Capstone Project-3

Mobile Price Range Prediction-ML Classification

Team Members

Mohd Sakib Quraishi

AlmaBetter

Contents :

1.Problem statement

2. Data Exploration

3.Data Pre-processing

4.Data Cleaning

5 Exploratory Data Analysis

6.Feature Engineering

8.Standardization

9.Logistic regression model

10.LightGBM

11.Random Forest

12.Decision tree

13.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. We can use the K-S chart to evaluate which customers will default on their credit card payments

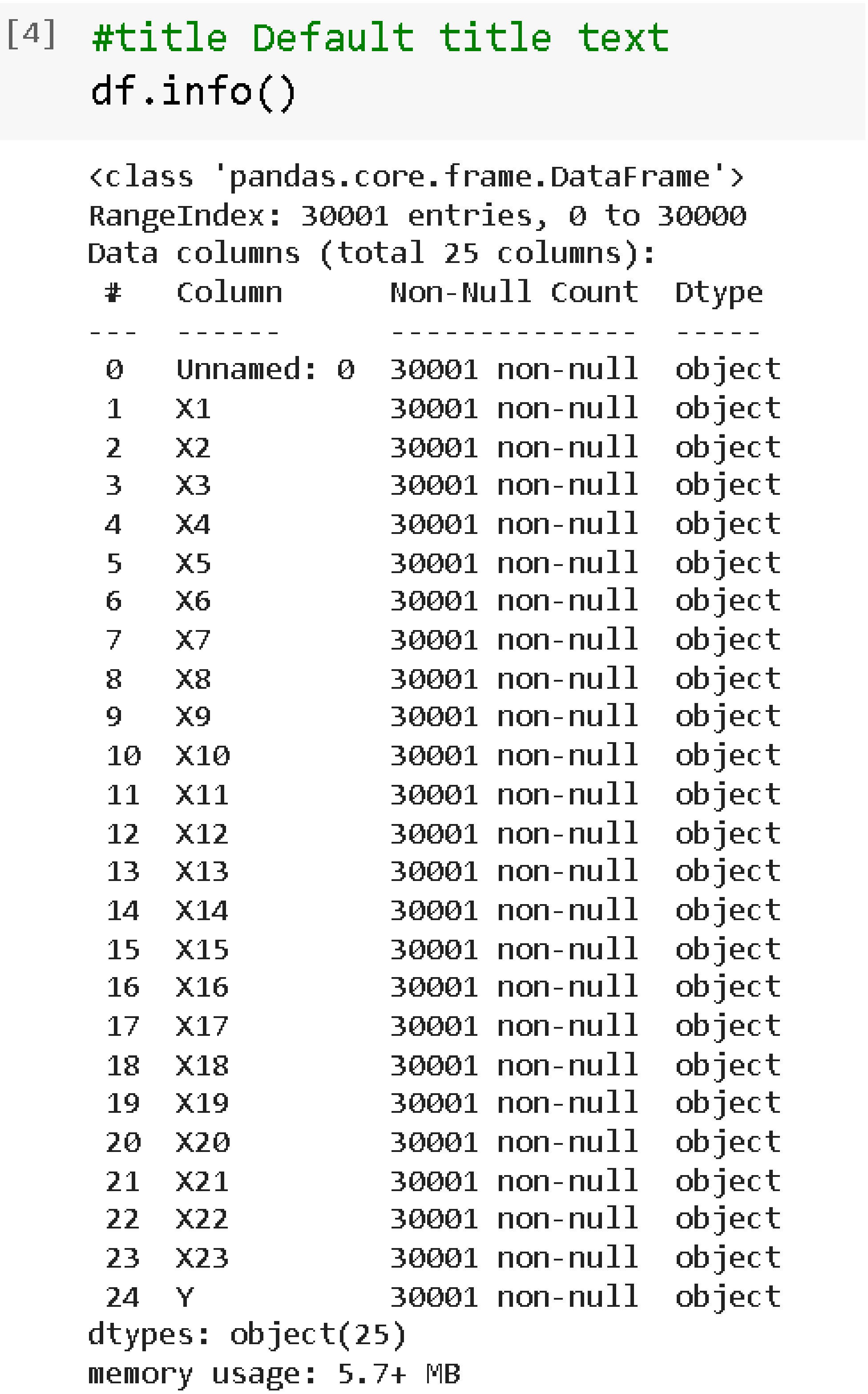
Attribute Information

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -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.

- X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

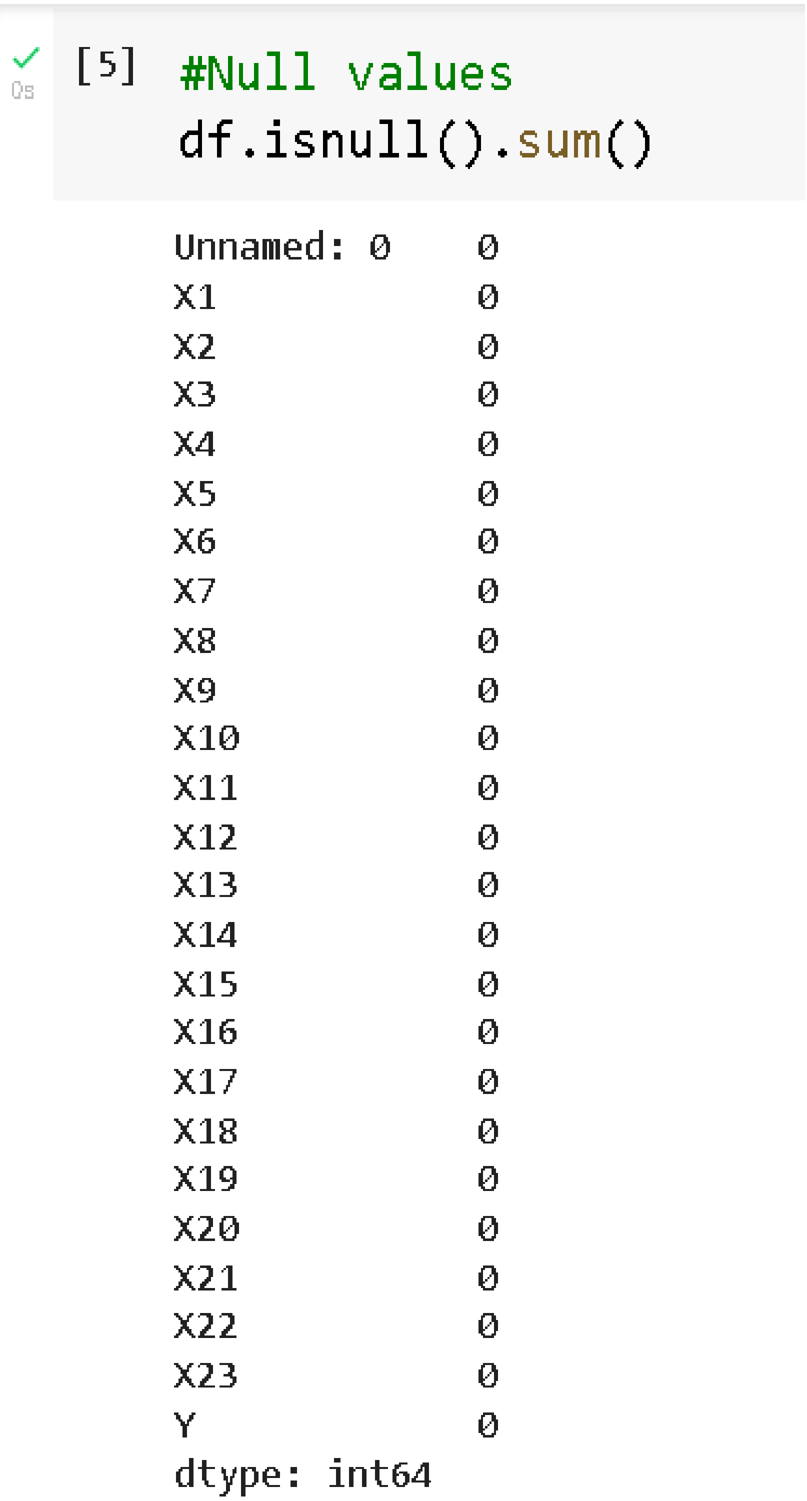
- X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Data Inspection:



```
#title Default title text
df.info()
```

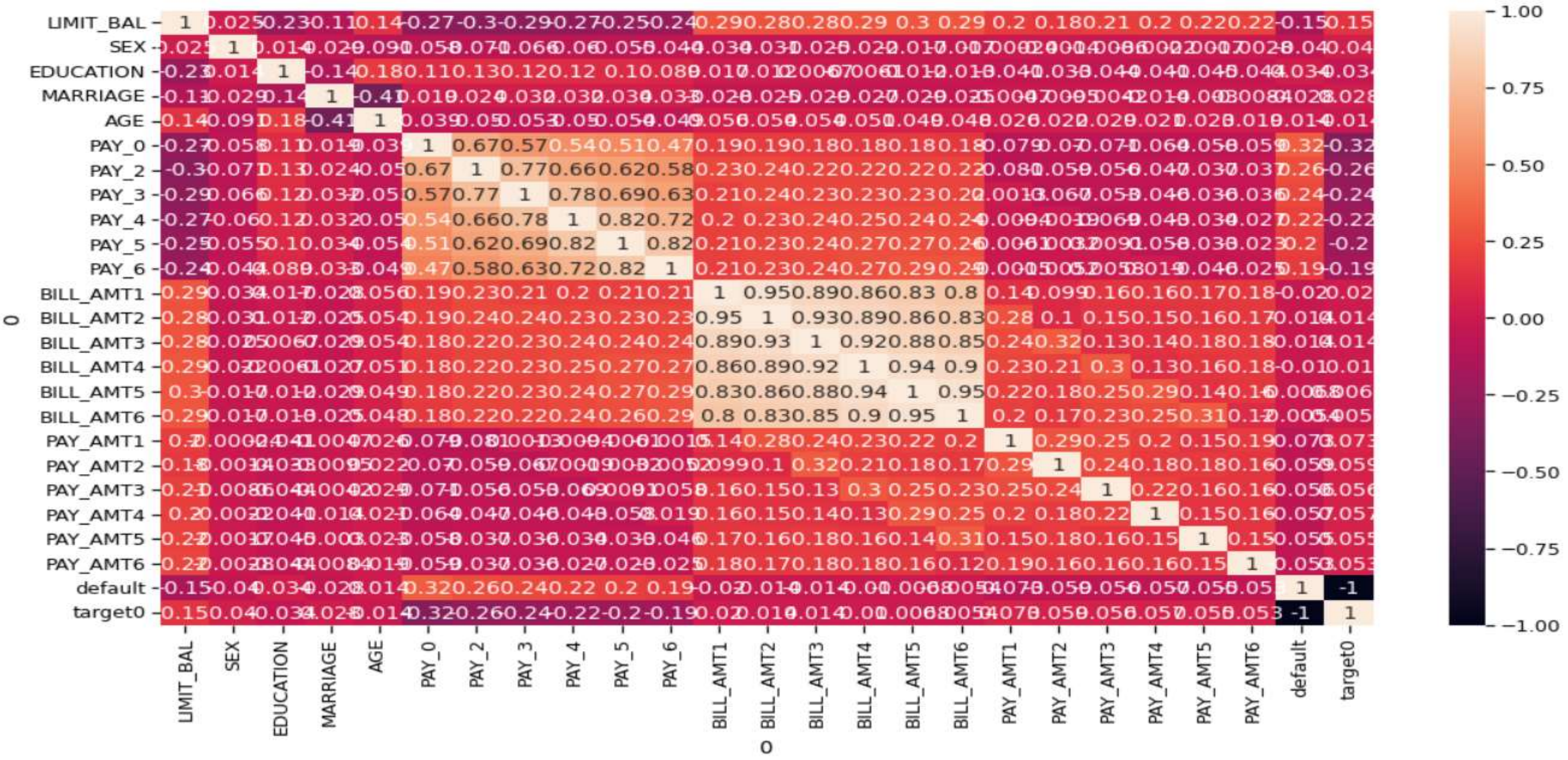
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30001 entries, 0 to 30000
Data columns (total 25 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  30001 non-null  object
 1   X1          30001 non-null  object
 2   X2          30001 non-null  object
 3   X3          30001 non-null  object
 4   X4          30001 non-null  object
 5   X5          30001 non-null  object
 6   X6          30001 non-null  object
 7   X7          30001 non-null  object
 8   X8          30001 non-null  object
 9   X9          30001 non-null  object
 10  X10         30001 non-null  object
 11  X11         30001 non-null  object
 12  X12         30001 non-null  object
 13  X13         30001 non-null  object
 14  X14         30001 non-null  object
 15  X15         30001 non-null  object
 16  X16         30001 non-null  object
 17  X17         30001 non-null  object
 18  X18         30001 non-null  object
 19  X19         30001 non-null  object
 20  X20         30001 non-null  object
 21  X21         30001 non-null  object
 22  X22         30001 non-null  object
 23  X23         30001 non-null  object
 24  Y           30001 non-null  object
dtypes: object(25)
memory usage: 5.7+ MB
```



