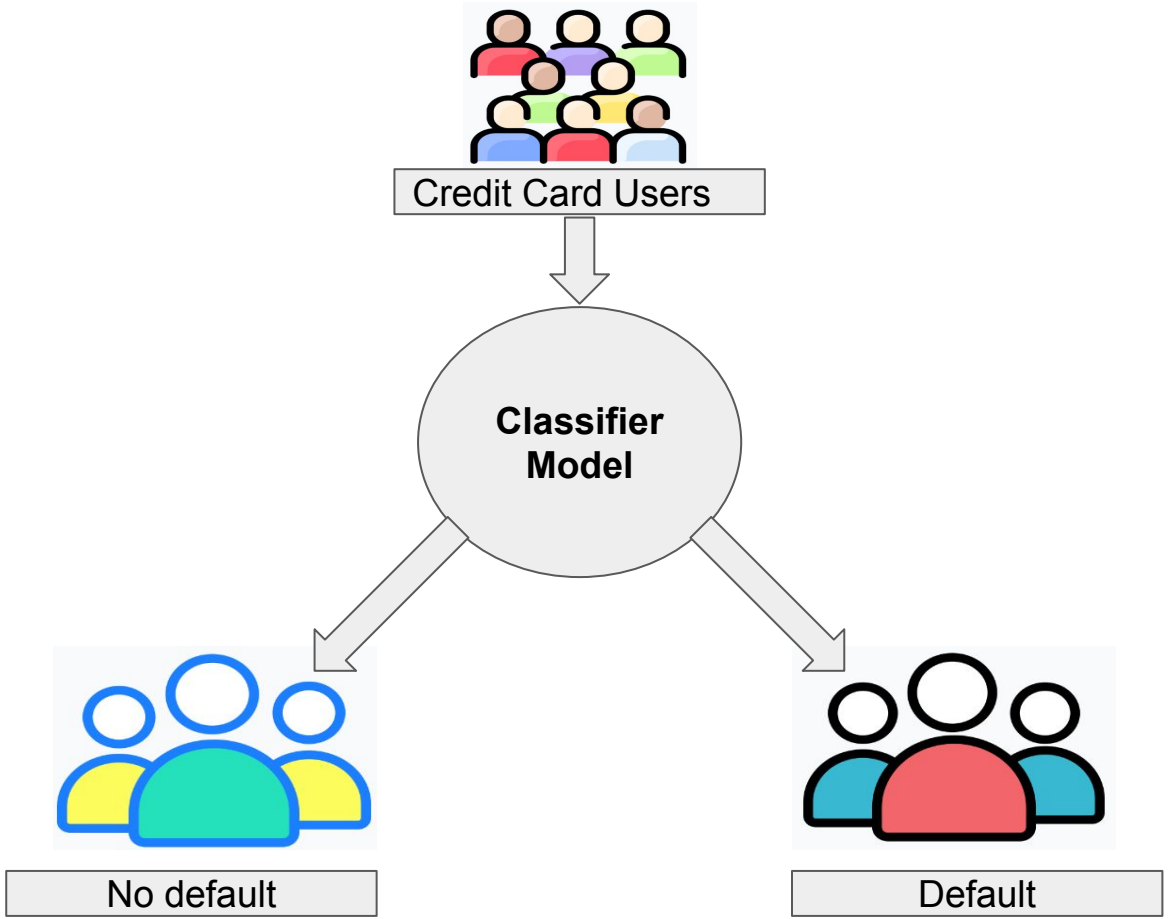
Supervised Machine Learning - Classification

