Capstone Project

(SUPERVISED ML – CLASSIFICATION)

Credit Card Default Prediction

TEAM

- IQBAL BABWANE
- SAMEER ANSARI
- LUKMAN HAIDER KHAN

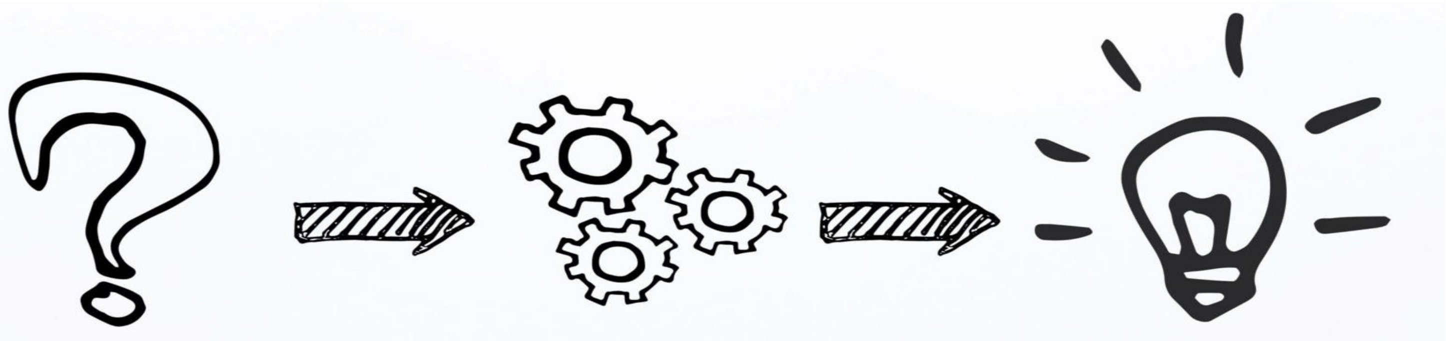
~ UNDER THE GUIDANCE OF TEAM ALMABETTER

CONTENT

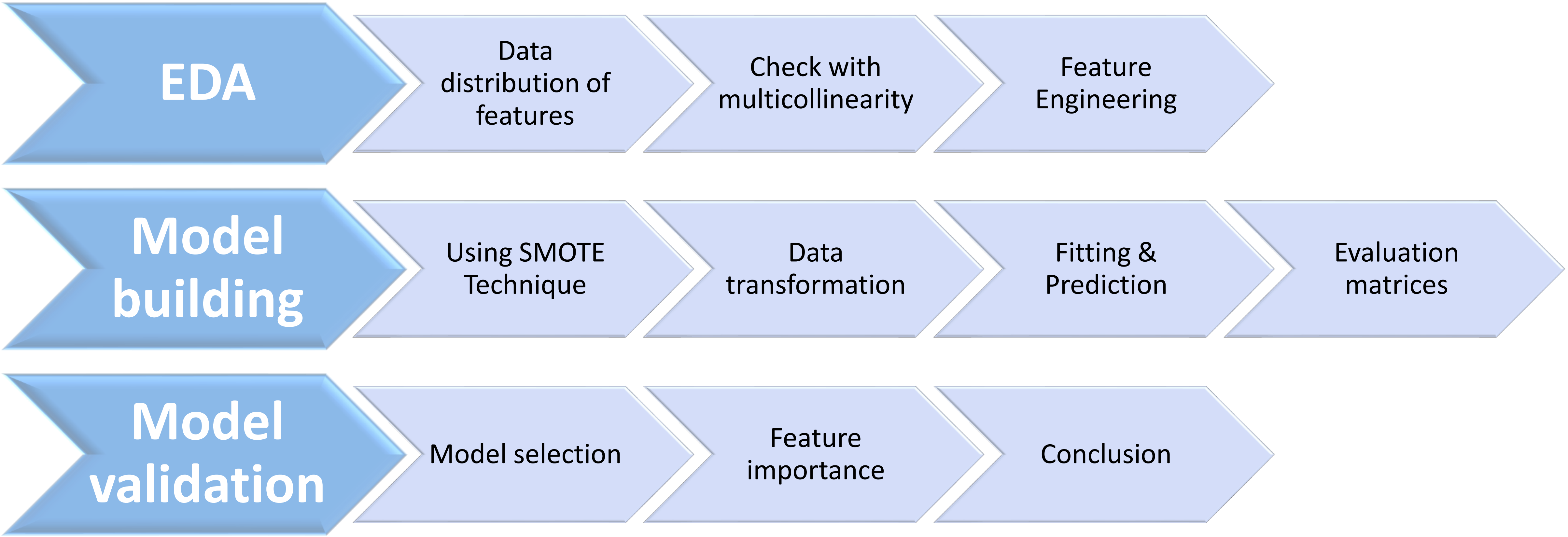
- PROBLEM STATEMENT
- METHODOLOGY
- INTRODUCTION OF PROJECT
- DATA DESCRIPTION
- EDA
- FITTING VARIOUS MODEL
- MODEL PERFORMANCE COMPARISION
- MODEL VALIDATION
- MODEL EXPLAINABILITY
- CONCLUSION

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.



METHODOLOGY



EDA

Data distribution of features

Check with multicollinearity

Feature Engineering

Model building

Using SMOTE Technique

Data transformation

Fitting & Prediction

Evaluation matrices

Model validation

Model selection

Feature importance

Conclusion

INTRODUCTION

The basic idea of this capstone project is to use the Supervised Machine Learning - Classification to predict customers default payments in Taiwan. Here we have previous 6 month transaction bills and statements as our major information to classify defaulter.

Based on these features we will be predicting our target variable i.e. credit card defaulters. By using concepts like model validation, we will came to know which features are important and how much they contribute to our target variable.

DATA DESCRIPTION

- ***ID: ID of each client***
- ***LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)***
- ***SEX: Gender (1 = male, 2 = female)***
- ***EDUCATION: (1 = graduate school, 2 = university, 3 = high school, 0,4,5,6 = others)***
- ***MARRIAGE: Marital status (1 = married, 2 = single, 3 = others)***
- ***AGE: Age in years***
- **Scale for PAY_0 to PAY_6 :**

  (-2 = No consumption, -1 = paid in full, 0 = use of revolving credit (paid minimum only),

  1 = payment delay for one month, 2 = payment delay for two months,

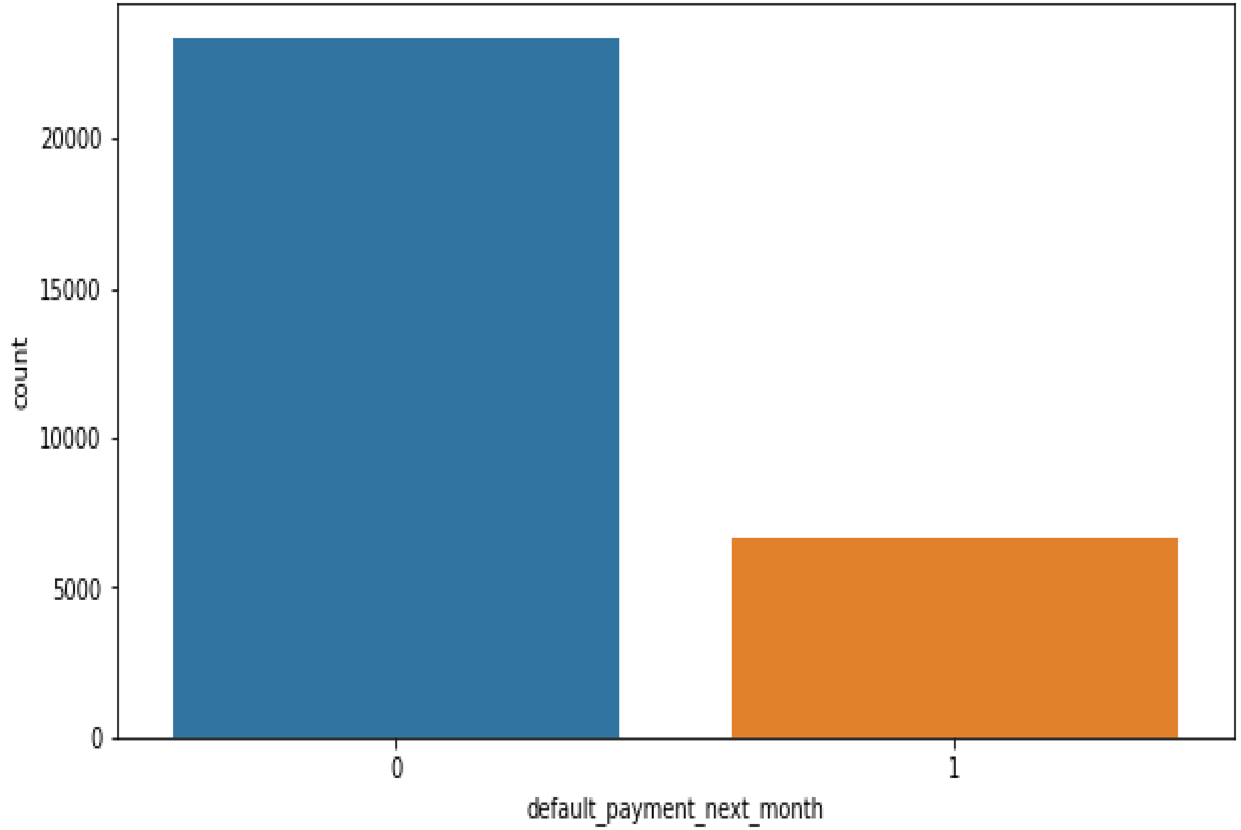
  ***... 8 = payment delay for eight months, 9 = payment delay for nine months and above*)**
- ***PAY_0 to PAY_6: Repayment status in (September, 2005), (August, 2005).....(April, 2005)***
- ***BILL_AMT1 to BILL_AMT6: Amount of bill statement in (September, 2005), (August, 2005).....(April, 2005)***
- ***PAY_AMT1 to PAY_AMT6: Amount of previous payment in (September, 2005), (August, 2005).....(April, 2005)***
- ***Default payment next month: Default payment (1=yes, 0=no)***