```
#Null values
df.isnull().sum()
```

```
Unnamed: 0    0
X1            0
X2            0
X3            0
X4            0
X5            0
X6            0
X7            0
X8            0
X9            0
X10           0
X11           0
X12           0
X13           0
X14           0
X15           0
X16           0
X17           0
X18           0
X19           0
X20           0
X21           0
X22           0
X23           0
Y             0
dtype: int64
```

Correlation Heatmap



The only feature with a notable positive correlation with the dependent variable ‘Default’ is re-payment status during the last month (September).

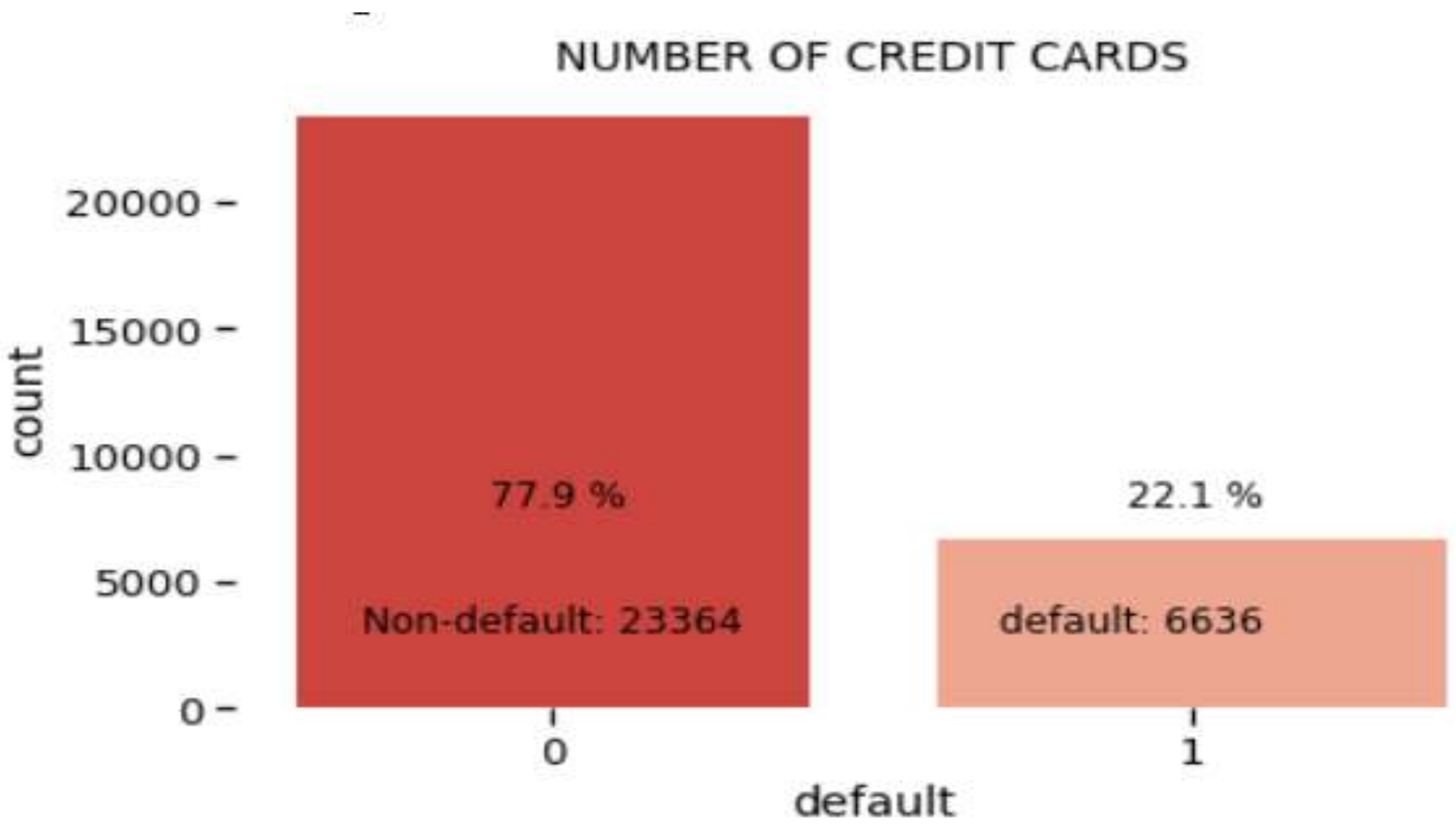
The highest negative correlation with default occurs with Limit_Balance, indicating that customers with lower limit balance are more likely to default.

Feature Engineering

1. In EDUCATION column the indication 4-others,5-unknown ,0 and 6-unknown can be merged in single class 4.

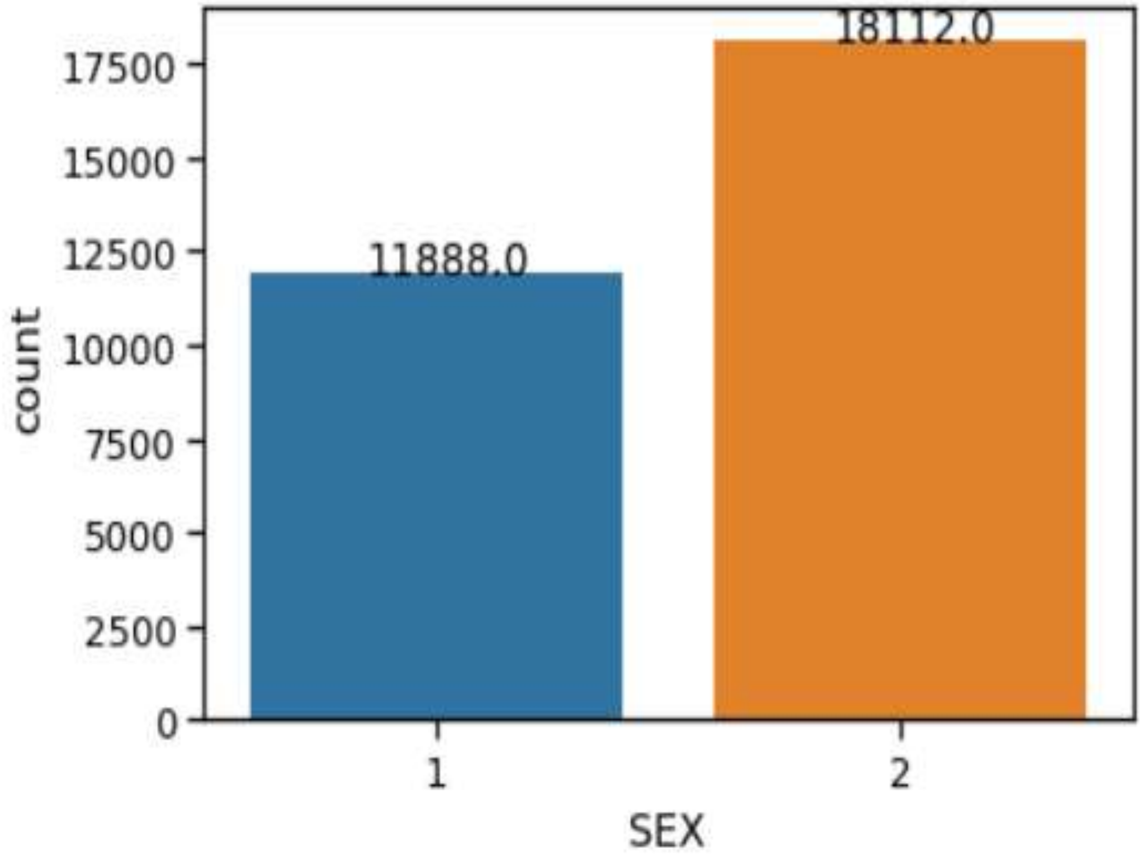