Topic - Credit Card Default Prediction

By:-

Ajit kumar Toppo

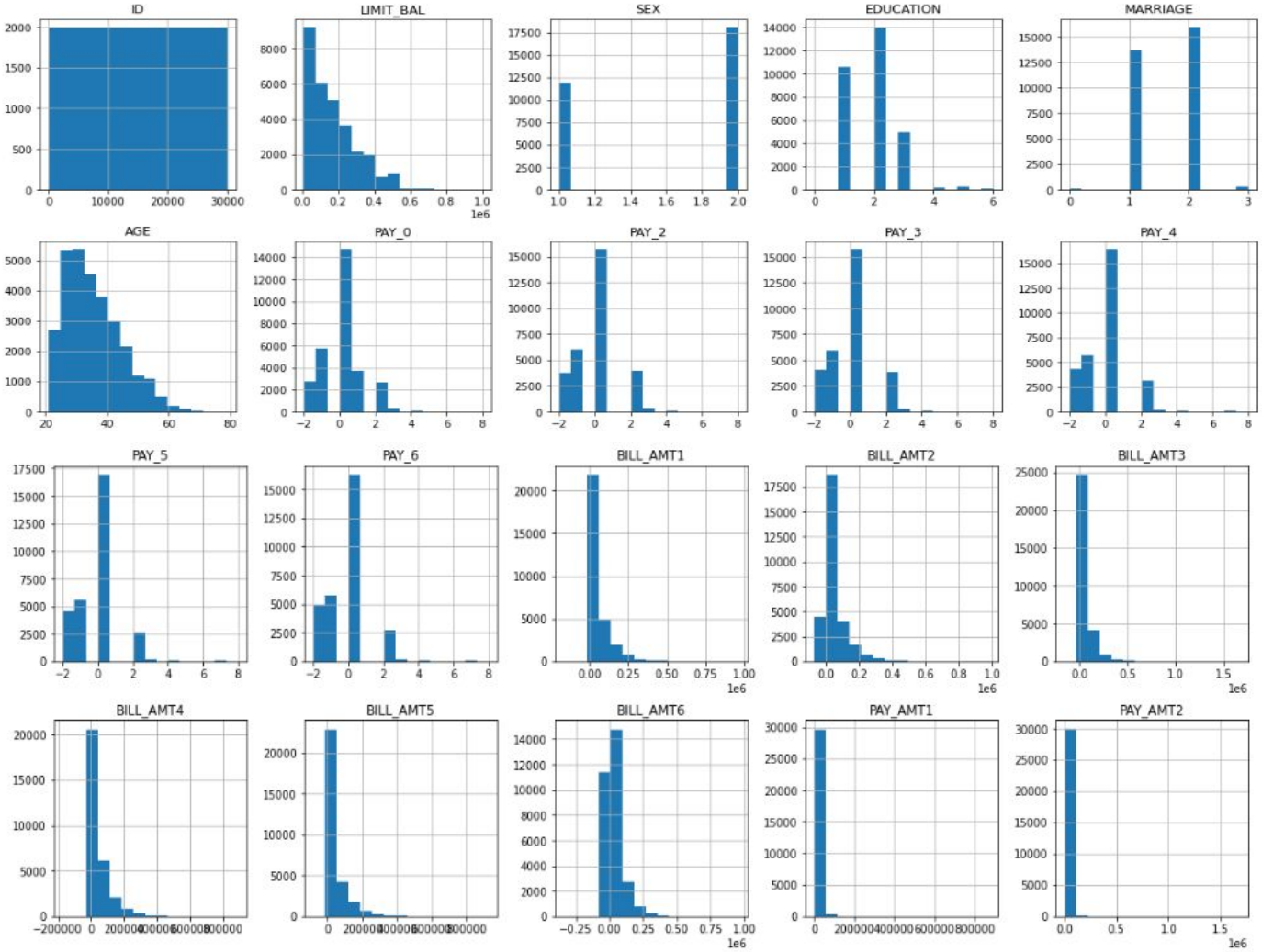

Credit Card Users
Classifier Model
No default
Default

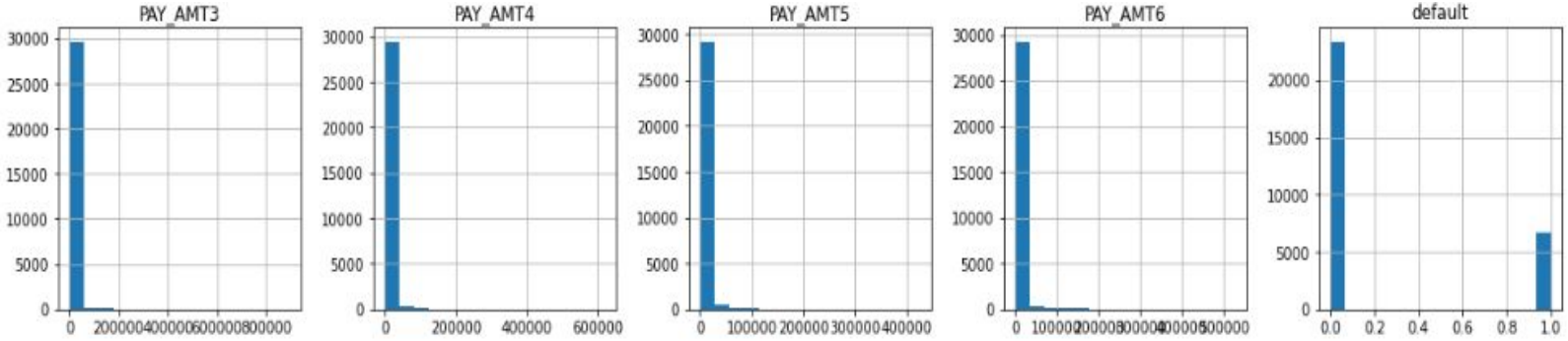
Data Description

1. **ID:** ID of each client
2.LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
3.SEX: Gender (1=Male, 2=Female)
4.EDUCATION: (1=Graduate school, 2=University, 3=High school, 0,4,5,6=others)
5.MARRIAGE: Marital status (1=married, 2=single, 0,3=others)
6.AGE: Age in years
7.PAY_0-6: Repayment status in September 2005 - April, 2005
8.BILL_AMT1-6: Amount of bill statement in September- April. 2005 (NT dollar)
9.PAY_AMT1-6: Amount of previous payment in September- April. 2005 (NT dollar)
10.Default payment next month: Default payment (1=yes, 0=no)

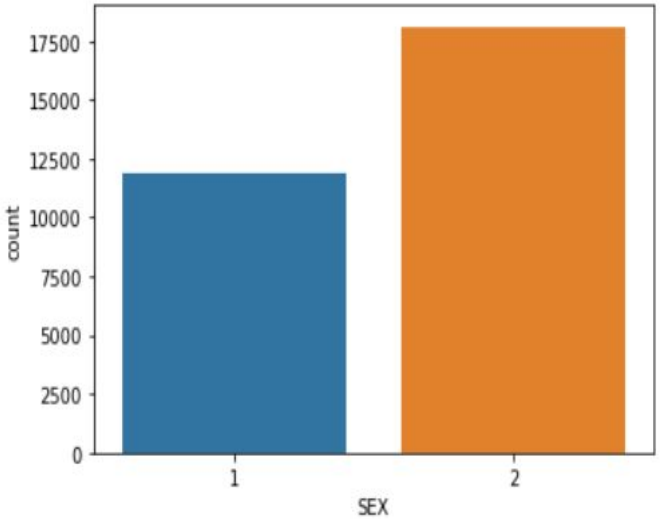
Dataset(rows-30000, columns-25)

