Machine Learning

Capstone Classification Project

Credit Card

Default Prediction

Fully Explained

BY

Diksha Shejao

CONTENT

1. Business Understanding

In today's world credit card have becomes a lifeline to a lot of people so banks provide us with credit card. 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".

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 to the credit card for a few consecutive months

Objective of our project is to predict which customer might default in upcoming months.

The research aims at developing a mechanism to predict the credit card default beforehand and to identity the potential customer base that can be offered various credit instruments so as to invite minimum default.

2. Data Summary

Unnamed: 0	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X15	X16	X17	X18	X19	X20
0	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	P
1	1	20000	2	2	1	24	2	2	-1	-1	...	0	0	0	0	689	0	
2	2	120000	2	2	2	26	-1	2	0	0	...	3272	3455	3261	0	1000	1000	
3	3	90000	2	2	2	34	0	0	0	0	...	14331	14948	15549	1518	1500	1000	
4	4	50000	2	2	1	37	0	0	0	0	...	28314	28959	29547	2000	2019	1200	
5	5	50000	1	2	1	57	-1	0	-1	0	...	20940	19146	19131	2000	36681	10000	
6	6	50000	1	1	2	37	0	0	0	0	...	19394	19619	20024	2500	1815	657	
7	7	500000	1	1	2	29	0	0	0	0	...	542653	483003	473944	55000	40000	38000	
8	8	100000	2	2	2	23	0	-1	-1	0	...	221	-159	567	380	601	0	

X20	X21	X22	X23	Y
PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
0	0	0	0	1
1000	1000	0	2000	1
1000	1000	1000	5000	0
1200	1100	1069	1000	0
10000	9000	689	679	0
657	1000	1000	800	0
38000	20239	13750	13770	0
0	581	1687	1542	0

1. Feature 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 payment 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

Y-Default payment

Insights From Our Dataset

1. This dataset is from Taiwan.

2. In our dataset there are 30001 rows, 25 columns.

3. There are No missing values present

4. There are No duplicate values present

5. There are No null values

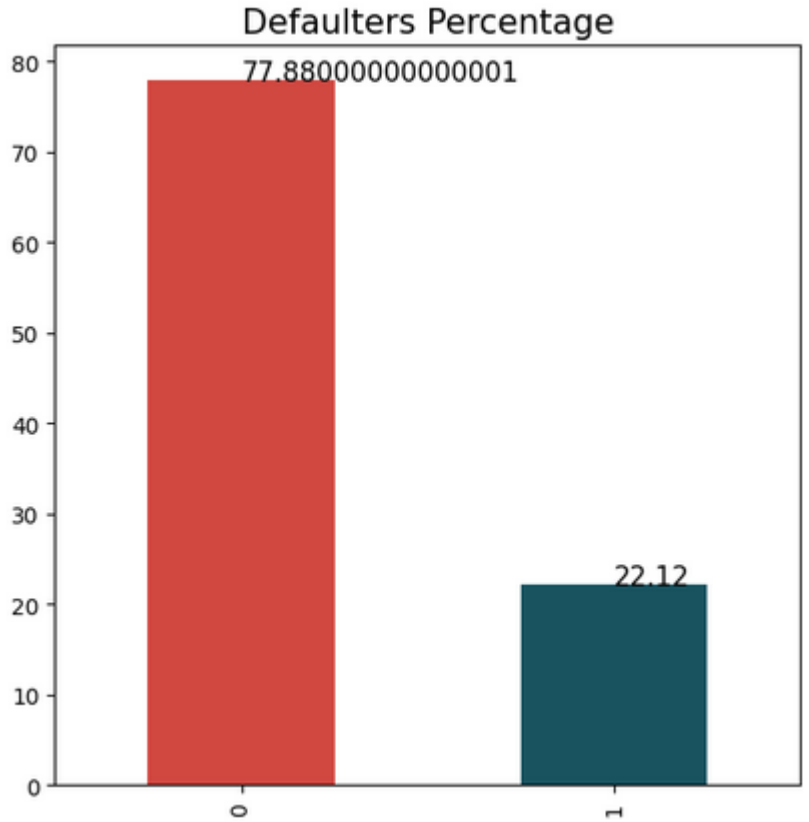
6. And finally we have ‘default payment next month’ variable which we need to predict for new observations.