EDA

Data distribution of target variable

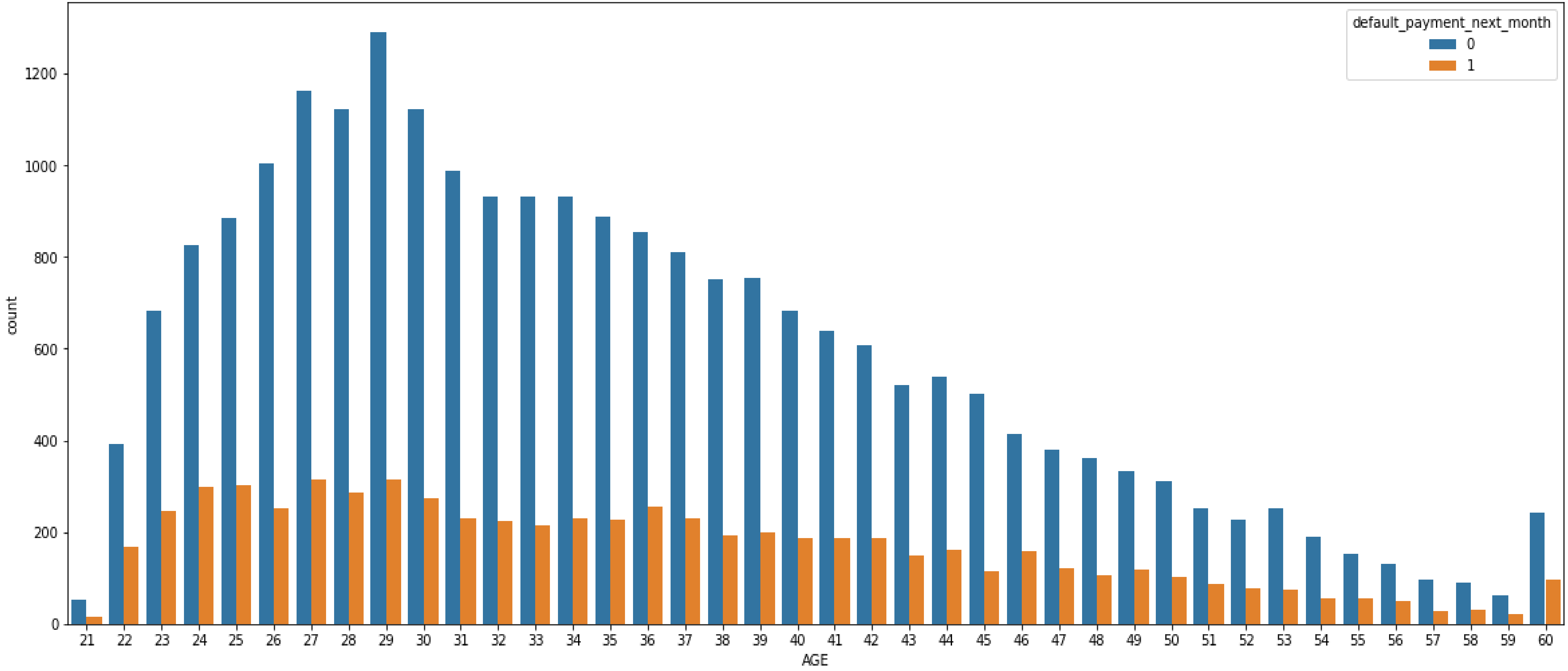
- ❑ Non-Defaulter(0) -**0.7788**
- ❑ Defaulter(1) - **0.2212**

22% of customers has default payment next month



EDA

Analysis on AGE feature



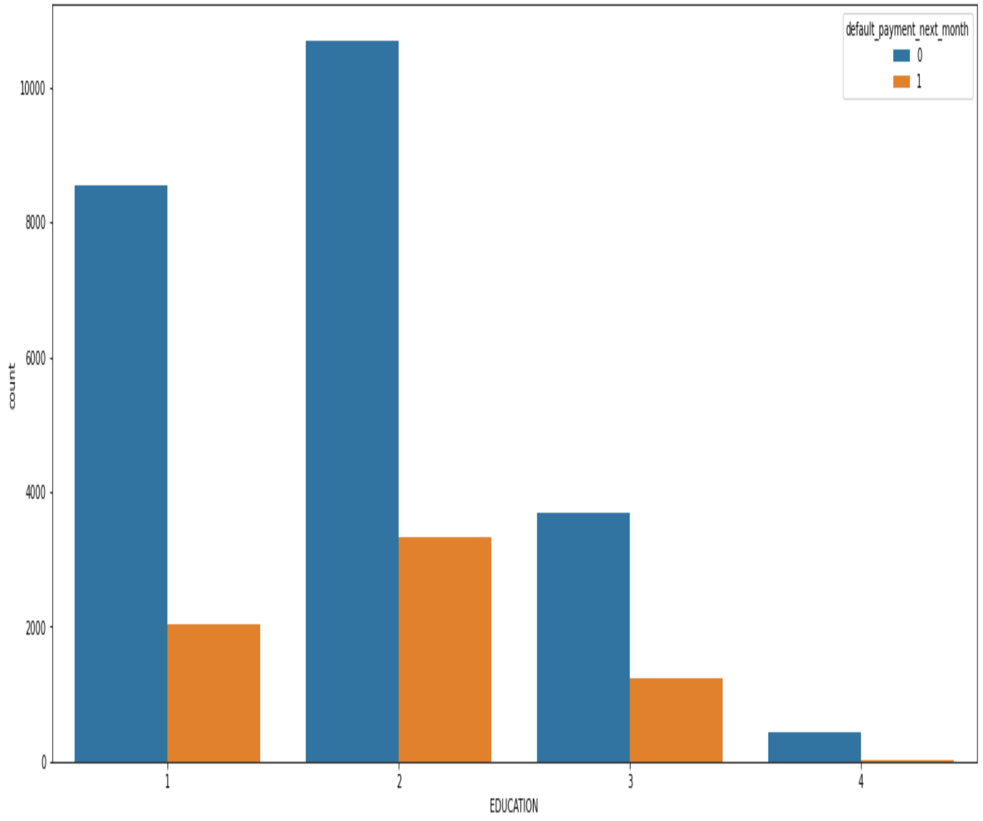
- ❑ 20 to 40 years customer are on average for defaulters
- ❑ Age above 60 years are almost defaulters