2. In MARRIAGE column we can merge 0 in value 3-others which will valid according to given data.

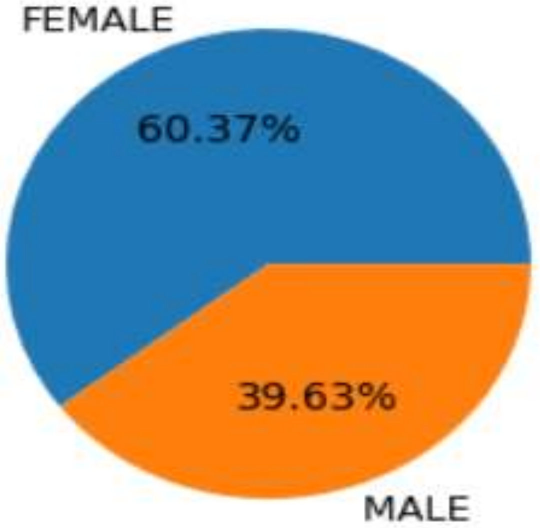
Exploratory Data Analysis



We can see that there are 6,636 default credit cards and 23364 non-default credit cards that is, the proportion of default in the data is 22,1% and 77.9% for non-default.

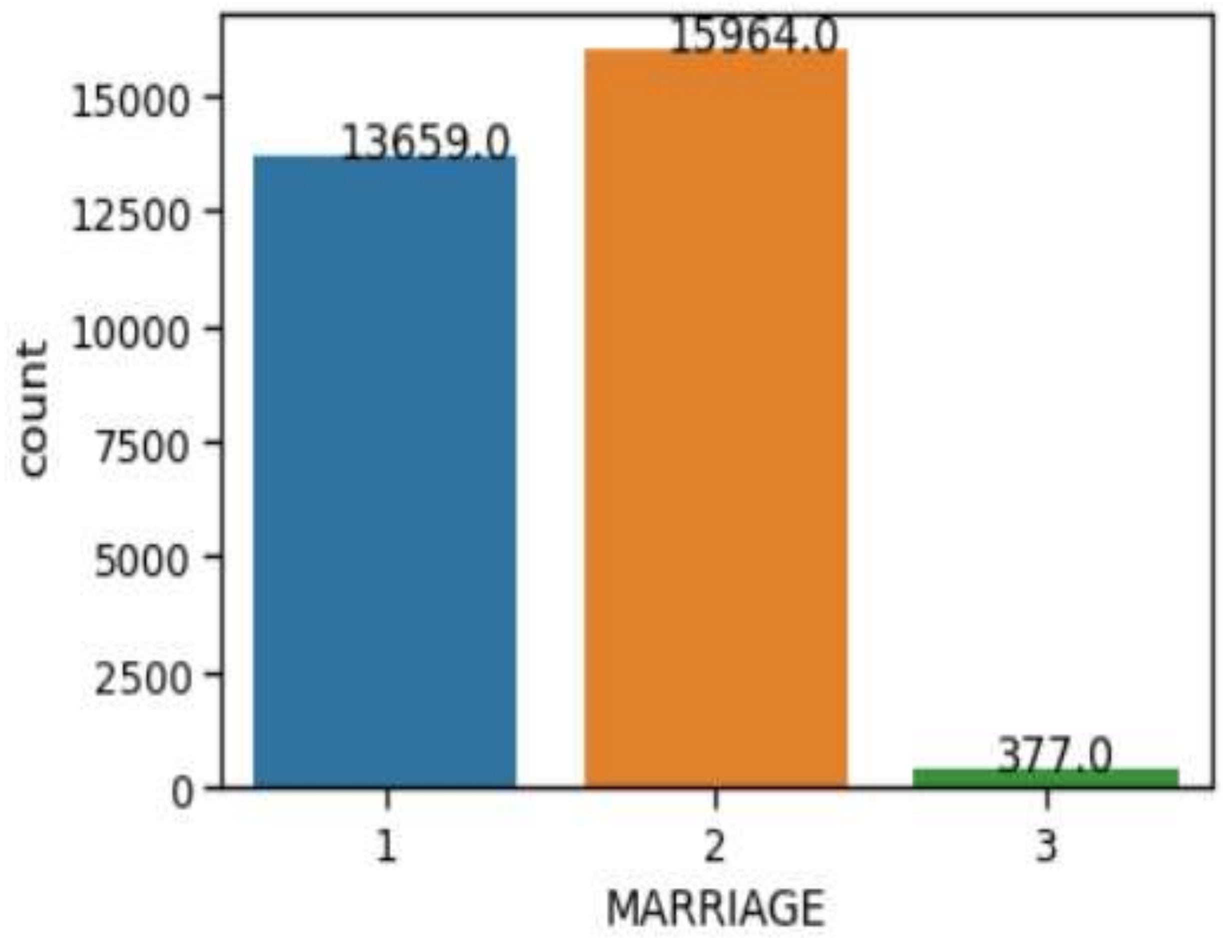
Exploratory Data Analysis on Gender

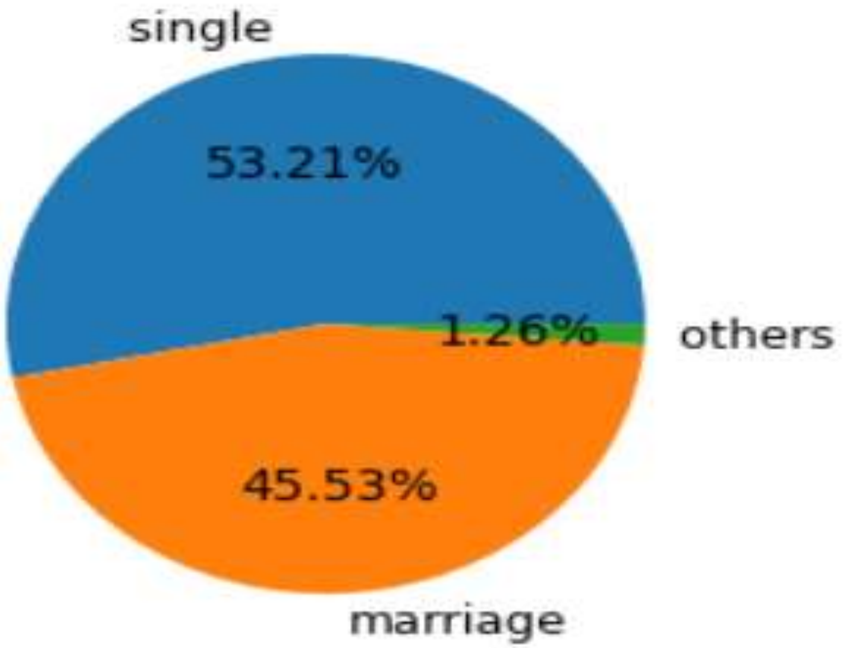




Females are more in number than male in using credit cards.

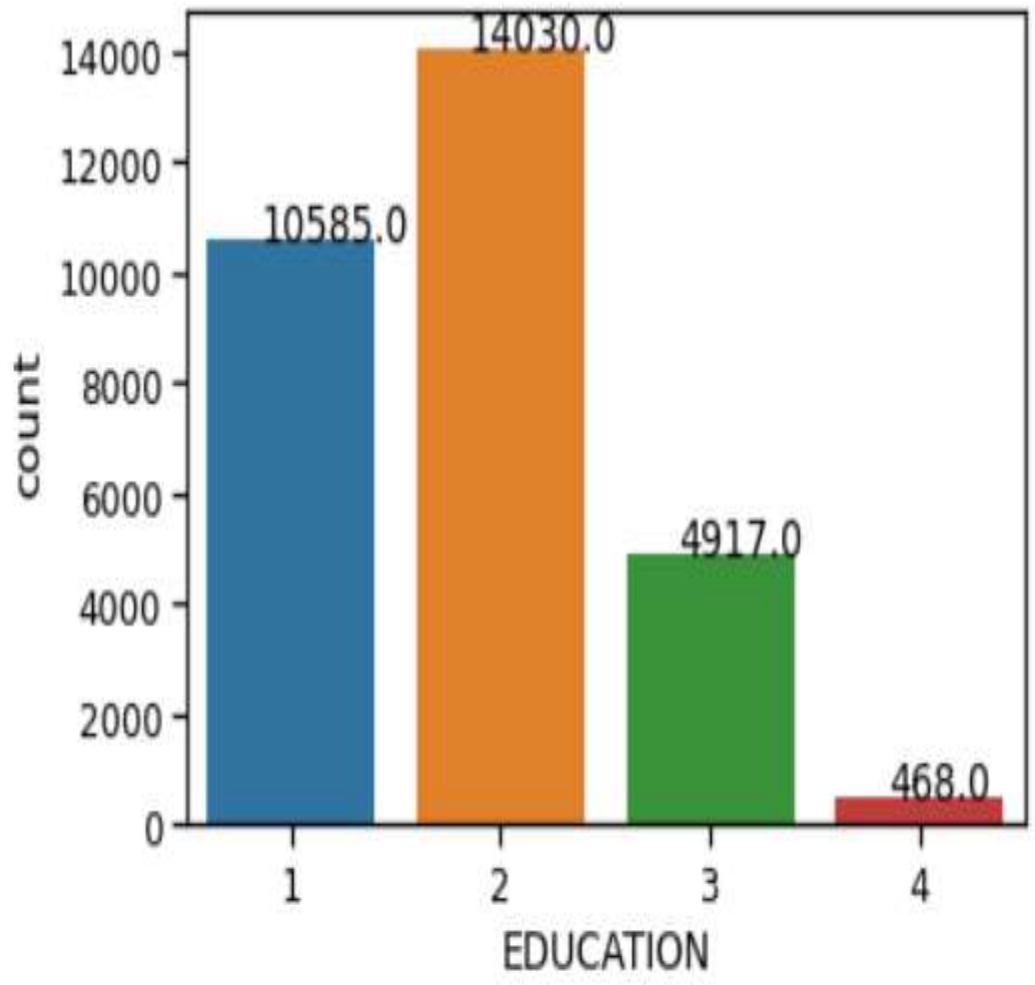
Exploratory Data Analysis on Marriage

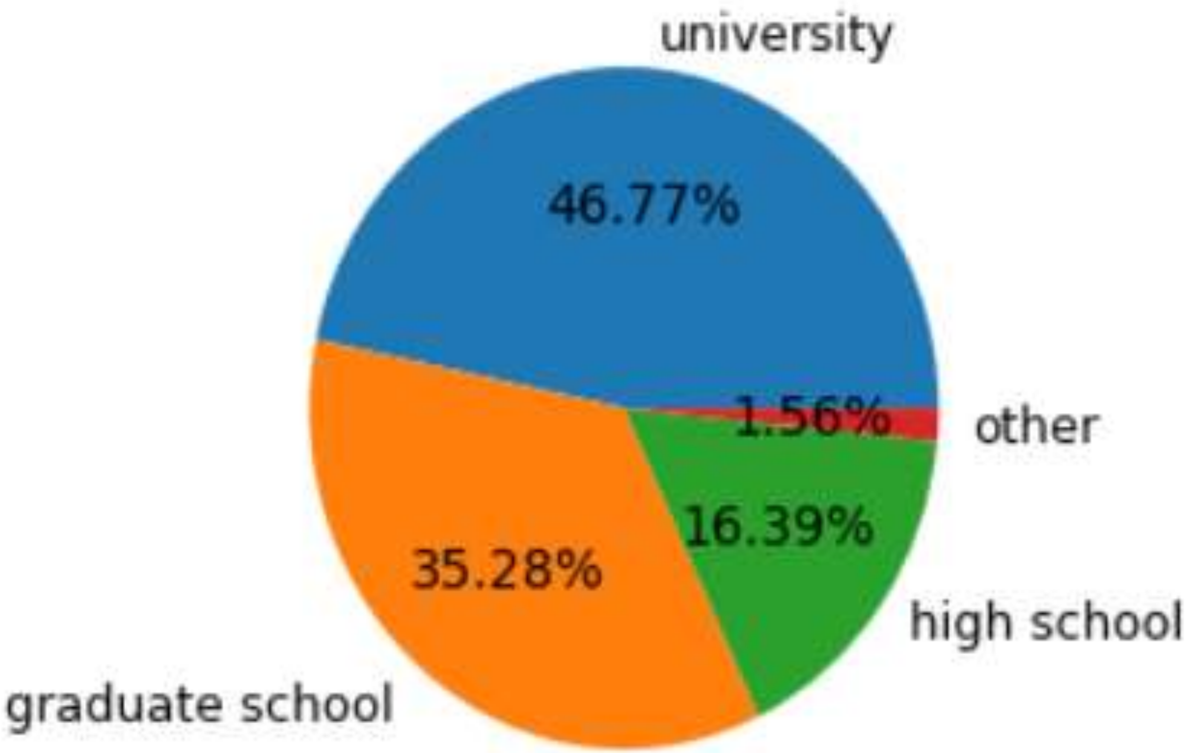




→It is clear from the above plot that more than half of credit card holder are single and very less are others like divorce or seperated..

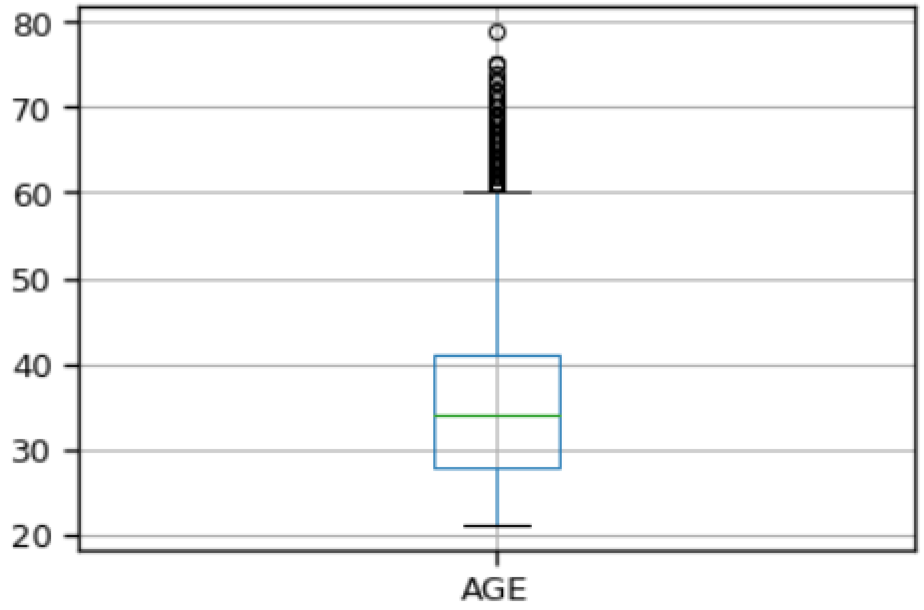
Exploratory Data Analysis on Education





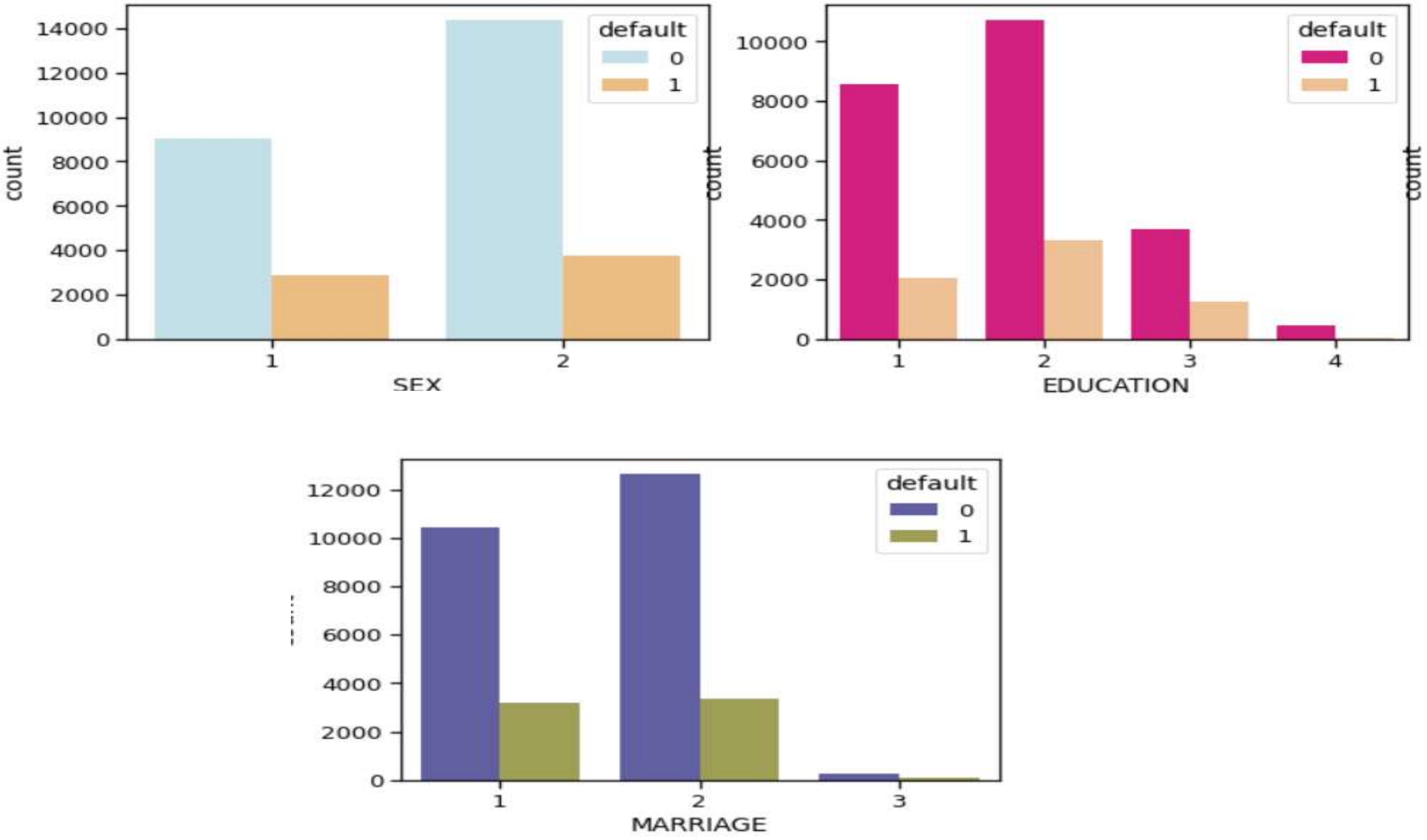
we can say that most the credit card holder are educated ,approx half of the credit card holder are educated at university level

