbers was 3341 and defaulters in married people were 3206 and defaulters in others were 89.

Age Wise Defaulters



- **Most of 27th age people defaulters least defaulters were above 60age people were defaulters.**

Heatmap Correlation



In this heatmap we can conclude that correlation relationship between any two columns,some part of data having highly correlated,some part of data having negative correlated and some data have zero correlated.*In heatmap, Highly correlated items "PAY_SEPT","BILL_AMT_SEPT","PAY_AMT_SEP Removed*

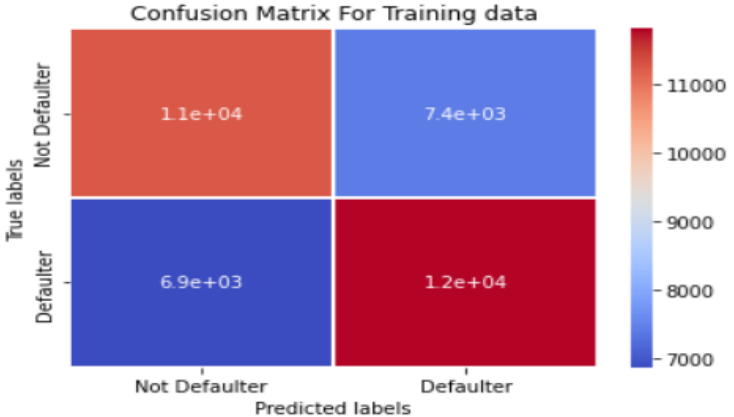
Applying SMOTE(Synthetic Minority Oversampling Technique)

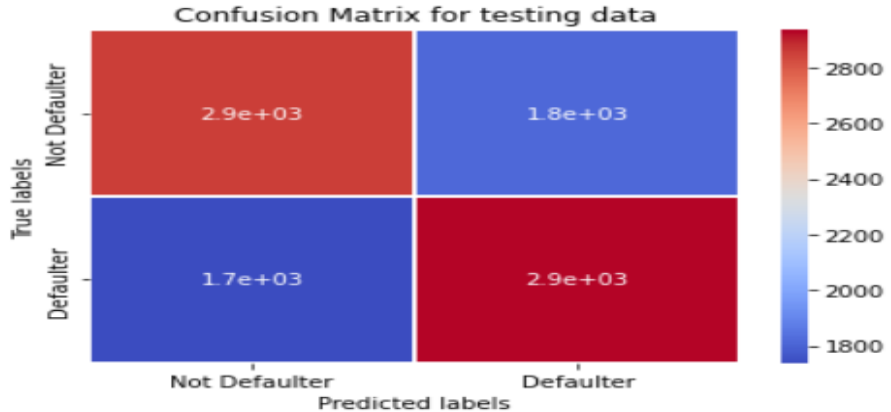
After dataset is imbalanced dataset so we need to do the balance using SMOTE(Synthetic Minority Oversampling Technique)

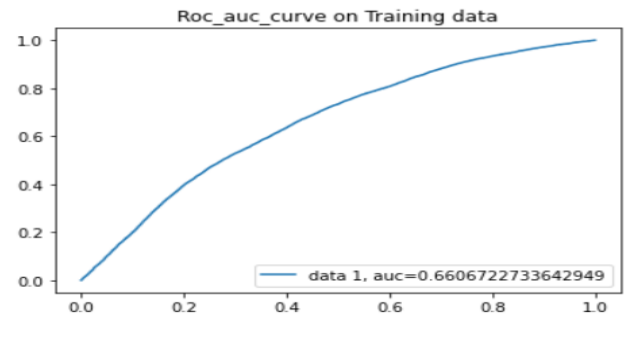
We got,

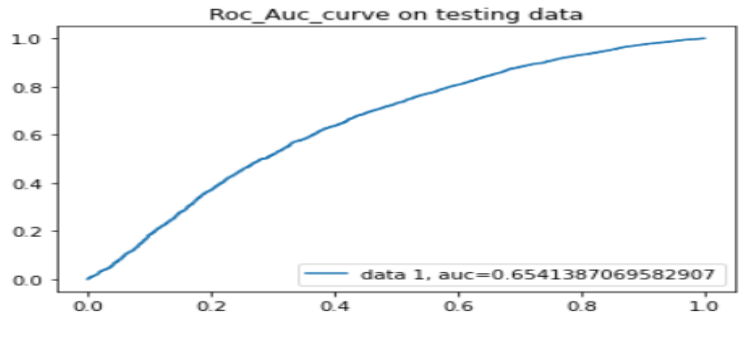
- **Original dataset shape Counter({0: 18691, 1: 5309})**
- **Resample dataset shape Counter({1: 23364, 0: 23364})**
- **Counter({0: 23364, 1: 23364})**