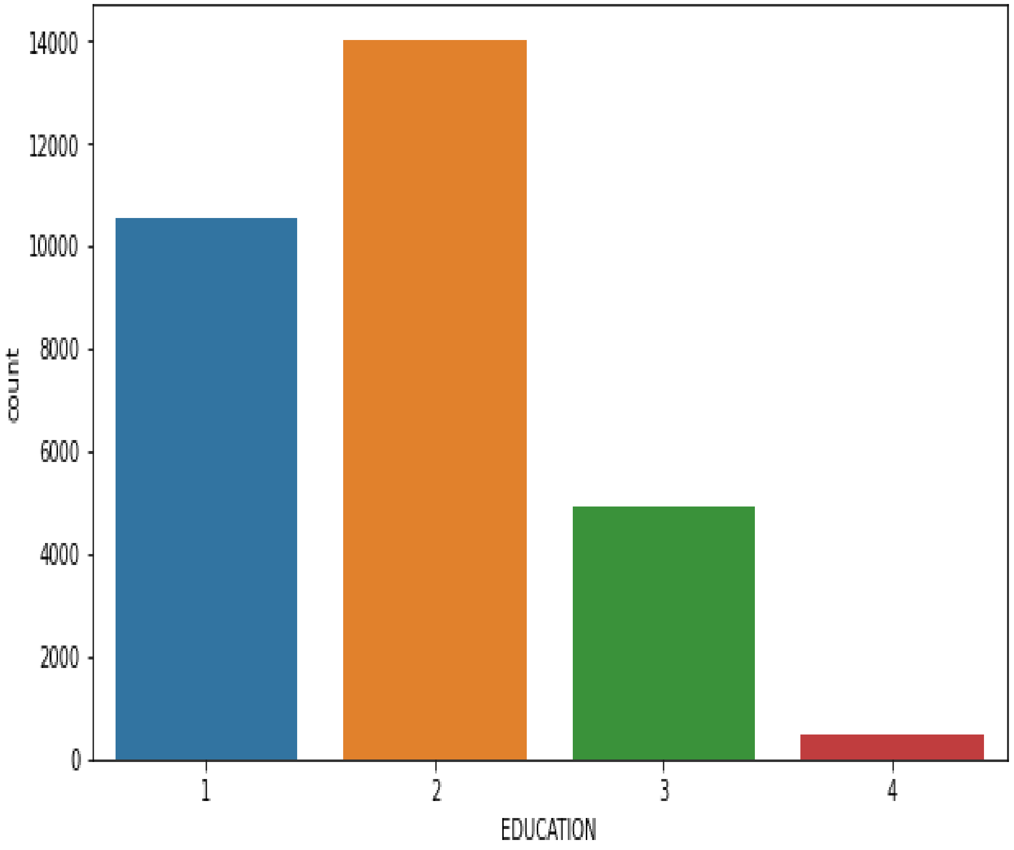
EDA

Analysis on Education feature

EDUCATION

1. Graduate School
2. University
3. Highschool
4. Others





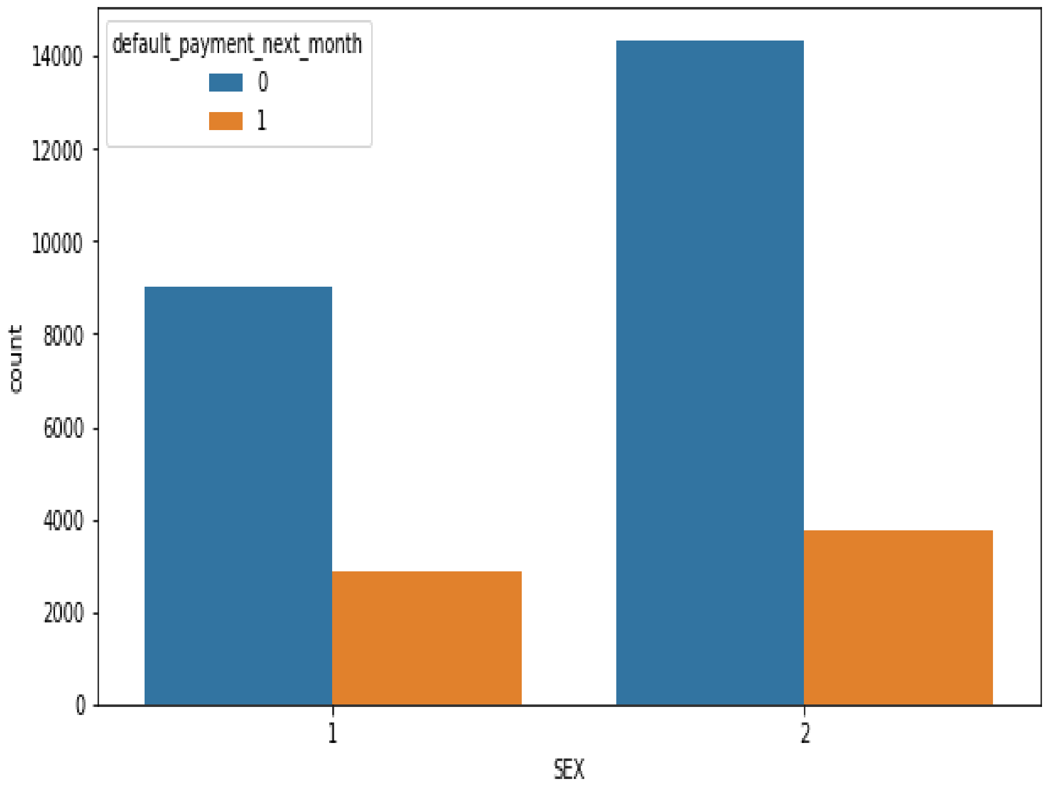
□ Customer which had education at University level has more user as well as defaulters