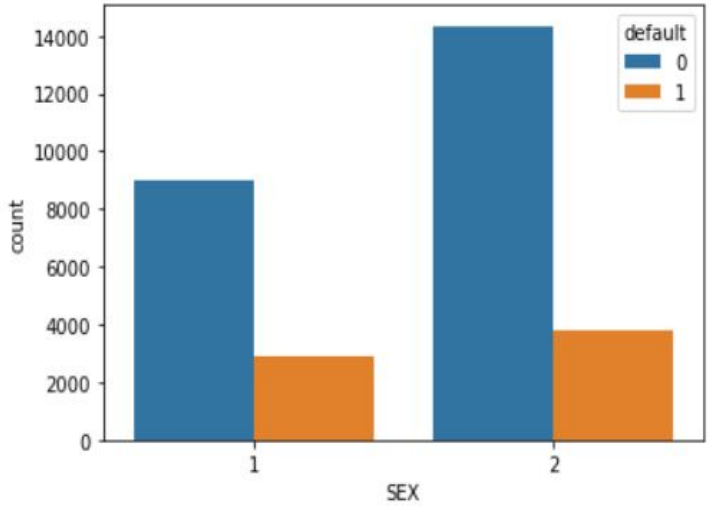
#	Column	Non-Null Count	Dtype
0	ID	30000 non-null	int64
1	LIMIT_BAL	30000 non-null	int64
2	SEX	30000 non-null	int64
3	EDUCATION	30000 non-null	int64
4	MARRIAGE	30000 non-null	int64
5	AGE	30000 non-null	int64
6	PAY_0	30000 non-null	int64
7	PAY_2	30000 non-null	int64
8	PAY_3	30000 non-null	int64
9	PAY_4	30000 non-null	int64
10	PAY_5	30000 non-null	int64
11	PAY_6	30000 non-null	int64
12	BILL_AMT1	30000 non-null	int64
13	BILL_AMT2	30000 non-null	int64
14	BILL_AMT3	30000 non-null	int64
15	BILL_AMT4	30000 non-null	int64
16	BILL_AMT5	30000 non-null	int64
17	BILL_AMT6	30000 non-null	int64
18	PAY_AMT1	30000 non-null	int64
19	PAY_AMT2	30000 non-null	int64
20	PAY_AMT3	30000 non-null	int64
21	PAY_AMT4	30000 non-null	int64
22	PAY_AMT5	30000 non-null	int64
23	PAY_AMT6	30000 non-null	int64
24	default payment next month	30000 non-null	int64