Logistic Regression



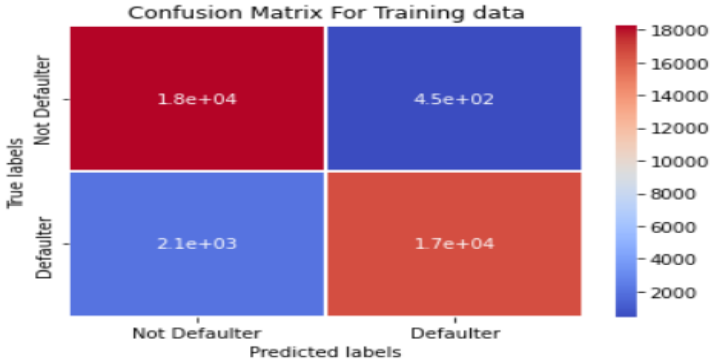


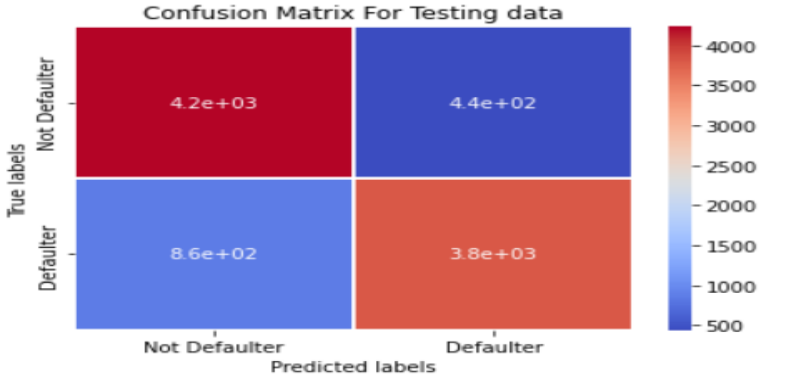


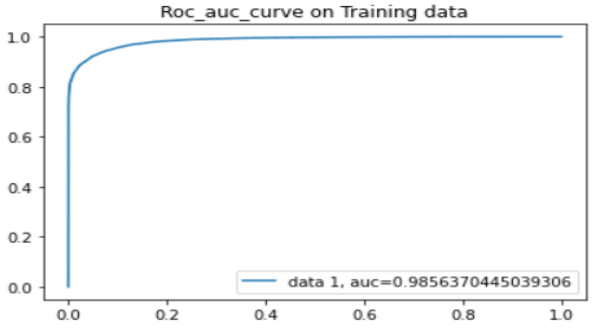


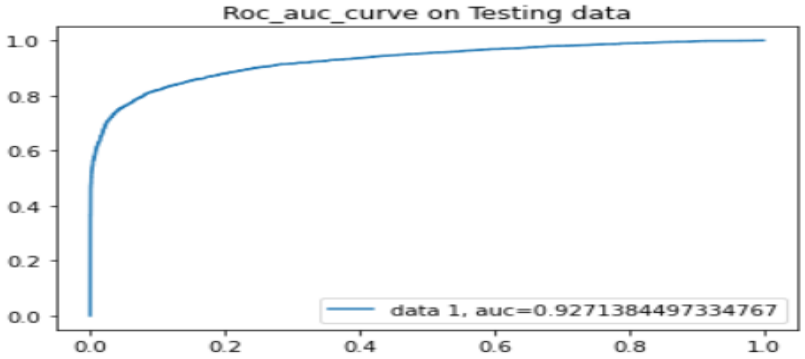
we have implemented logistic regression and we getting accuracy_score is approx 62%. and precision score approx is 62% and f1_score is 62%and roc_auc approx is 62% As we have imbalanced dataset, recall_score is approx 63% better parameter. Let's go ahead with other models and see if they can give better result