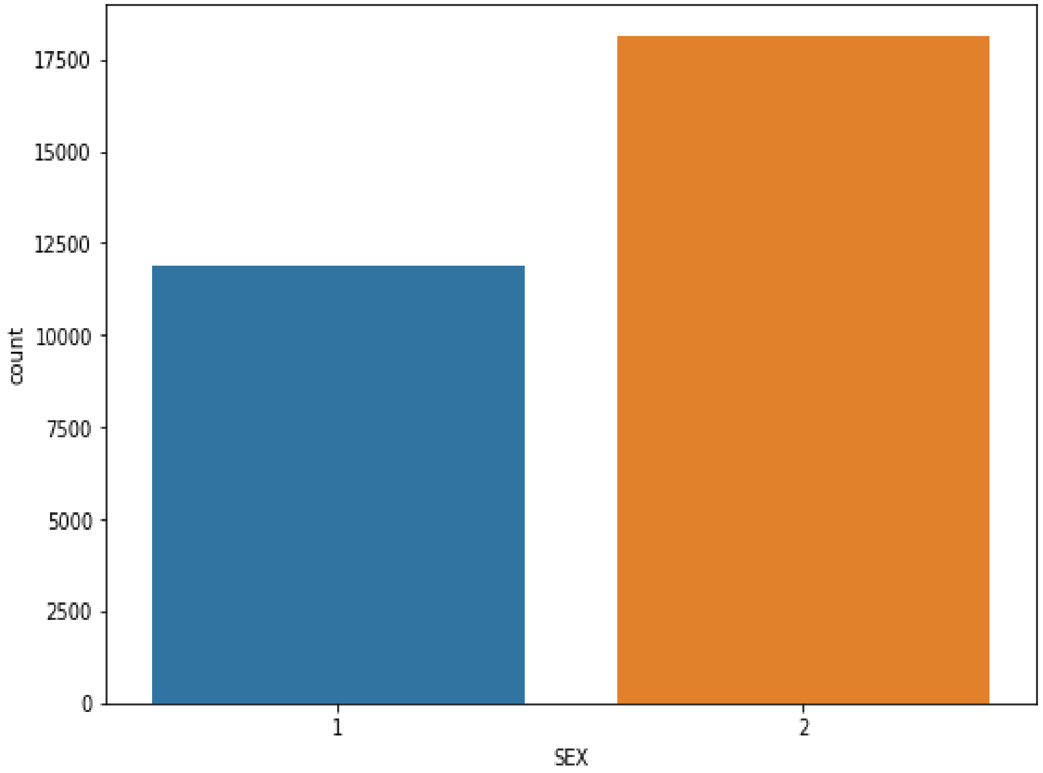
EDA

Analysis on SEX feature

SEX

1. Male
2. Female





❑ Female customer count are more as compared to male

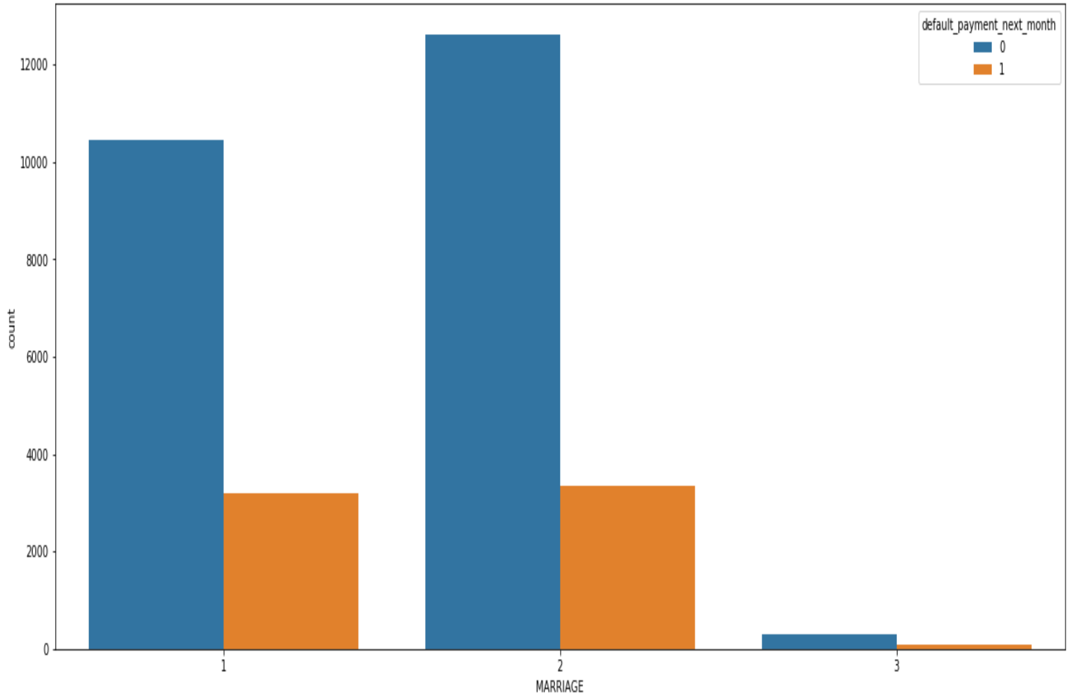
❑ From above graph it is clear that female customers are more defaulters