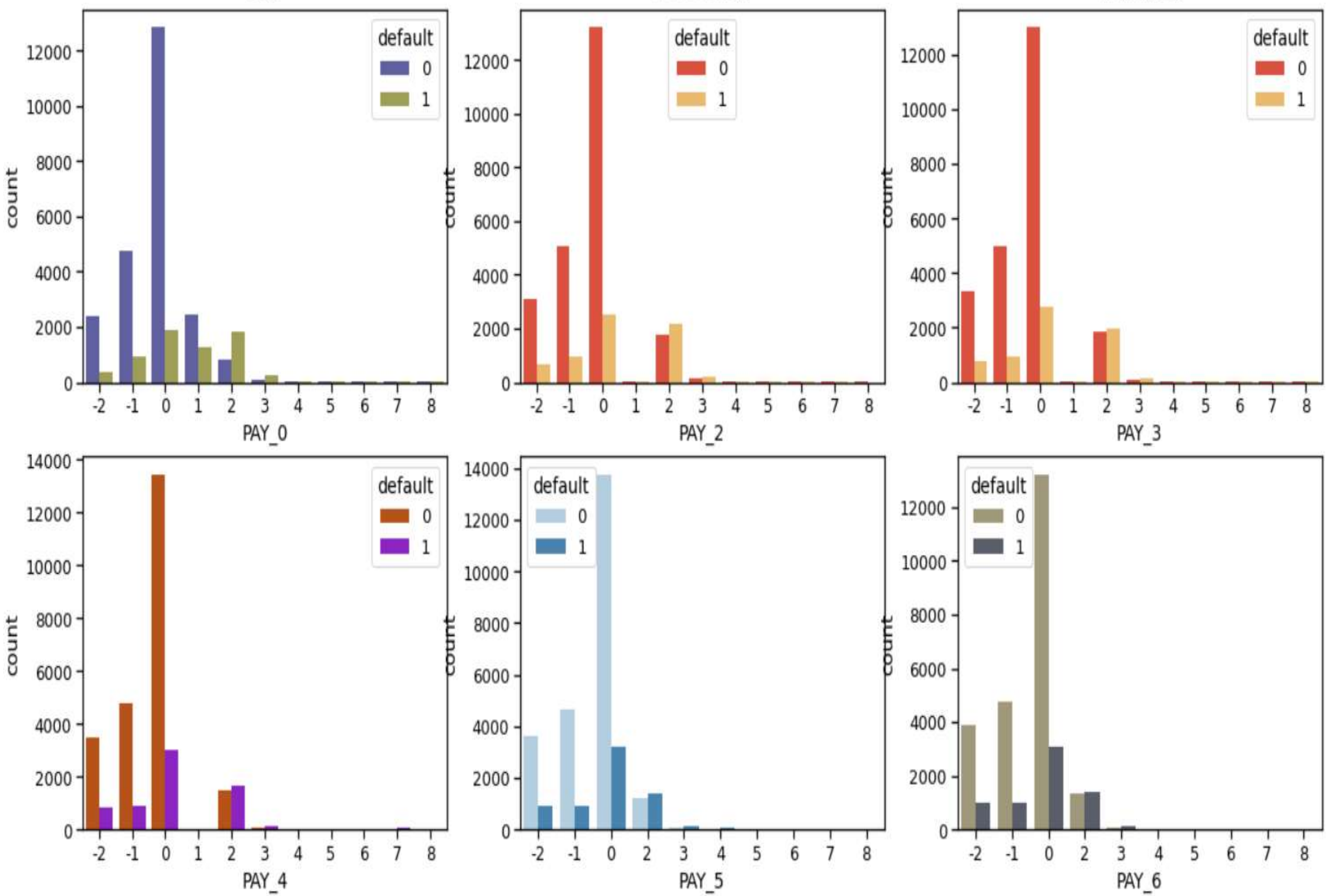
Exploratory Data Analysis on Age



- Mostly the credit card holder are in the age group of 30 to 40 years.

Explanatory variables by defaulted and non-defaulted cards



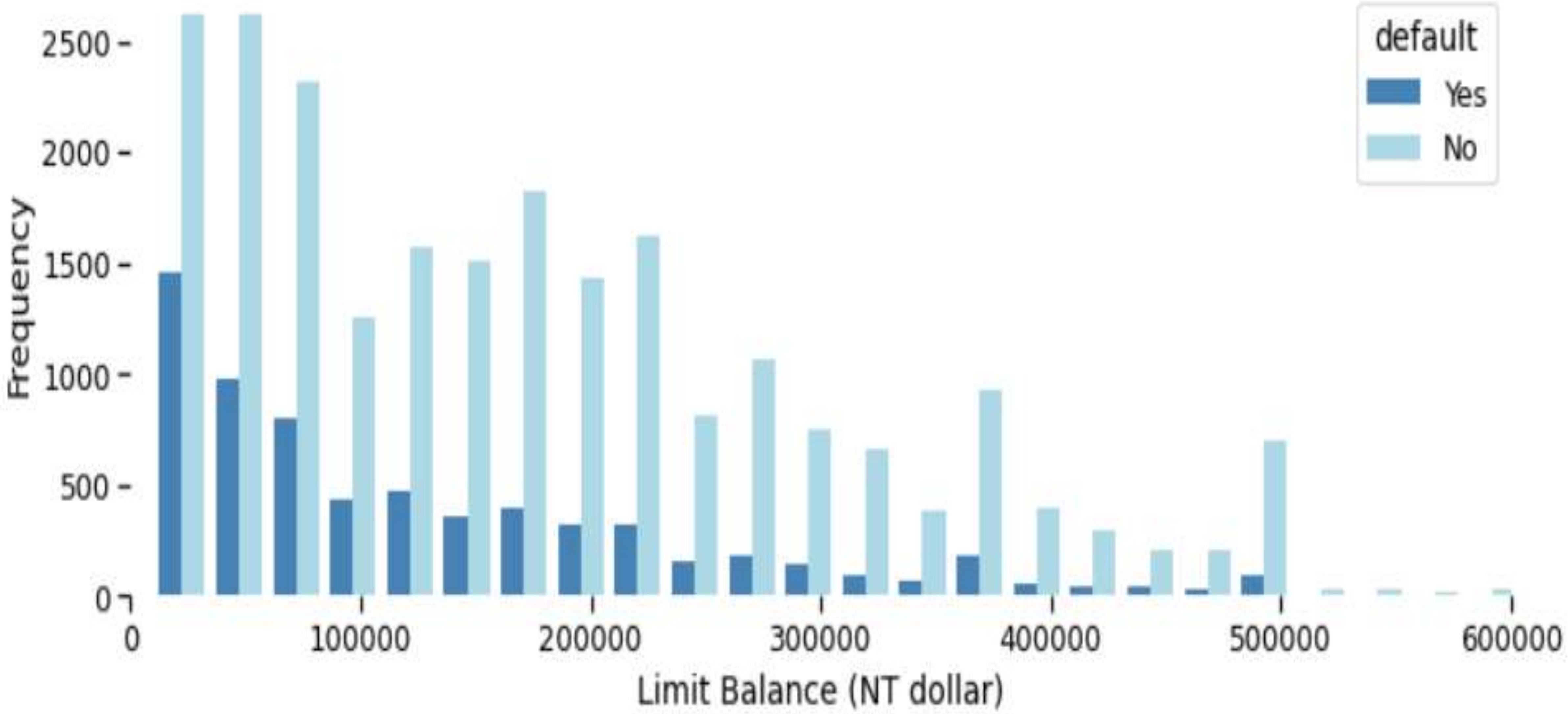

count
default
0
1
0
2000
4000
6000
8000
10000
12000
14000
-2
-1
0
1
2
3
4
5
6
7
8
PAY_0
PAY_2
PAY_3
PAY_4
PAY_5
PAY_6

1.Female are more in number than male in using credit cards,NonDefaults have a higher proportion of Females (Sex=2).

2.Most the credit card holder are educated ,approx half of the credit card holder are educated at university level,NonDefaults have a higher proportion of MoreEducated (EDUCATION=1 or 2).

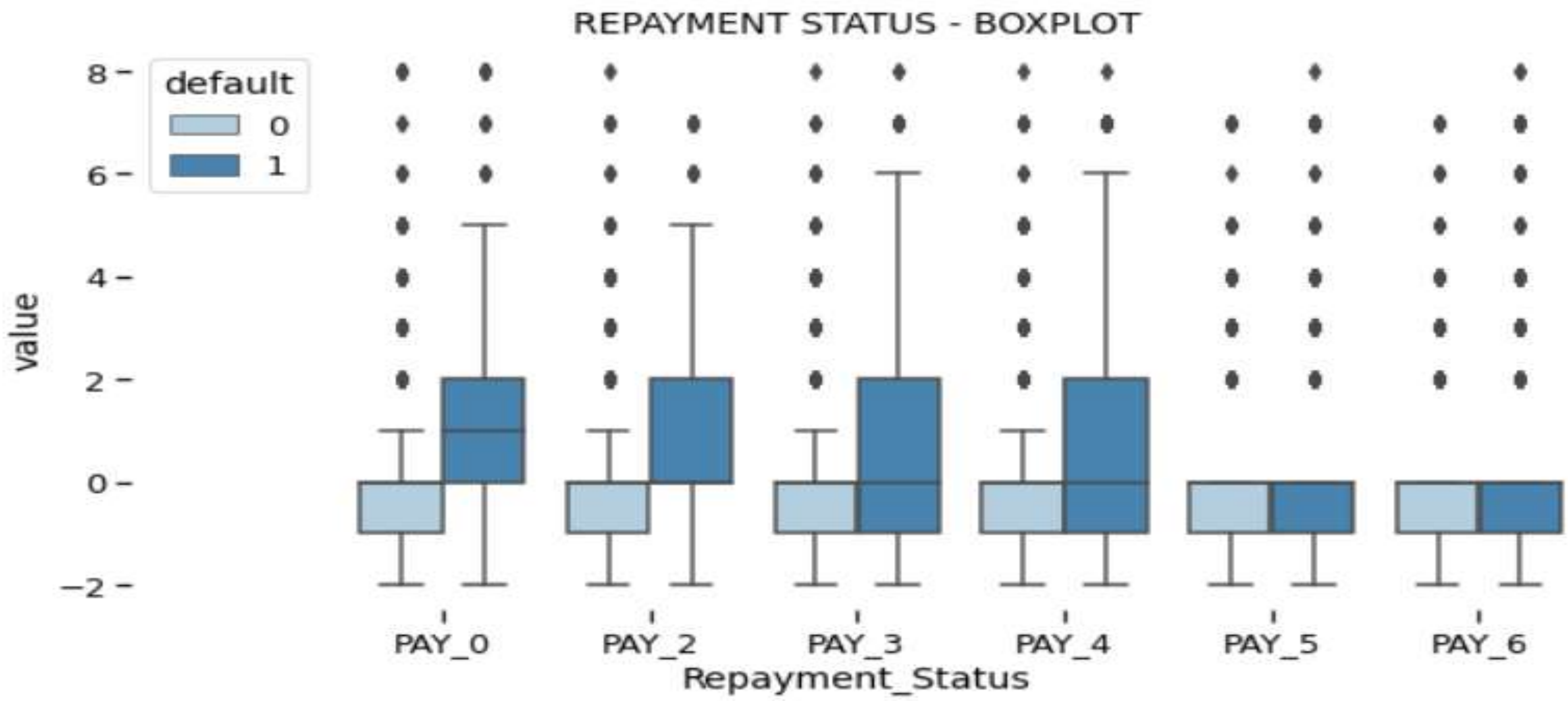
3. We see that being Female, More educated, Single and between 30-40years old means a customer is more likely to make payments on time.

Analysis of default based on credit card limit



You can see that for higher limit the default case are less

Boxplot for analysis of previous payment