Random Forest Classifier



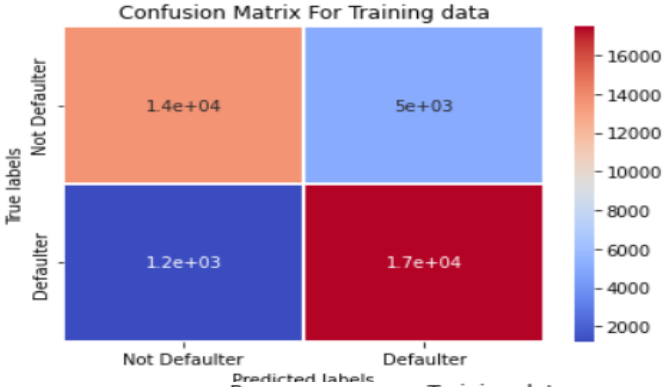


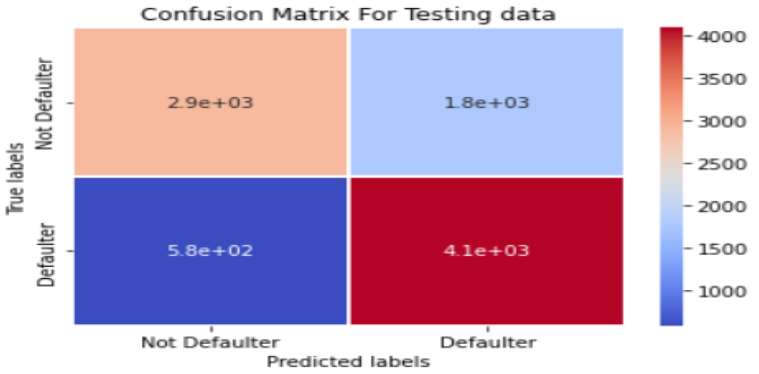


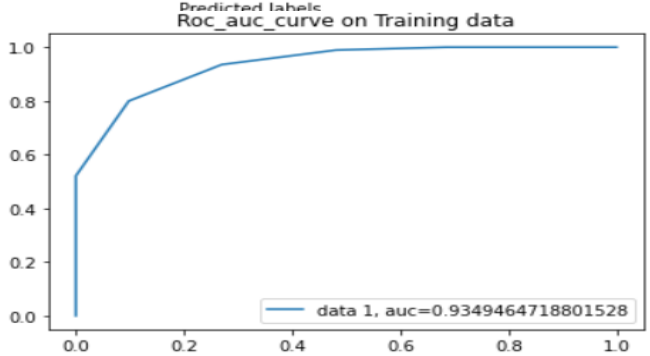


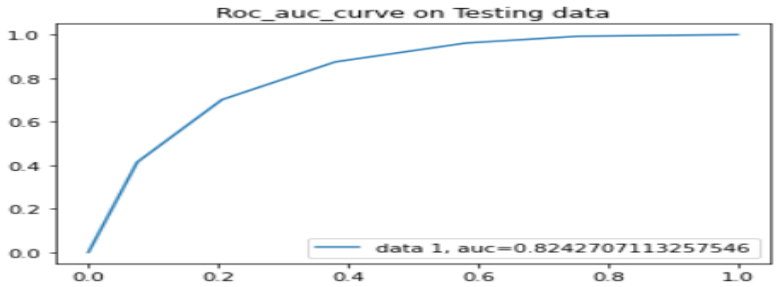
- **By implemented Random Forest and we getting accuracy_score is approx 86%. and recall_score is approx 82%,and f1_score is 85%,ROC_AUC score is 86%, precision score is 90% better parameter**