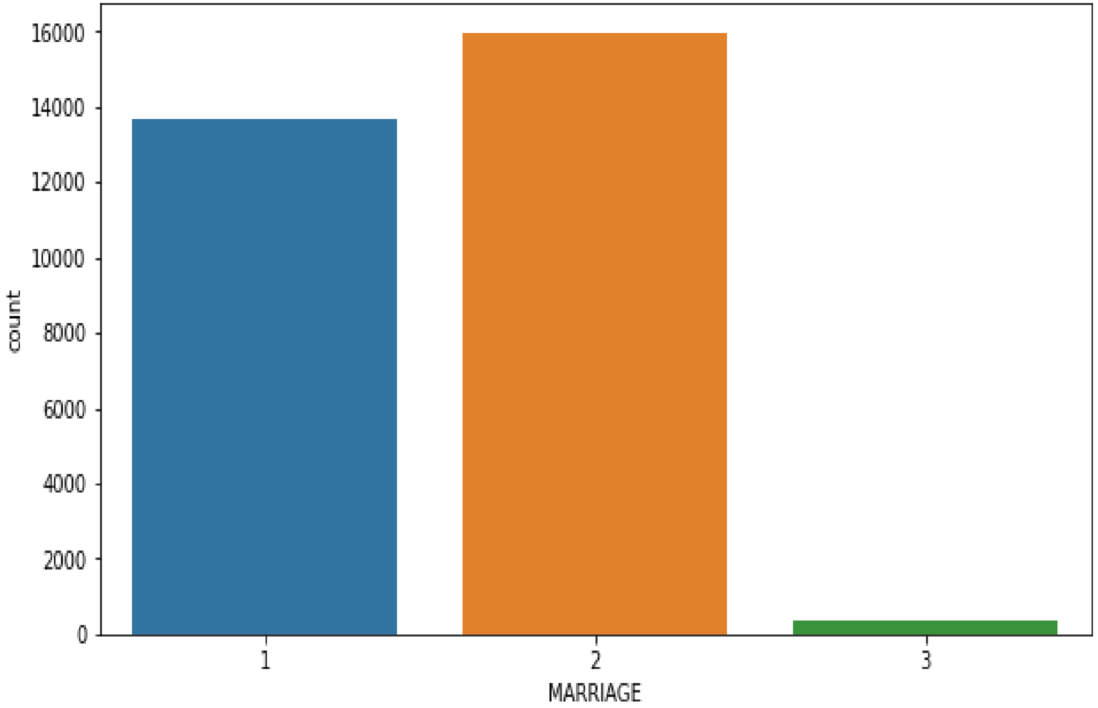
EDA

Analysis on MARRIAGE feature

MARRIAGE

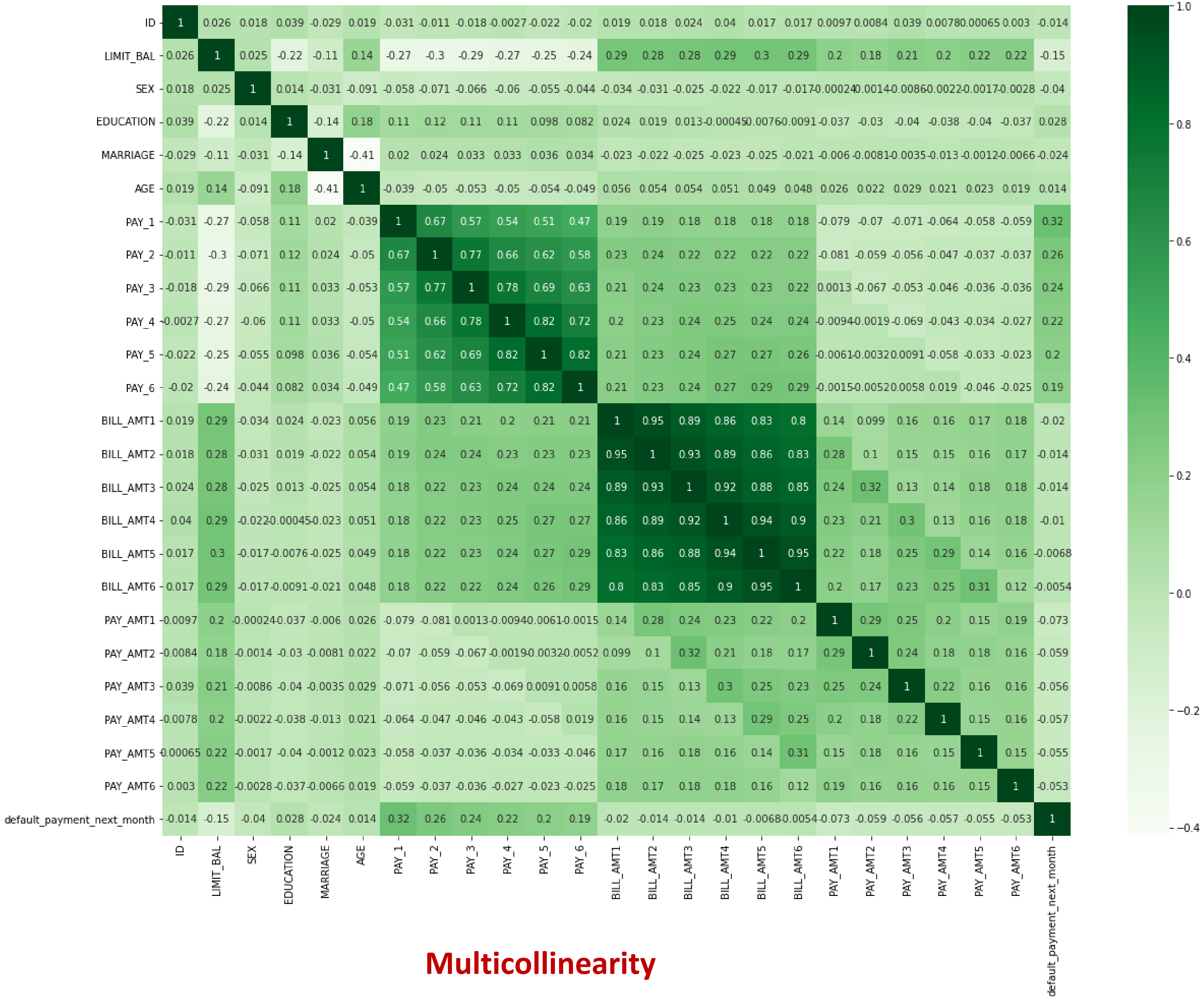
1. Married
2. Single
3. Others





❑ Married customer count is greater of all

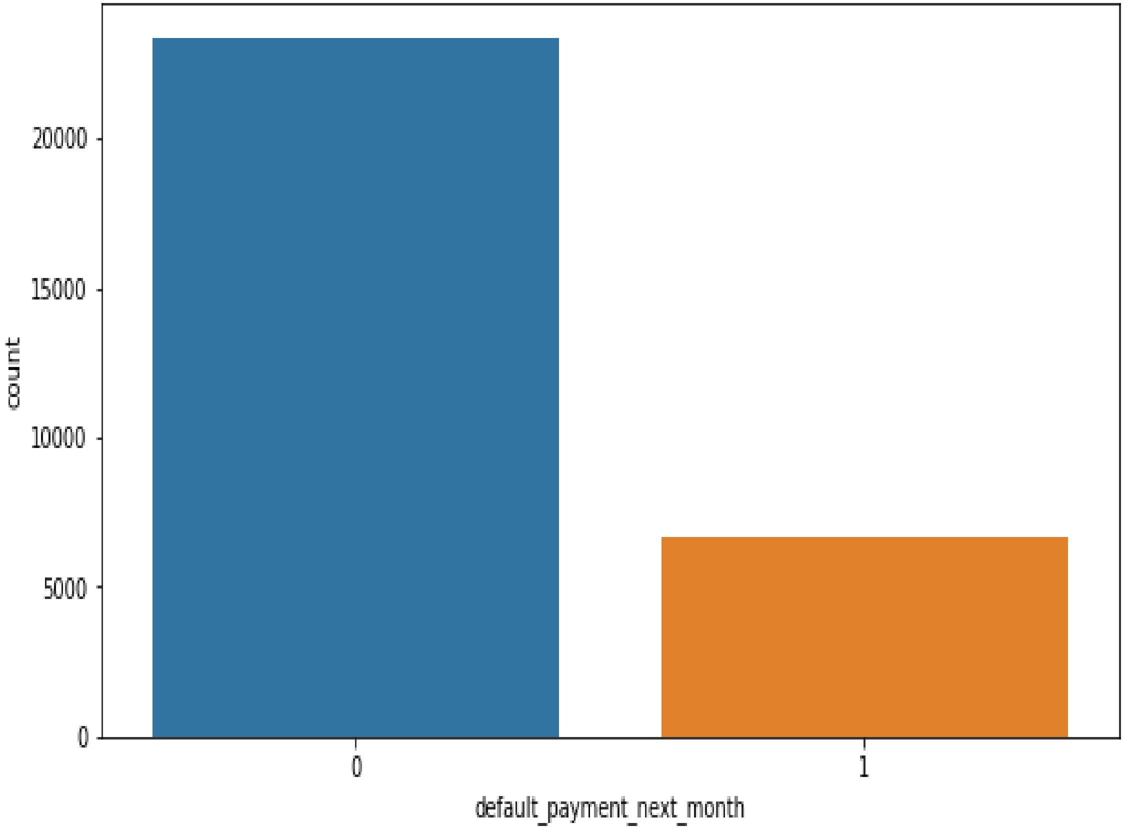
❑ Married and single defaulter customers does not have much difference but, married customers takes lead for defaulters



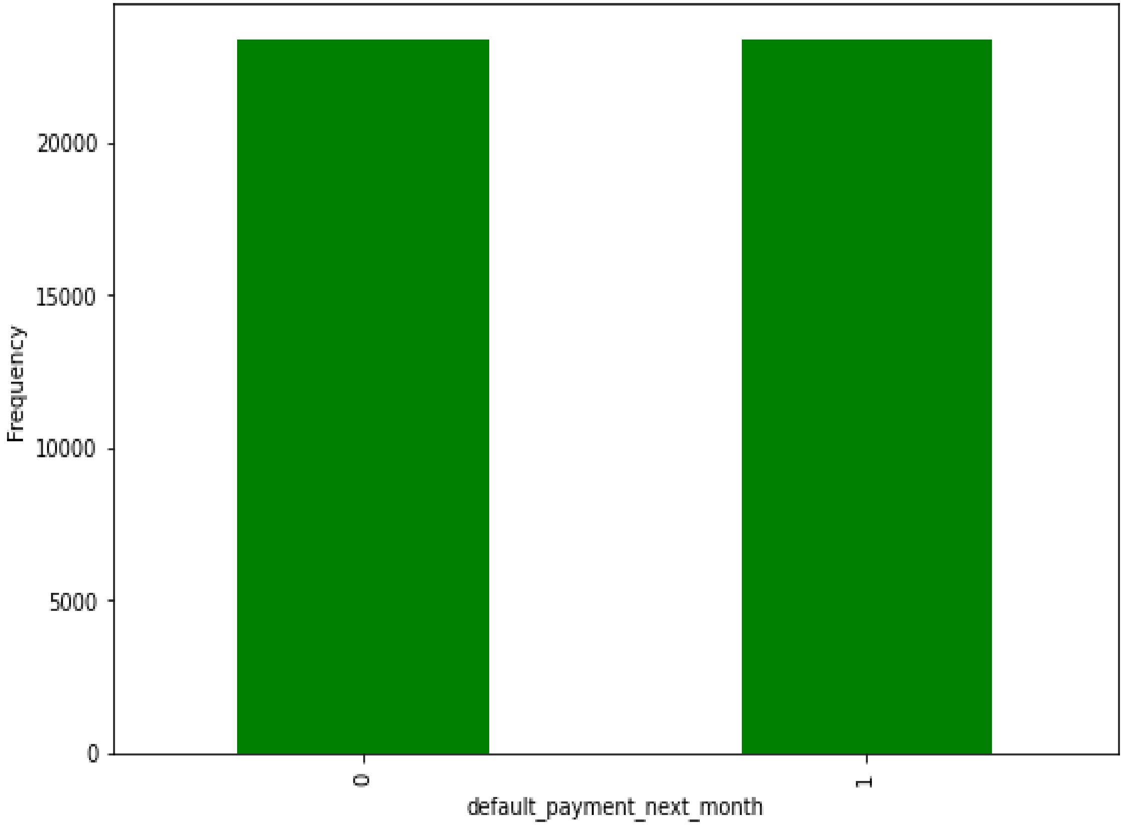
Multicollinearity

- □ Here most of the categories have correlated with each other, because all those are previous transaction of customer

SMOTE - Synthetic Minority Oversampling Technique



Target Variable before SMOTE



Target Variable after SMOTE

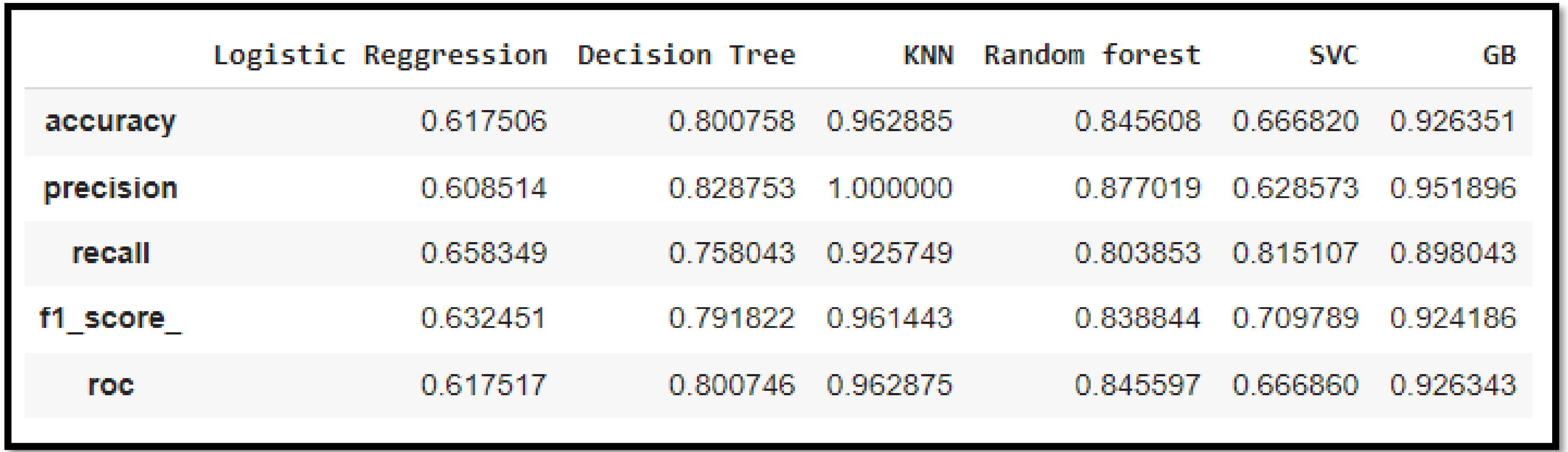
FITTING VARIOUS MODEL

1. Logistic Regression
2. Decision Tree Classifier
3. K-Nearest Neighbors Classifier
4. Random forest Classifier
5. Support Vector Classifier
6. Gradient Boosting Classifier