- PAY_0 (Repayment status in September) and PAY_2 (Repayment status in August) have more discriminatory power the repayment status in other months.

Machine Learning Modelling

Model Selection and Evaluation :

Before building a models we performed the train test split. We kept 15% of the data for test and remaining 85% of the data for training the model.

We compared 4 algorithms and evaluated them based on the overall accuracy score and the recall of the individual classes.

- Accuracy is the ratio of the total number of correct predictions and the total number of predictions.
- The recall is the measure of our model correctly identifying True Positives.

1)Logistic regression ML algorithm

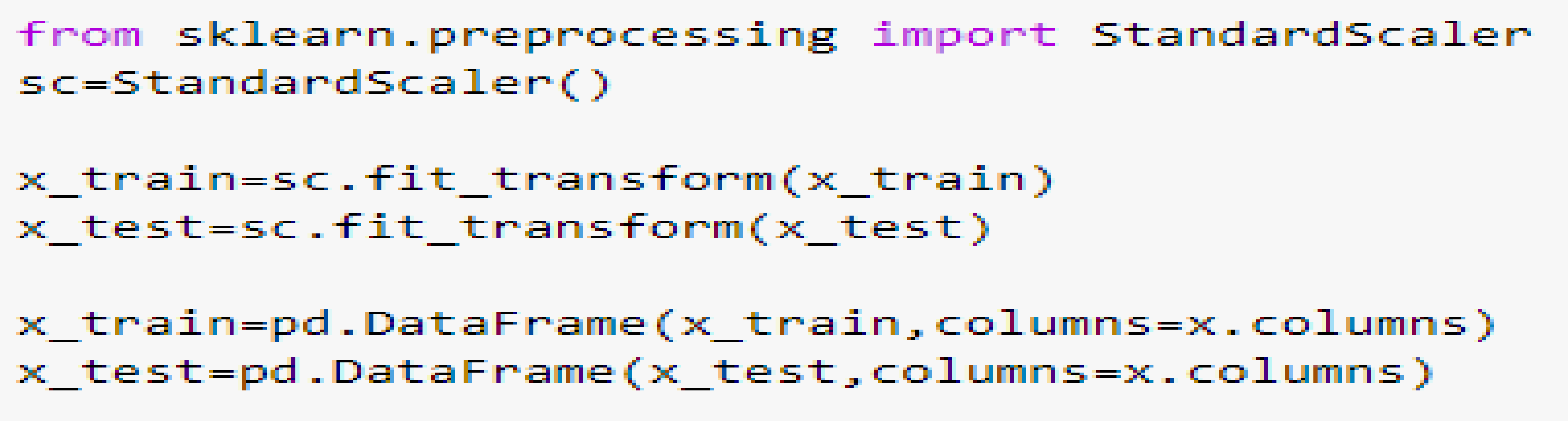
2)LightGBM Classifier

3) Decision Tree classifier

4) Random Forest classifier

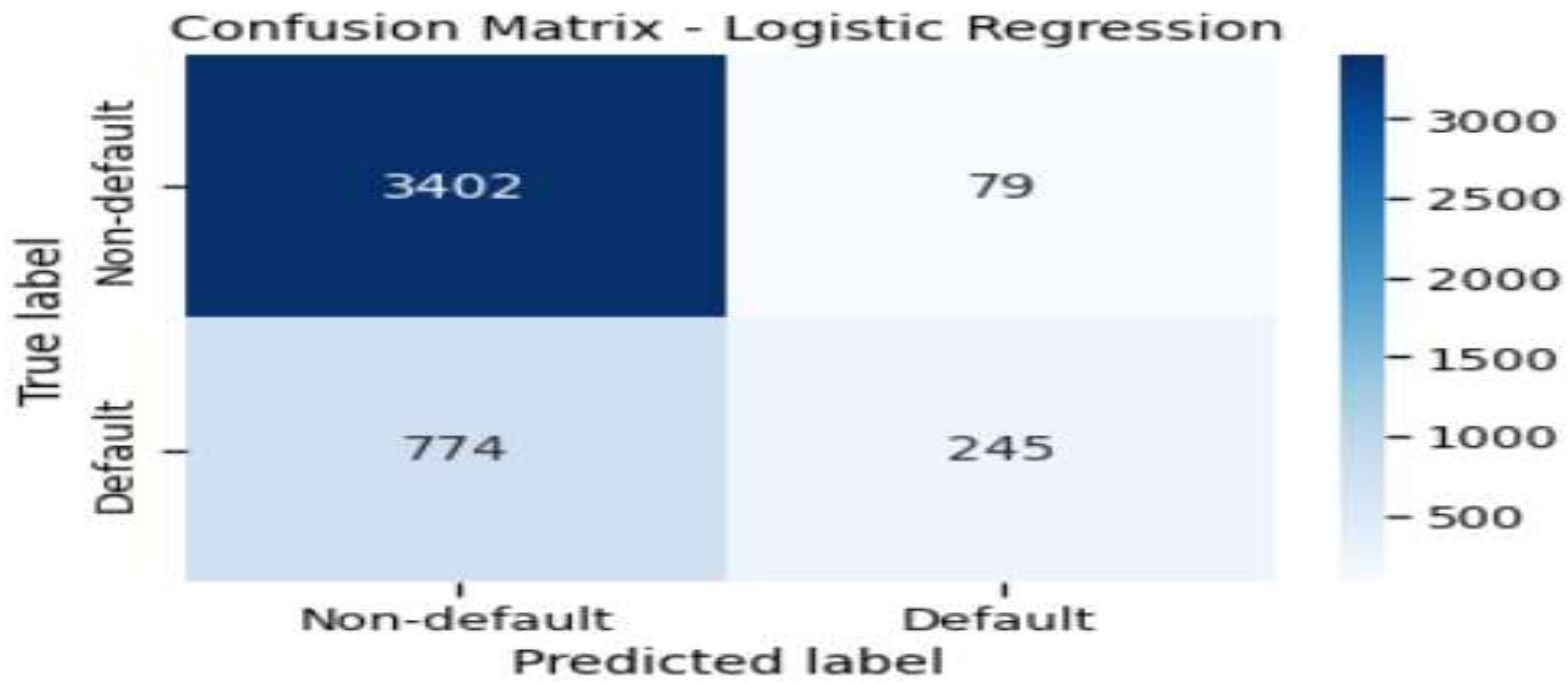
Standardization

Standardization is an important technique that is mostly performed as a preprocessing step before many Machine Learning models, to standardize the range of features of input data set.