KNN Classifier



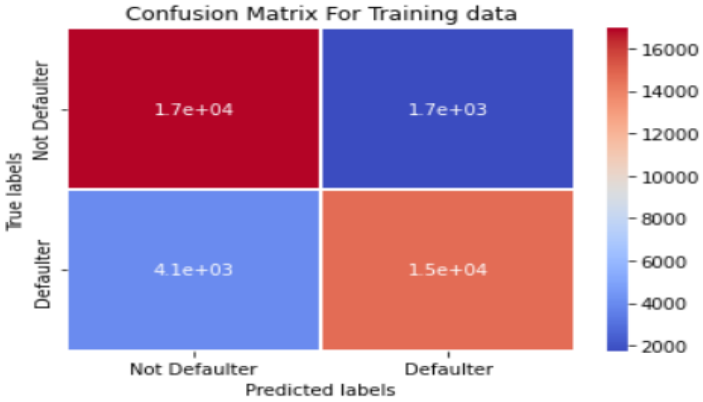


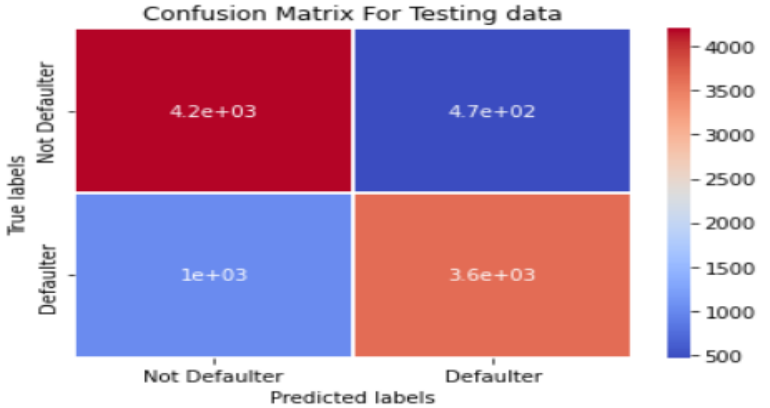


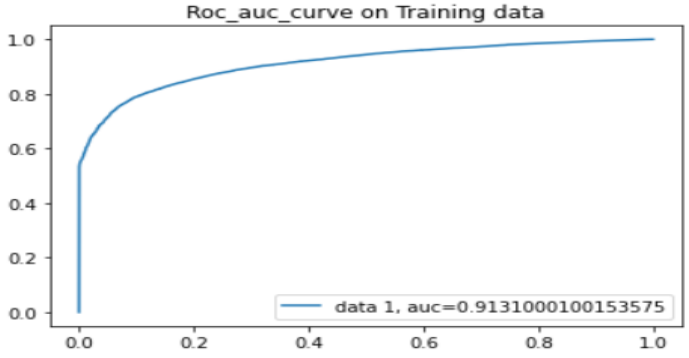


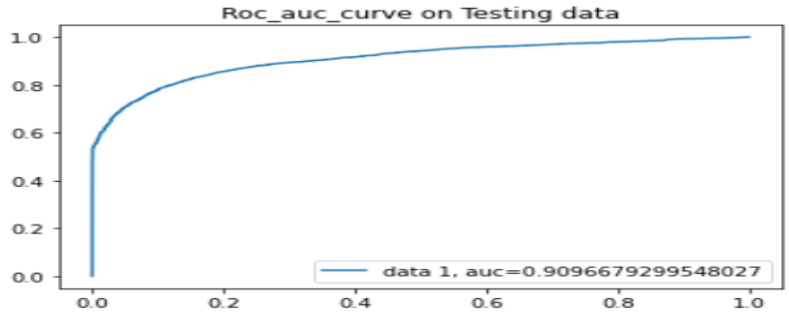
By implemented KNN and we getting accuracy_score is approx 75%. and precision score is approx 70% and f1_score is 76% and ROC_AUC score is 75% ,recall_score is approx 88% better parameter

XGBoost Classifier









By implemented XGBOOST WITH GRID SEARCH CV and we getting accuracy_score is approx 84%. and recall_score is approx 78% and f1_score is 83% and ROC_AUC score is 84% ,precision score is approx 90% better parameter

Summary For Train and Test Data

TRAIN SUMMARY

SL NO	MODEL_NAME	Train accuracy_score	Train Precision_score	Train Recall_score	Train f1_score	Train ROC_AUC_Score
1	Logistic Regression	0.6177572093520945	0.6144089822226843	0.6323899202824889	0.6232697935616546	0.6177572093520945
2	Random_Forest_classifier	0.9318121020812156	0.9737614463489082	0.8875394574929111	0.9286533993898173	0.9318121020812155
3	KNN_Classifier	0.8327804825852014	0.7761498845675724	0.9353164624685678	0.8483319179910227	0.8327804825852015
4	XGBoost_Classifier	0.844818361778396	0.8935637518319492	0.7828901610400728	0.8345738158382526	0.8448183617783961

TEST SUMMARY

SL NO	MODEL_NAME	Test accuracy_score	Test Precision_score	Test Recall_score	Test f1_score	Test ROC_AUC_Score
1	Logistic Regression	0.6196233682858977	0.6176346801346801	0.6280761823239889	0.6228116710875332	0.6196233682858977
2	Random_Forest_classifier	0.8607960624866253	0.896519285042333	0.815750053498823	0.8542296918767507	0.8607960624866253
3	KNN_Classifier	0.7472715600256794	0.6969490369865349	0.8750267494115129	0.7759013282732449	0.7472715600256794
4	XGBOOST_Classifier	0.8400385191525787	0.8864299610894941	0.7800128397175262	0.829823562891292	0.8400385191525788

Challenges

- Understanding the columns.
- Feature engineering.
- Getting a higher accuracy on the models.

Conclusion

- By used different type of Classification algorithms to train our model like, Logistic Regression, Random Forest Classifier,KNN_Classifier, XGboost_Classifier. alongwith tuned the parameters ,Out of them Random forest classifier with Grid search CV (tuned hyperparameters gave) the best result.

- Highest Precision score is approx 90%,
- ROC_Auc score is approx 86%,
- and Accuracy_score is approx 86%,
- and It's F1_score approx is 85%,
  Recall_score approx is 82%

Thank you