7. Nine categorical variables present.

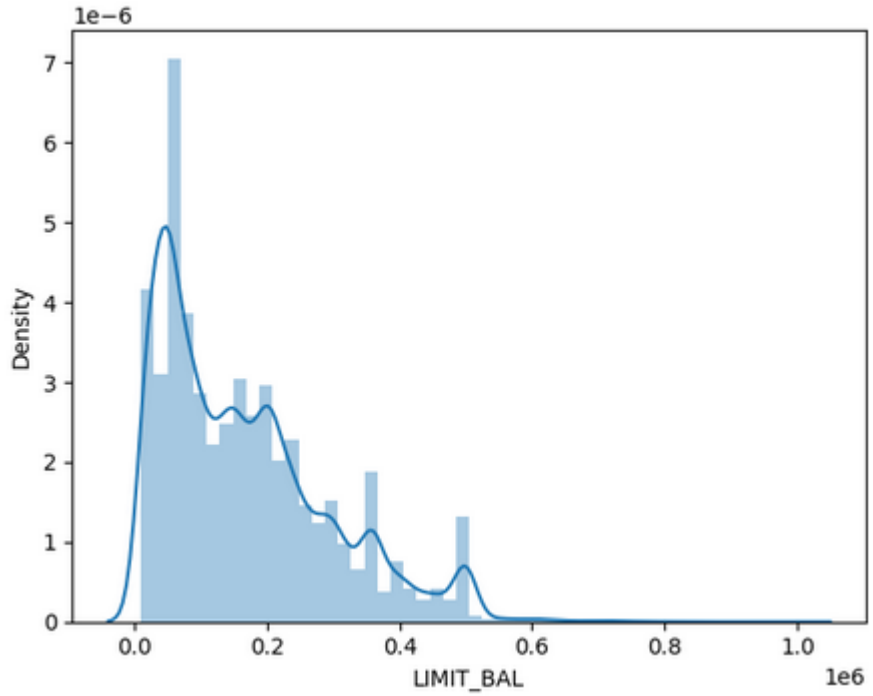
8. Six months payment and bill data available.

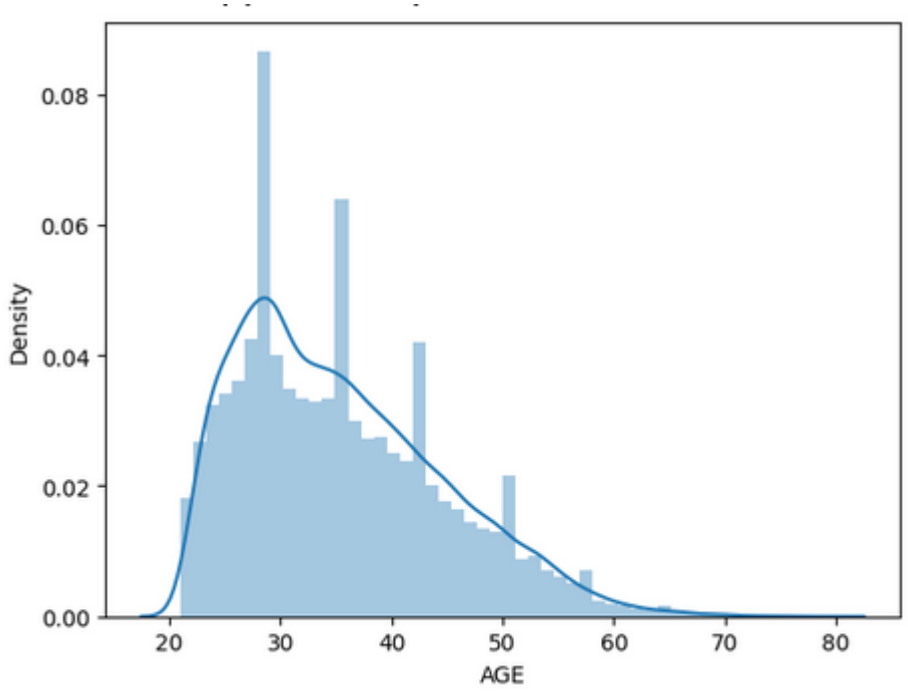
9. The columns are :- ‘ID’, ‘LIMIT_BAL’, ‘EDUCATION’, ‘MARRIAGE’, ‘AGE’, ‘PAY_0’, ‘PAY_2’, ‘PAY_3’, ‘PAY_4’, ‘PAY_5’, ‘PAY_6’, ‘BILL_AMT1’, ‘BILL_AMT2’, ‘BILL_AMT3’, ‘BILL_AMT4’, ‘BILL_AMT5’, ‘BILL_AMT6’, ‘PAY_AMT1’, ‘PAY_AMT2’, ‘PAY_AMT3’, ‘PAY_AMT3’, ‘PAY_AMT4’, ‘PAY_AMT5’, ‘PAY_AMT6’, ‘DEFAULTER’, ‘AGE_BIN’