MODEL PERFORMANCE COMPARISION

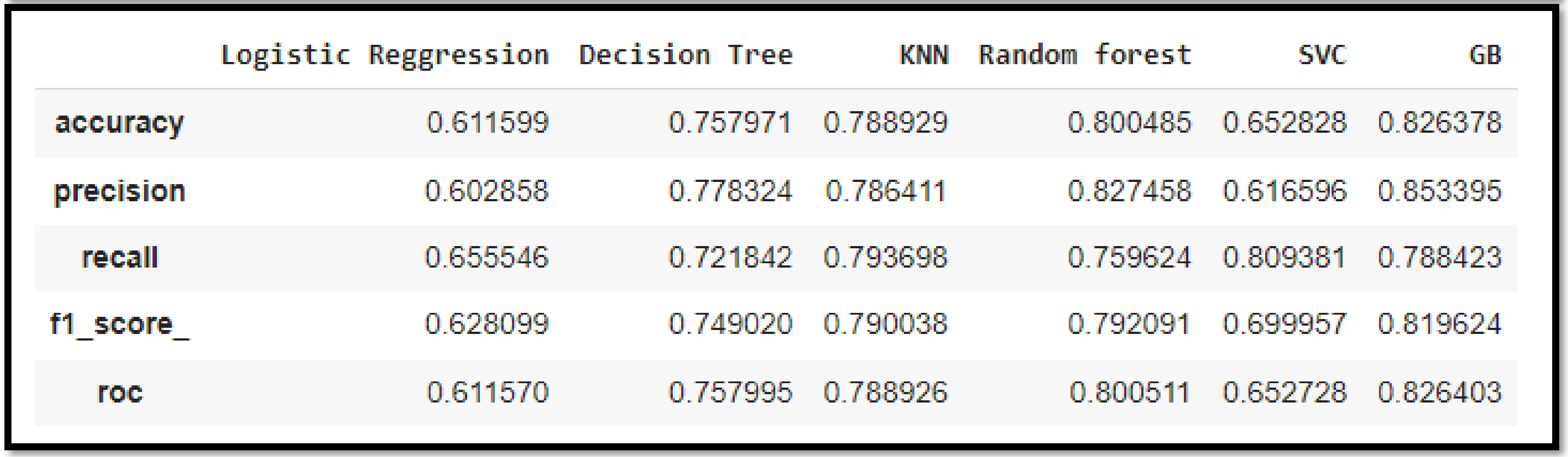
Evaluation matrices for all the models

Training Scores



	Logistic Reggression	Decision Tree	KNN	Random forest	SVC	GB
accuracy	0.617506	0.800758	0.962885	0.845608	0.666820	0.926351
precision	0.608514	0.828753	1.000000	0.877019	0.628573	0.951896
recall	0.658349	0.758043	0.925749	0.803853	0.815107	0.898043
f1_score_	0.632451	0.791822	0.961443	0.838844	0.709789	0.924186
roc	0.617517	0.800746	0.962875	0.845597	0.666860	0.926343

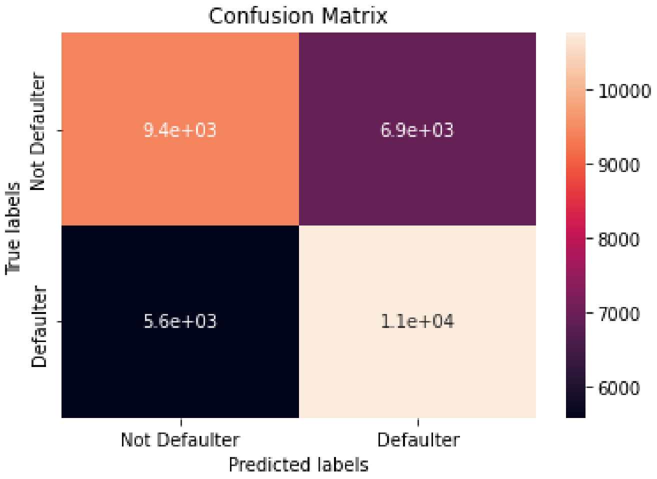
Testing Scores



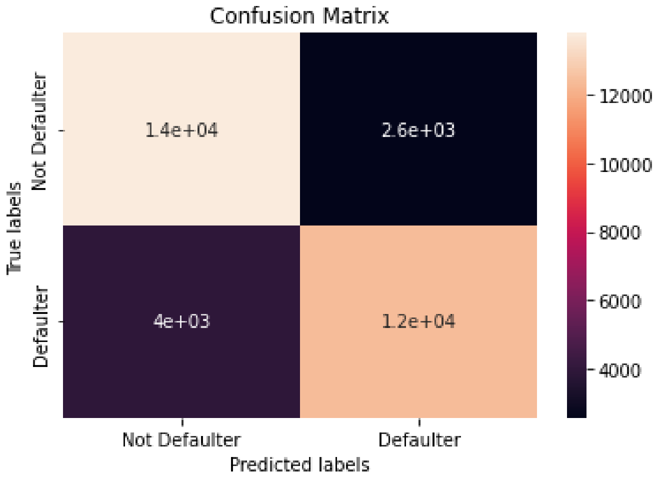
	Logistic Reggression	Decision Tree	KNN	Random forest	SVC	GB
accuracy	0.611599	0.757971	0.788929	0.800485	0.652828	0.826378
precision	0.602858	0.778324	0.786411	0.827458	0.616596	0.853395
recall	0.655546	0.721842	0.793698	0.759624	0.809381	0.788423
f1_score_	0.628099	0.749020	0.790038	0.792091	0.699957	0.819624
roc	0.611570	0.757995	0.788926	0.800511	0.652728	0.826403