```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()

x_train=sc.fit_transform(x_train)
x_test=sc.fit_transform(x_test)

x_train=pd.DataFrame(x_train,columns=x.columns)
x_test=pd.DataFrame(x_test,columns=x.columns)
```

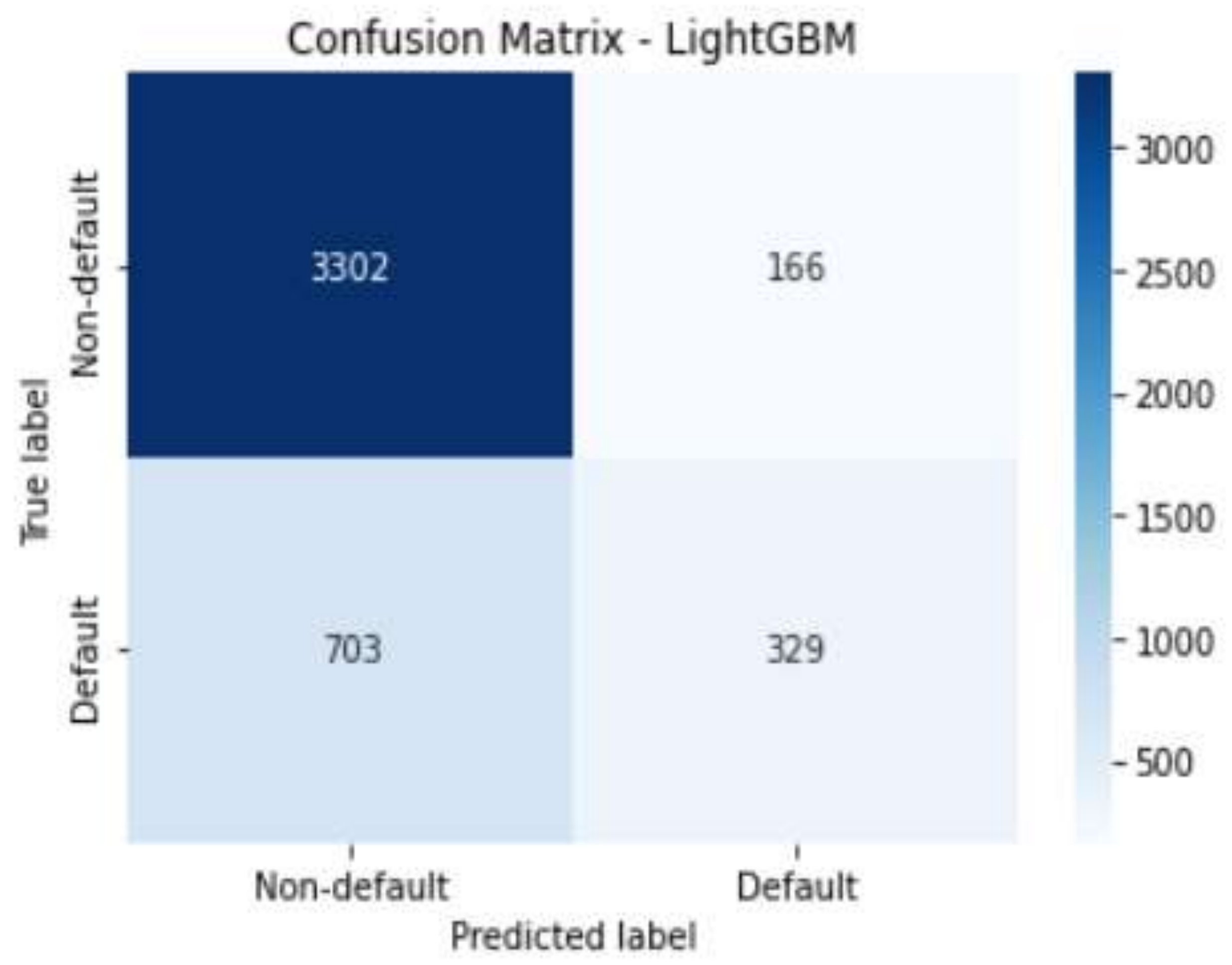
Implementing Logistic regression ML algorithm for classification

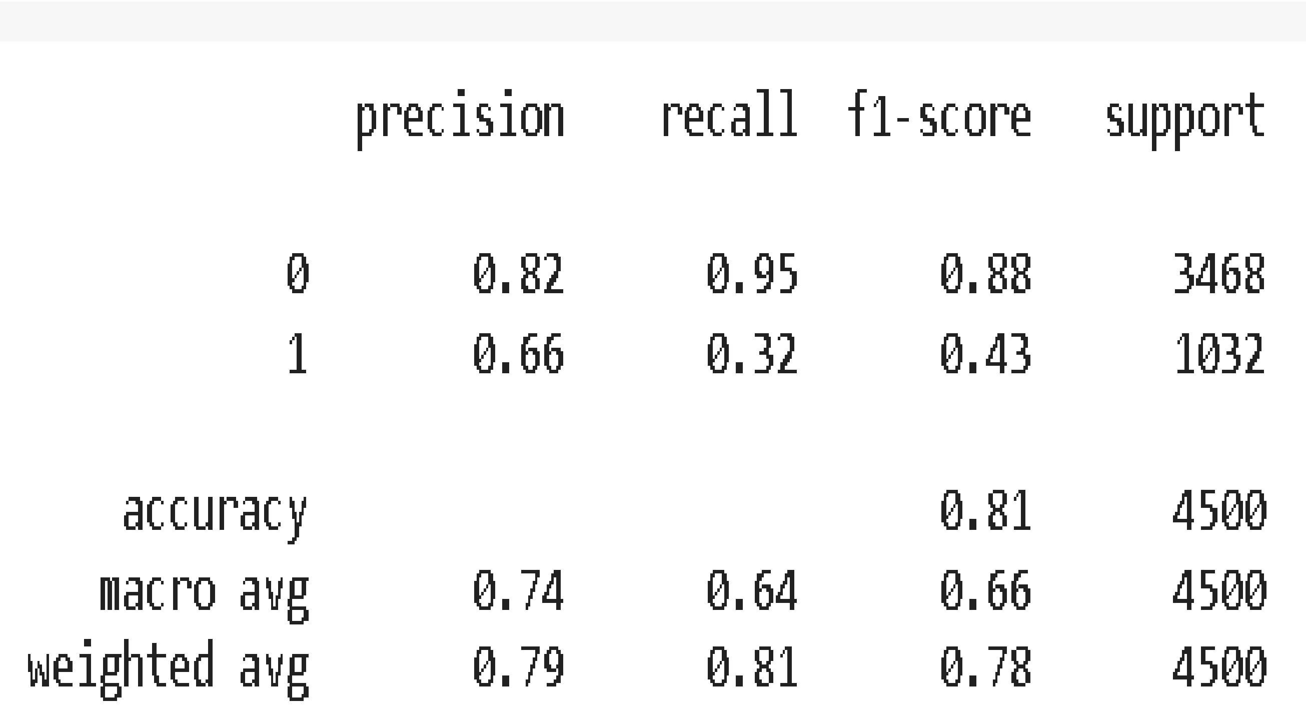


The Diagonal labels are true predicted labels .All other labels are falsely predicted variables .

we achieved 81 % accuracy with logistic regression.

Implementing LightGBM Classifier model

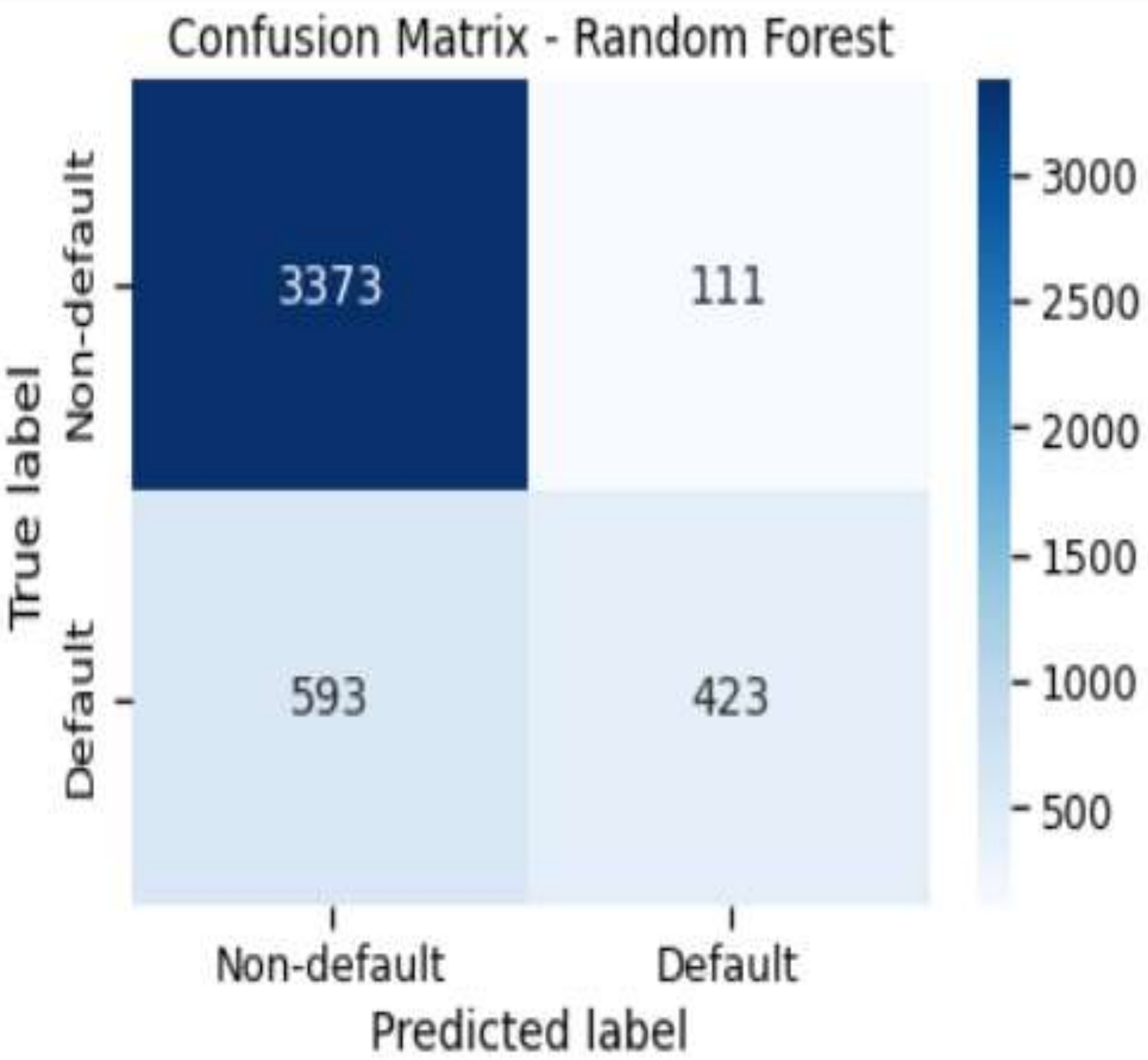


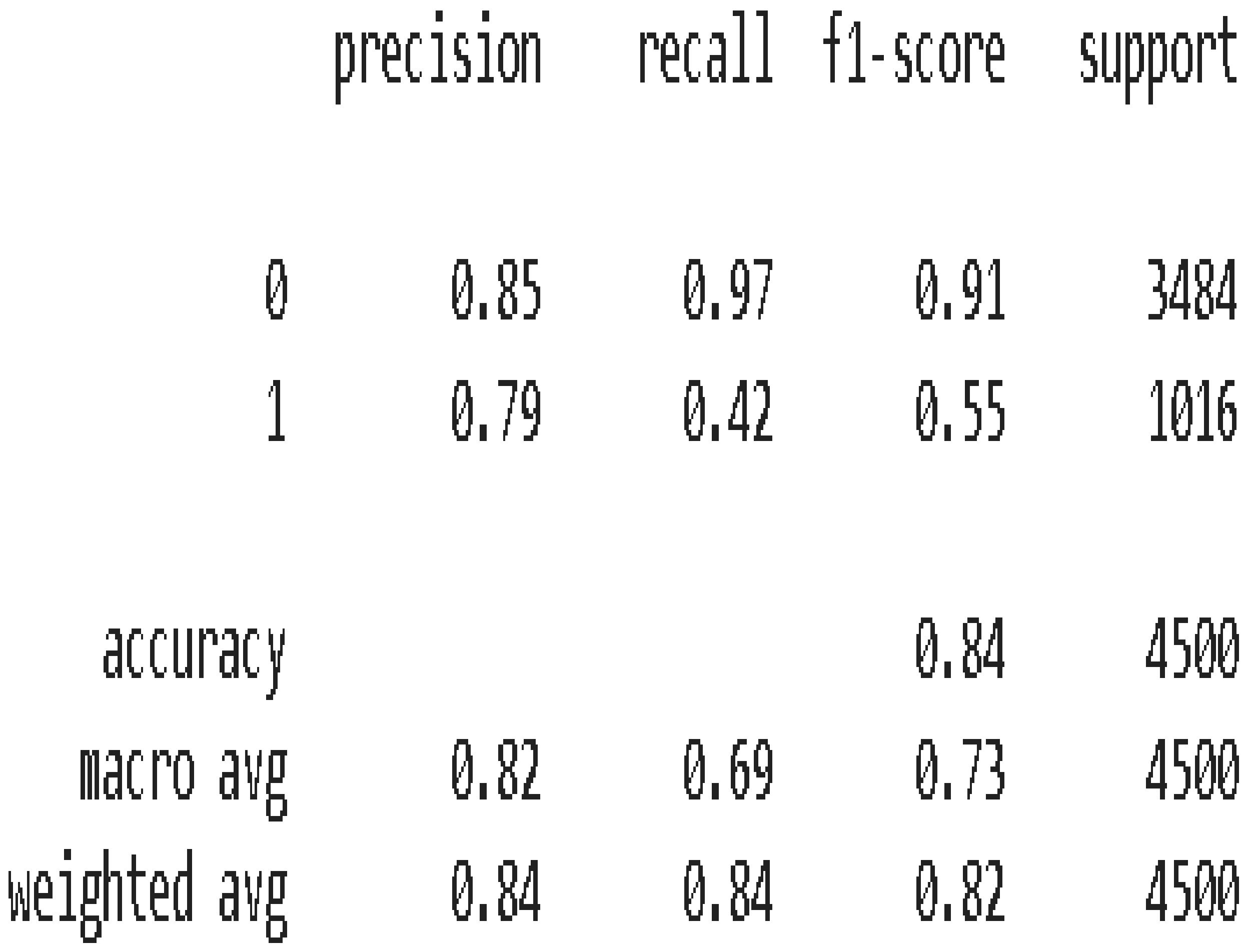