4.Exploratory data analysis

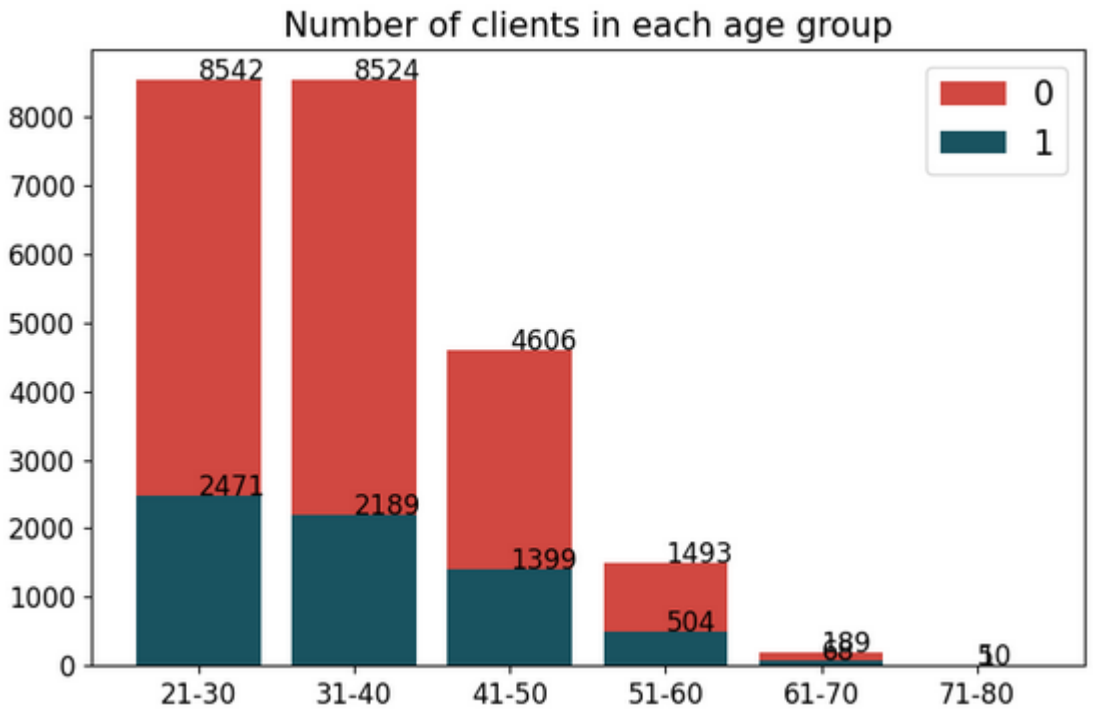


so we have 22% defaulters in our dataset and 77% persons are non defaulters



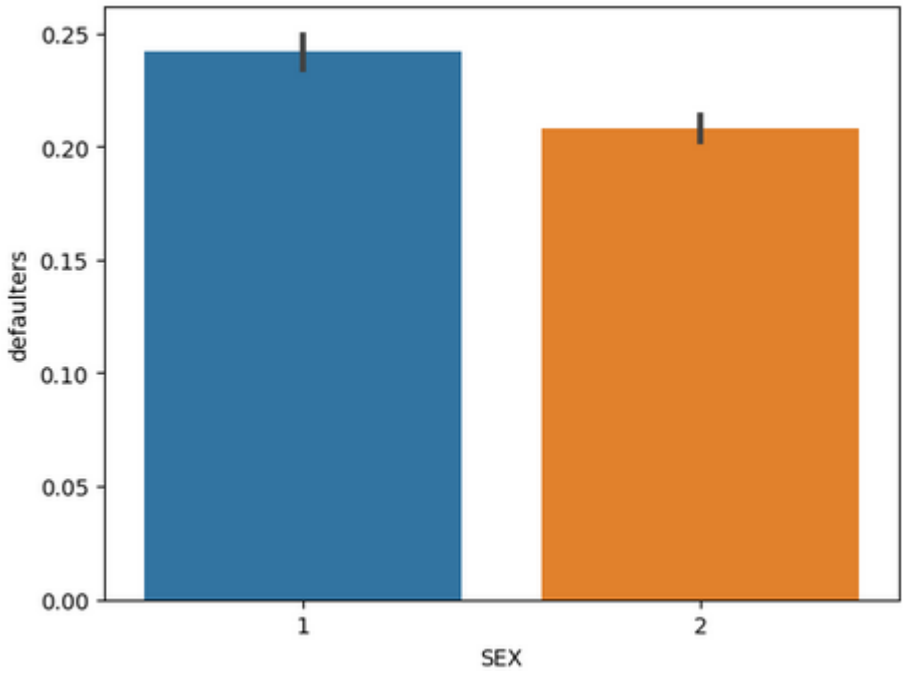


The data shows that most people are of age range 20-40 and a few only from 50-60 group



We have maximum clients from 21-30 age group followed by 31-40.

Hence with increasing age group the number of clients that will default the payment next month is decreasing



So we have more male defaulters .

5.Model Building

Classifiers

Random Forest

KNN

Implementing SMOT

XG BOOST

Classifiers

Random FOREST

Testing Accuracy

Accuracy score 0.82

roc_auc_scor 0.66

The recall on test data is 0.36

```
print(classification_report(pred, y_test))
print(confusion_matrix( y_test, pred))
```

```
              precision    recall  f1-score   support

           0       0.94      0.84      0.89      5248
           1       0.37      0.65      0.47       752

    accuracy                           0.82      6000
   macro avg       0.66      0.75      0.68      6000
weighted avg       0.87      0.82      0.84      6000

[[4413  260]
 [ 835  492]]
```

KNN

```
Accuracy score 0.7716666666666666
Classification report
              precision    recall  f1-score   support

           0       0.80      0.95      0.87      4673
           1       0.45      0.15      0.22      1327

    accuracy                           0.77      6000
   macro avg       0.62      0.55      0.54      6000
weighted avg       0.72      0.77      0.72      6000
```

The roc_auc_score 0.54

The recall score is 0.14

Large unbalanced dataset,so KNN is giving low accuracy

```
Implementing SMOT
```

roc_auc_score 0.87

```
Random Forest with SMOT
```

accuracy_score 0.87

recall score 0.83

KNN with SMOT

roc_auc_score 0.80

recall_score 0.8065482559383693

XG BOOST

ruc_auc_score 0.65

recall score 0.37

XG BOOST with SMOT

ruc_auc_score 0.81

recall score 0.79

Conclusion

In this project , we worked on credible or not credible clients.

The dataset contained about 30001 rows, 25 columns.

We began by dealing with the dataset's missing values and doing EDA.

Using the model we built above, it can predicts whether a customer is likely to default on their credit card payment based on their credit card usage history. This can benefits us in multiple ways:

- Risk Assesement
- Portfolio Management
- Marketing Strategy
- Collection Management
- Credit Scoring

Overall, the output of the machine learning model can help us make more informed decisions about credit card lending and management, ultimately leading to better business outcomes