ID
LIMIT_BAL
SEX
EDUCATION
MARRIAGE
AGE
PAY_0
PAY_2
PAY_3
PAY_4
PAY_5
PAY_6
BILL_AMT1
BILL_AMT2
BILL_AMT3
BILL_AMT4
BILL_AMT5
BILL_AMT6
PAY_AMT1
PAY_AMT2
1e6


PAY_AMT3
PAY_AMT4
PAY_AMT5
PAY_AMT6
default

Number of male and female credit card holders



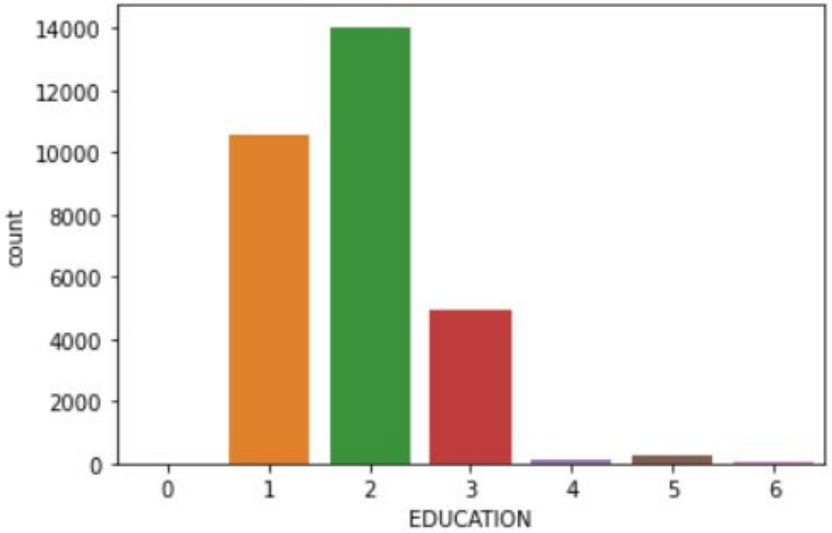


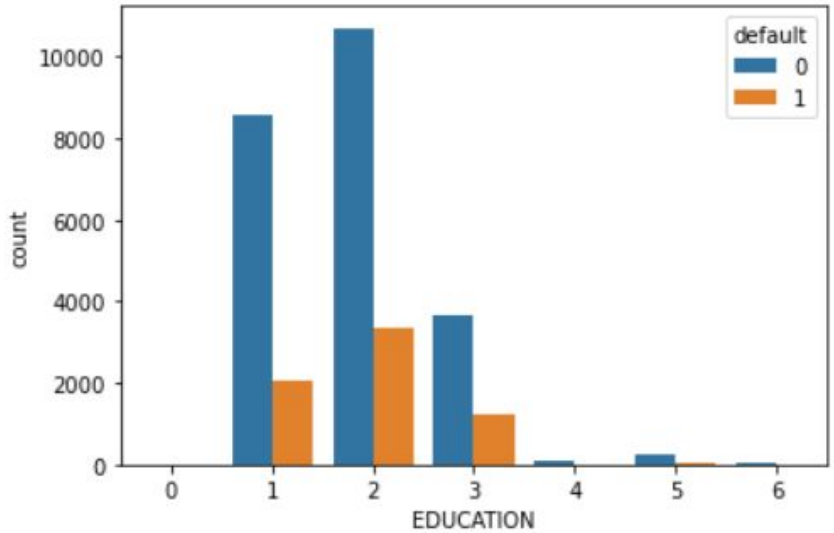
Female(2) -18112
Male(1) -11888

1= Male, 2= Female

- Number of females credit holder are higher than male.