```
              precision    recall  f1-score   support

           0       0.82      0.95      0.88      3468
           1       0.66      0.32      0.43      1032

    accuracy                           0.81      4500
   macro avg       0.74      0.64      0.66      4500
weighted avg       0.79      0.81      0.78      4500
```

Accuracy with LightGBM is equal to 81% and F1 score is 88% which is good for our model.

Implementing Random Forest Classifier model

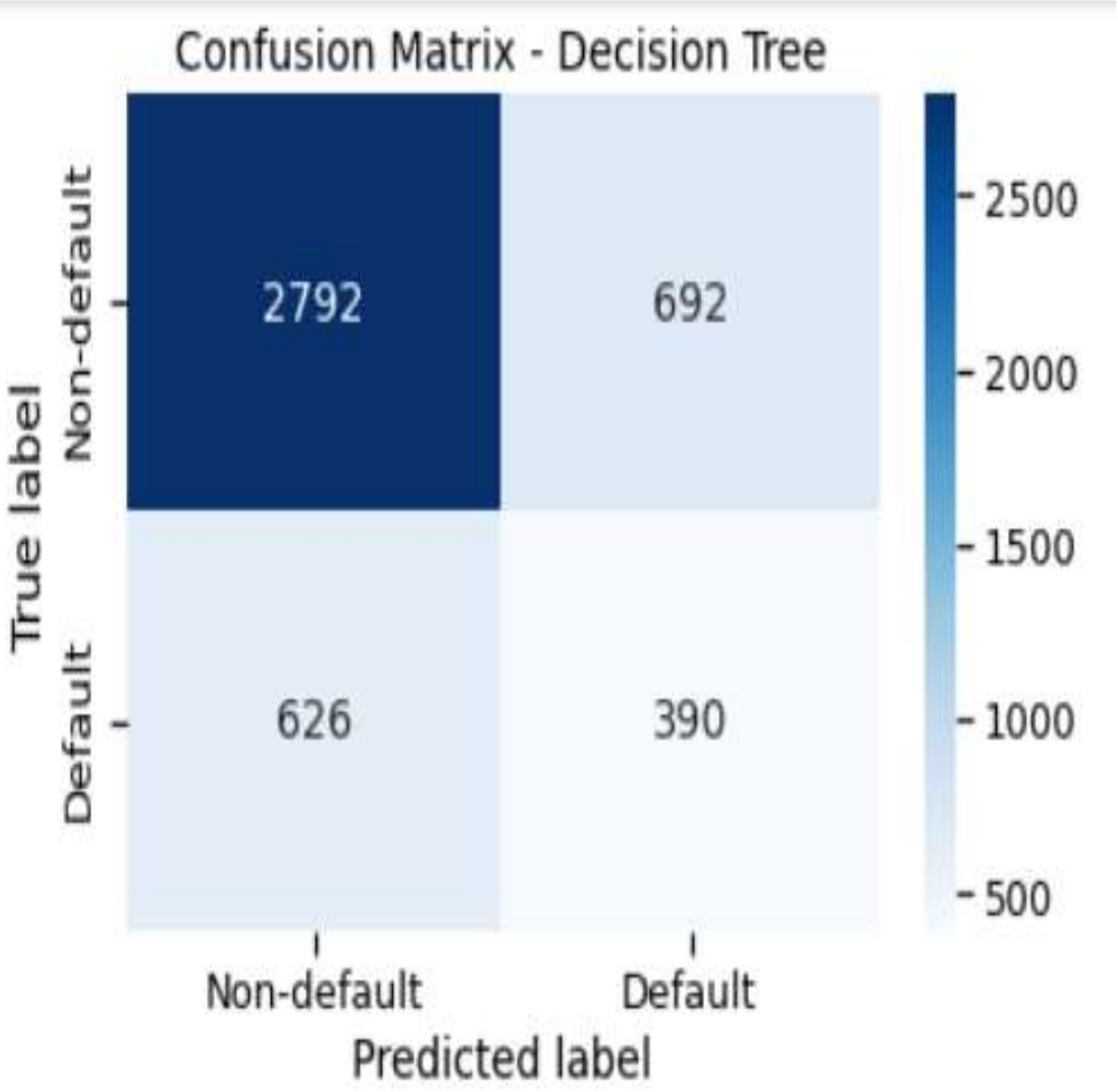


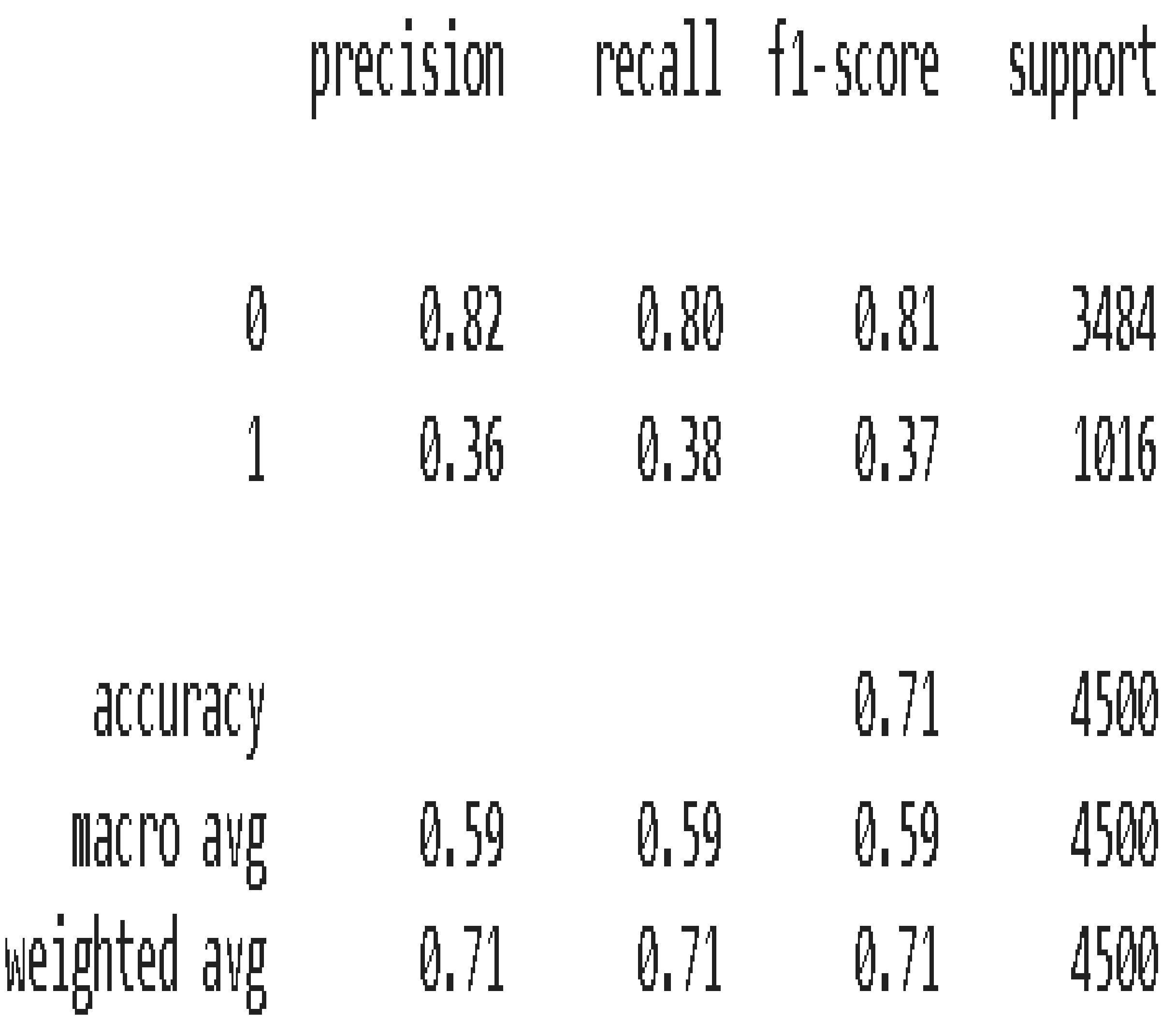