Confusion matrices of all training model

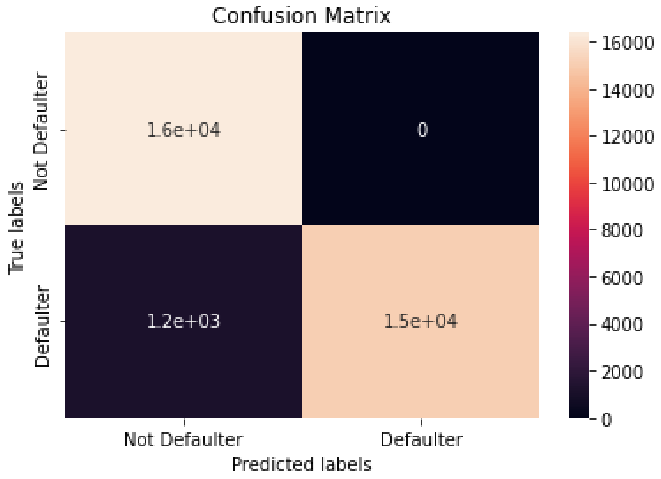
LR



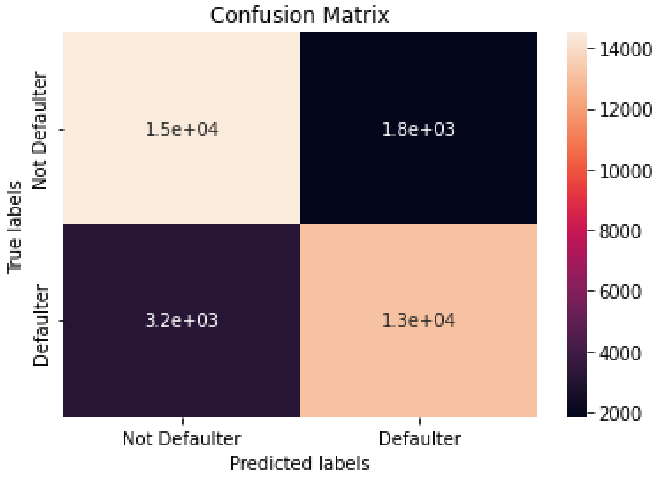
DT



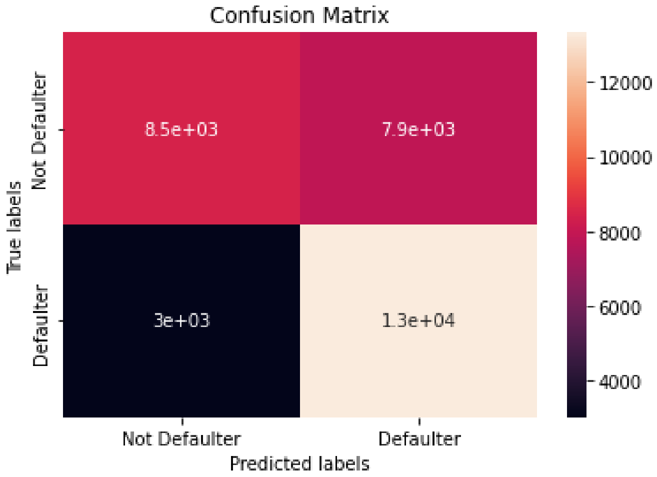
KNN



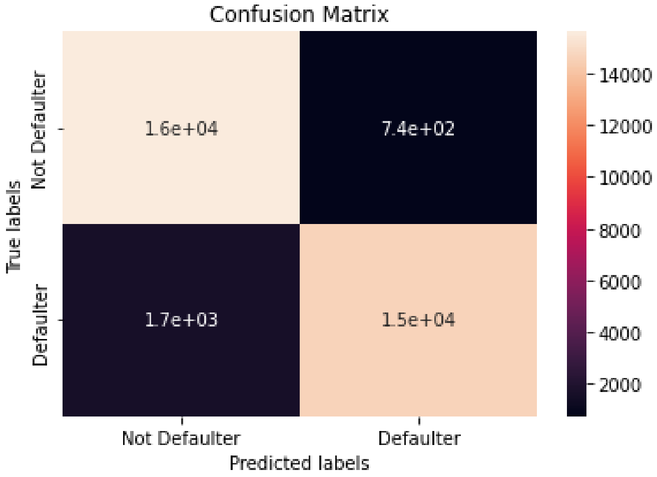
RF



SVC



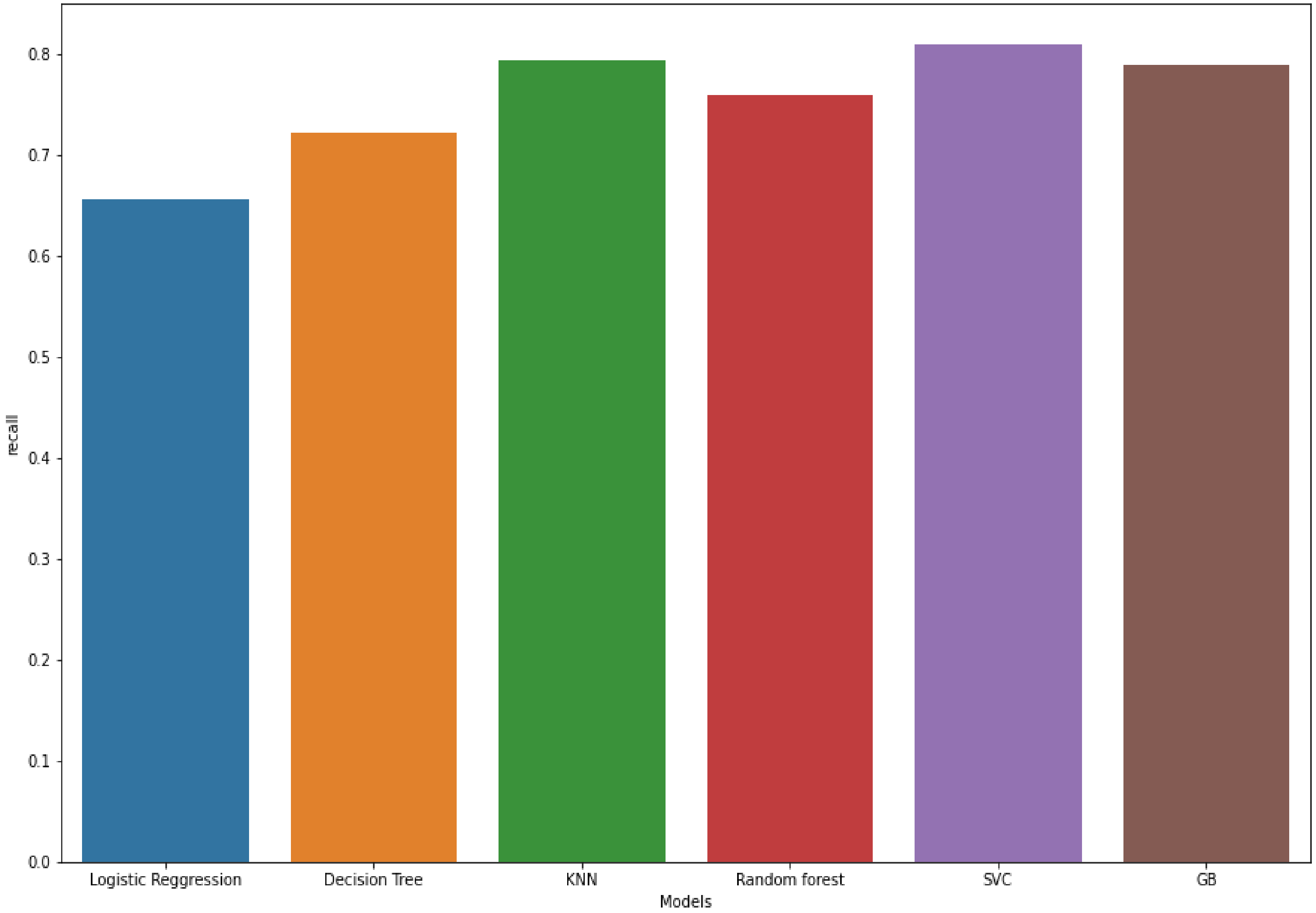
GB



Recall score for all the corresponding models

- ❑ In this classification problem there is a high cost for the bank when a default credit card is predicted as non-default, since no actions can be taken. Thus, we will give **recall** more importance .

- ❑ Here we find **SVC Model** as best performer in our case

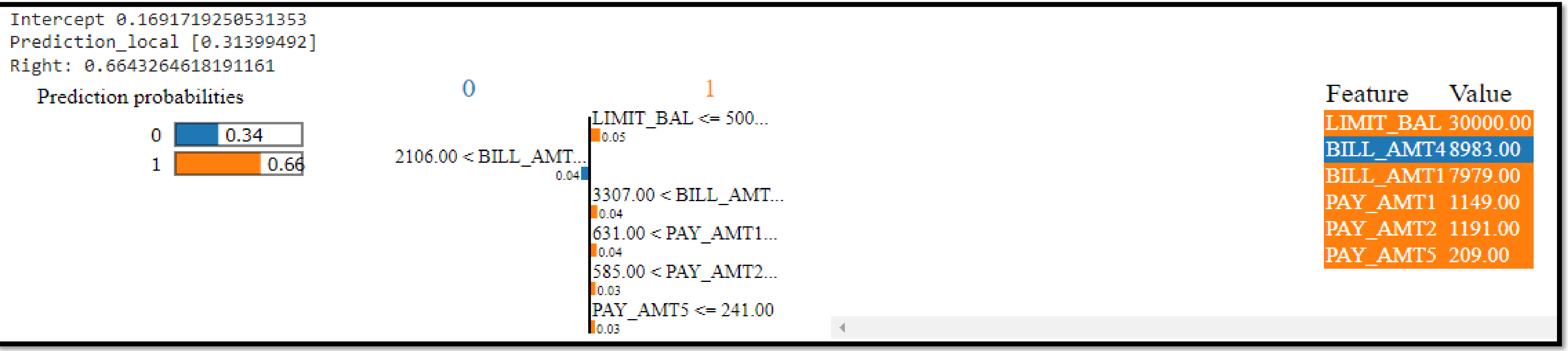


MODEL VALIDATION

- By observing Evaluation matrices for all the models-
  - ❑Logistic Regression model scores very bad as compared to others
  - ❑Decision Trees, KNN, Random Forest and Gradient boosting are quite good with linear models and gives better accuracy, but looking at the scores this models are over fitting and we can’t conclude w.r.t. accuracy.
  - ❑SVC Model are not so good with accuracy, but they are best with its recall score and that is what we wanted, so we will go with it.