Credit card holders based on education



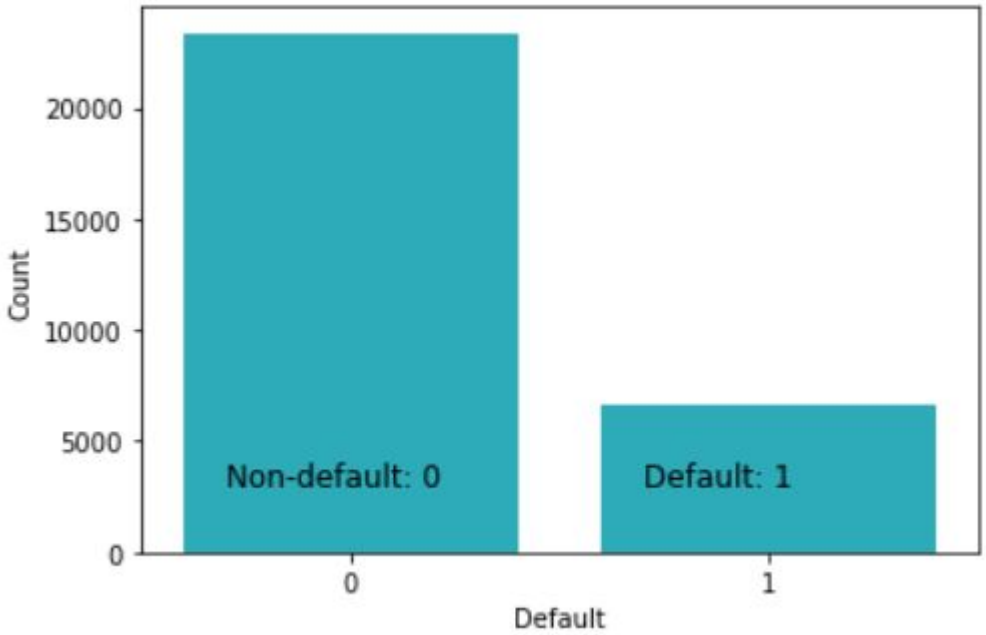


(1=Graduate school, 2=University, 3=High school, 0,4,5,6=others)

Credit card holders based on age

age	Number of users
29	1605
27	1477
28	1409
30	1395
26	1256
31	1217
25	1186
34	1162
32	1158
33	1146

Number of defaulters vs non-defaulters data present



This is an imbalanced dataset because the target column contains more number of observations for non-default cases and less number of observations for default cases.

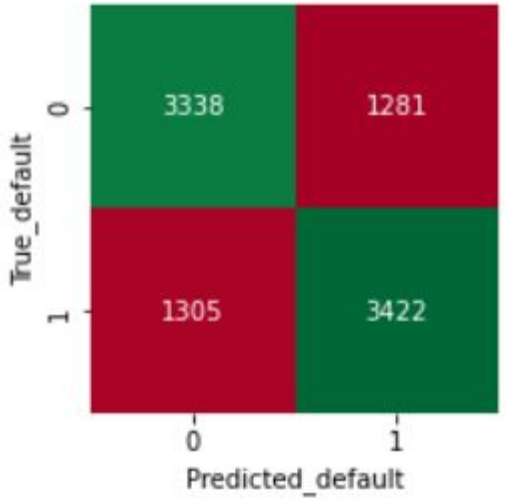
Data Preprocessing

- Drop the column ID
- Applied SMOTE (synthetic minority oversampling technique)
- Applied One- hot Encoding
- Applied StandardScaler
- Applied Train Test Split

Logistic Regression:

logistic regression classification report

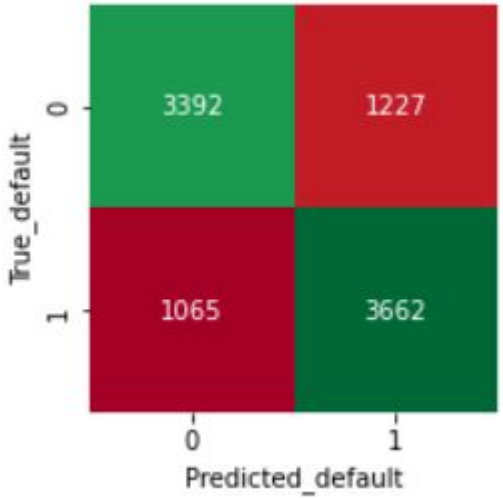
	precision	recall	f1-score	support
0	0.72	0.72	0.72	4619
1	0.73	0.72	0.73	4727
accuracy			0.72	9346
macro avg	0.72	0.72	0.72	9346
weighted avg	0.72	0.72	0.72	9346