```
              precision    recall  f1-score   support

           0       0.85      0.97      0.91      3484
           1       0.79      0.42      0.55      1016

    accuracy                           0.84      4500
   macro avg       0.82      0.69      0.73      4500
weighted avg       0.84      0.84      0.82      4500
```

The Diagonal labels are true predicted labels .All other labels are falsely predicted variables .

WE achieved accuracy of 84% in test dataset with Random forest.

Implementing Decision tree classification ML Algorithm



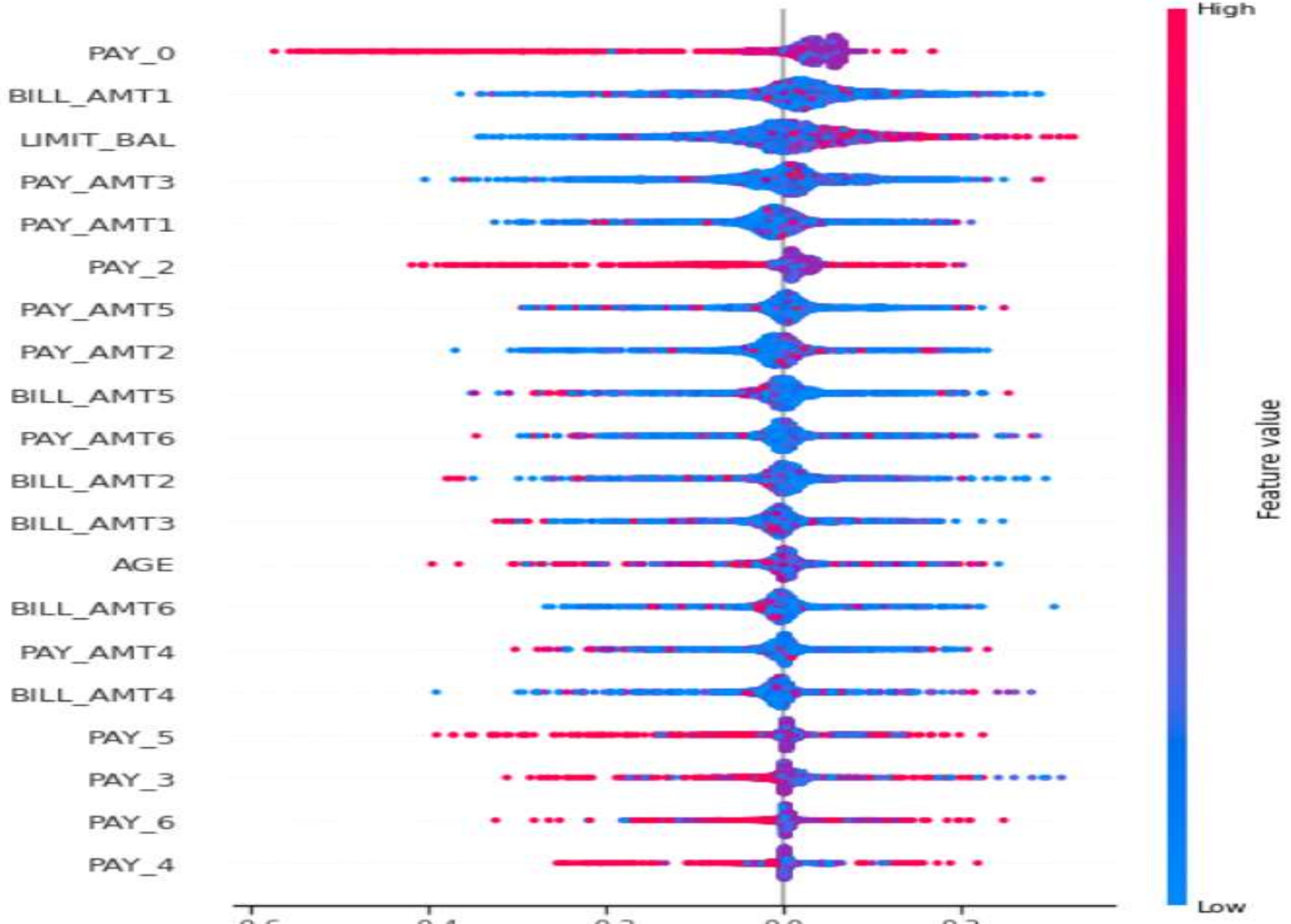


```
              precision    recall  f1-score   support

           0       0.82      0.80      0.81      3484
           1       0.36      0.38      0.37      1016

    accuracy                           0.71      4500
   macro avg       0.59      0.59      0.59      4500
weighted avg       0.71      0.71      0.71      4500
```

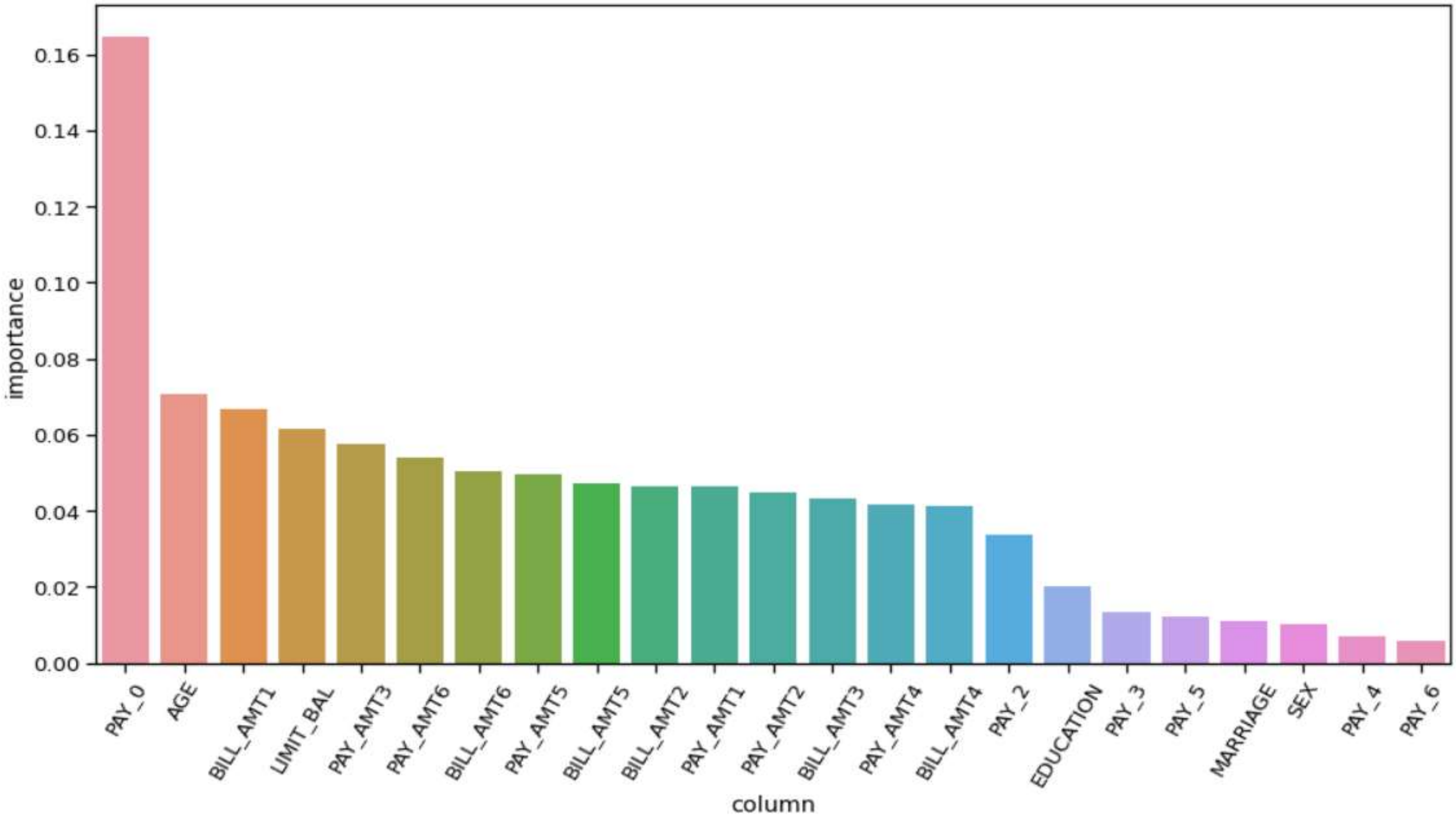
Evaluation of models:

Algorithm	Accuracy
Logistic regression	0.80
Light gbm	0.81
Random Forest	0.84
Decision tree	0.71

Interpretation or Justification of features



feature importance of model



Conclusion:

1.We have 25 columns and 30001 rows in this dataset ,The given data was cleaned and balanced ,no need to clean data.

2.The only feature with a notable positive correlation with the dependent variable ‘Default’ is re-payment status during the last month (September). The highest negative correlation with default occurs with Limit_Balance, indicating that customers with lower limit balance are more likely to default.

3.The average value for the amount of credit card limit is 167,484 NT dollars. The standard deviation is 129,747 NT dollars, ranging from 10,000 to 1M NT dollars.

4.There are 6,636 default credit cards and 23364 non-default credit cards that is, the proportion of default in the data is 22,1% and 77.9% for non-default.

5.Female are more in number than male in using credit cards,NonDefaults have a higher proportion of Females (Sex=2).

6.Most the credit card holder are educated ,approx half of the credit card holder are educated at university level,NonDefaults have a higher proportion of MoreEducated (EDUCATION=1 or 2).

Conclusion:

7. Defaults have a higher proportion of Lower LIMIT_BAL values.
8. We splitted the data as to train our model with 85% and test our model with 15% of the total data available.
9. Highest accuracy is achieved by Random Forest model equal to 81 % and with Logistic Regression accuracy is 81%.
10. PAY_X ,AGE and BILL_AMT are the most important feature which give result of default..

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

Thank You !