MODEL EXPLAINABILITY

1. Using LIME



```
Intercept 0.16917192505313353
Prediction_local [0.31399492]
Right: 0.6643264618191161
```

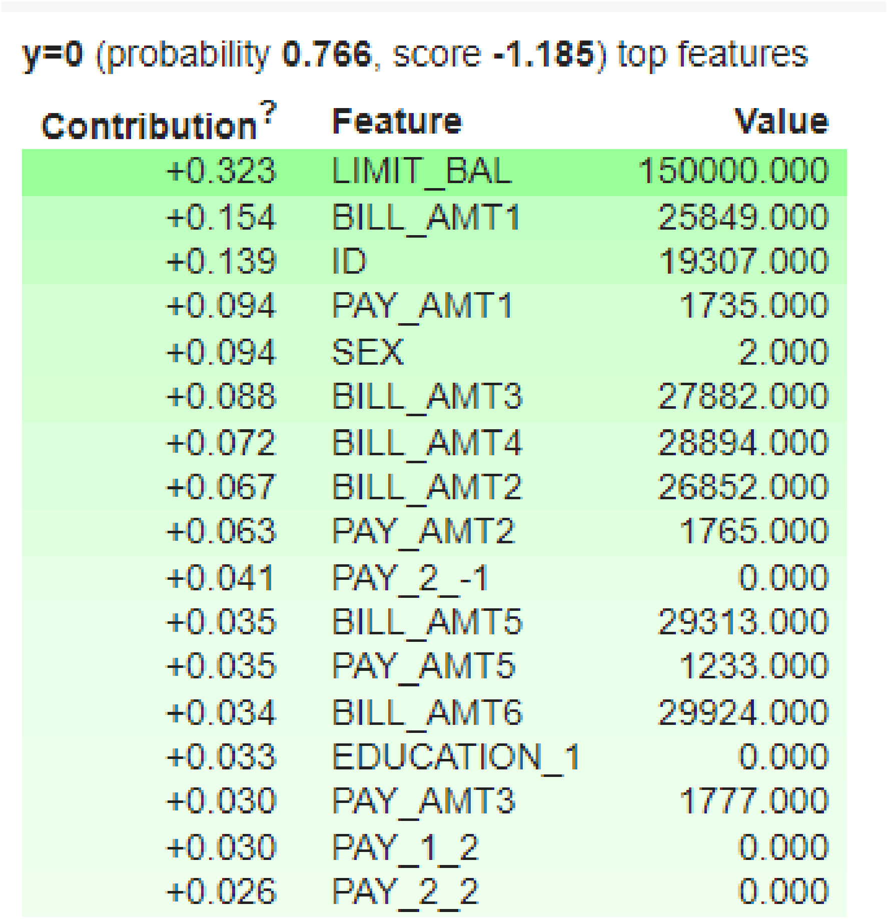
Feature	Value
LIMIT_BAL	30000.00
BILL_AMT4	8983.00
BILL_AMT1	7979.00
PAY_AMT1	1149.00
PAY_AMT2	1191.00
PAY_AMT5	209.00

For a data point in SVC, we got-

- ❑ Non -Defaulter probability – 0.34
- ❑ Defaulter probability – 0.66

2. Using ELI5

❑ By using ELI5, for a particular data point, we got some of the best feature with its corresponding values

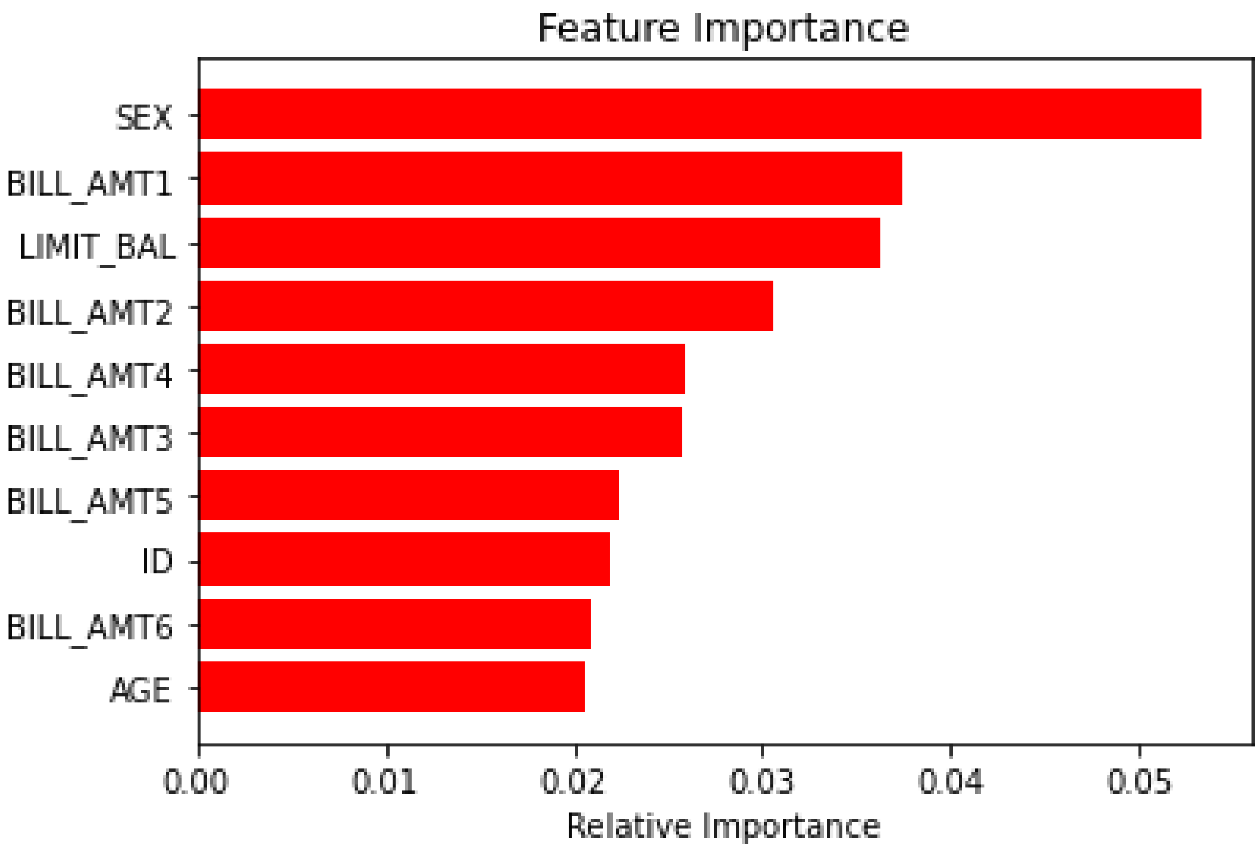


y=0 (probability **0.766**, score **-1.185**) top features

Contribution?	Feature	Value
+0.323	LIMIT_BAL	150000.000
+0.154	BILL_AMT1	25849.000
+0.139	ID	19307.000
+0.094	PAY_AMT1	1735.000
+0.094	SEX	2.000
+0.088	BILL_AMT3	27882.000
+0.072	BILL_AMT4	28894.000
+0.067	BILL_AMT2	26852.000
+0.063	PAY_AMT2	1765.000
+0.041	PAY_2_-1	0.000
+0.035	BILL_AMT5	29313.000
+0.035	PAY_AMT5	1233.000
+0.034	BILL_AMT6	29924.000
+0.033	EDUCATION_1	0.000
+0.030	PAY_AMT3	1777.000
+0.030	PAY_1_2	0.000
+0.026	PAY_2_2	0.000

Top features which helping to make our prediction

- ❑ Apart from Gender of Customer, transaction of last months and limit balance are the top features for prediction



CONCLUSION

❑From the previous slides we got some evident that SVC will perform better among all the models for the Credit Card Default Prediction, since the recall score was best for this model.

❑Limit balance and previous last months bills are the features contributes heavily to predict our target variable.

THANK
YOU