k-nearest neighbors:

k nearest neighbor classification report

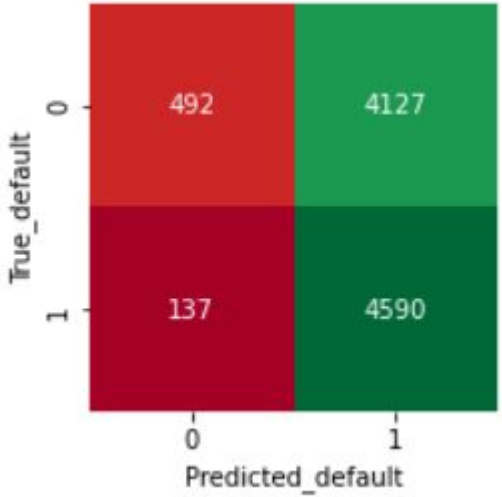
	precision	recall	f1-score	support
0	0.76	0.73	0.75	4619
1	0.75	0.77	0.76	4727
accuracy			0.75	9346
macro avg	0.76	0.75	0.75	9346
weighted avg	0.75	0.75	0.75	9346



Gaussian Naive Bayes:

Naive Bayes classification report

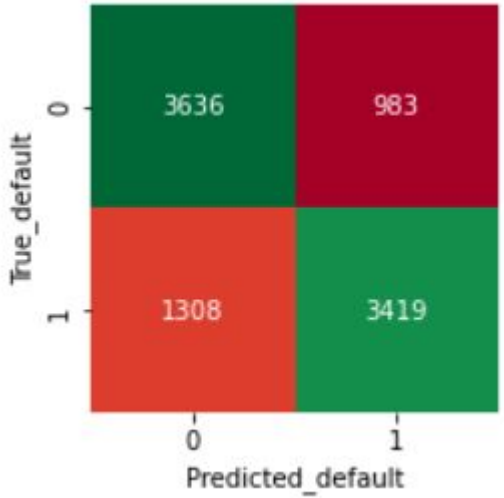
	precision	recall	f1-score	support
0	0.78	0.11	0.19	4619
1	0.53	0.97	0.68	4727
accuracy			0.54	9346
macro avg	0.65	0.54	0.44	9346
weighted avg	0.65	0.54	0.44	9346



Support Vectors Classifier:

Support Vector Machine classification report

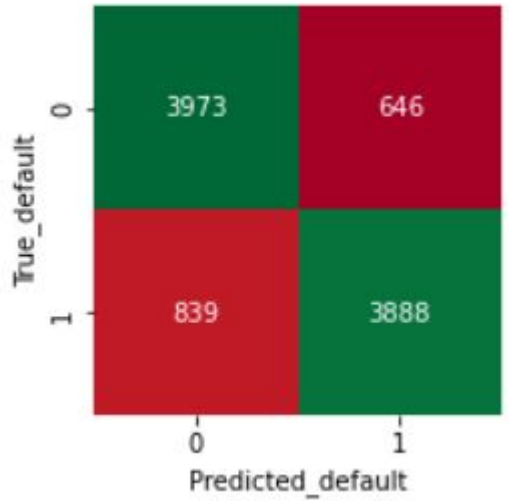
	precision	recall	f1-score	support
0	0.74	0.79	0.76	4619
1	0.78	0.72	0.75	4727
accuracy			0.75	9346
macro avg	0.76	0.76	0.75	9346
weighted avg	0.76	0.75	0.75	9346



Random Forest Classifier:

Random Forest classification report

	precision	recall	f1-score	support
0	0.83	0.86	0.84	4619
1	0.86	0.82	0.84	4727
accuracy			0.84	9346
macro avg	0.84	0.84	0.84	9346
weighted avg	0.84	0.84	0.84	9346



Model Performances

	Model Name	accuracy
0	logistic regression	0.723304
1	k nearest neighbor	0.754761
2	Naive Bayes	0.543762
3	Support Vector Machine	0.754868
4	Random Forest	0.841108


THANK
YOU