CAPSTONE PROJECT - 3

CREDIT CARD DEFAULT PREDICTION

(Supervised Machine Learning Classification)

By

VIKASKUMAR SHARMA

(Cohort Tosh)

❑Problem Statement

- **In today's world credit cards have become a lifeline to a lot of people so banks provide us with credit cards. 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 on the credit card for a few consecutive months**
- **The research aims at developing a mechanism to predict the credit card default beforehand and to identify the potential customer base that can be offered various credit instruments so as to invite minimum default.**
- **Objective of our project is to predict which customer might default in upcoming months.**

DATA SUMMARY

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6
0	1	20000	2	2	1	24	2	2	-1	-1	-2	-2
1	2	120000	2	2	2	26	-1	2	0	0	0	2
2	3	90000	2	2	2	34	0	0	0	0	0	0
3	4	50000	2	2	1	37	0	0	0	0	0	0
4	5	50000	1	2	1	57	-1	0	-1	0	0	0

BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
3913	3102	689	0	0	0	0	689	0	0	0	0	1
2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0

FEATURE SUMMARY

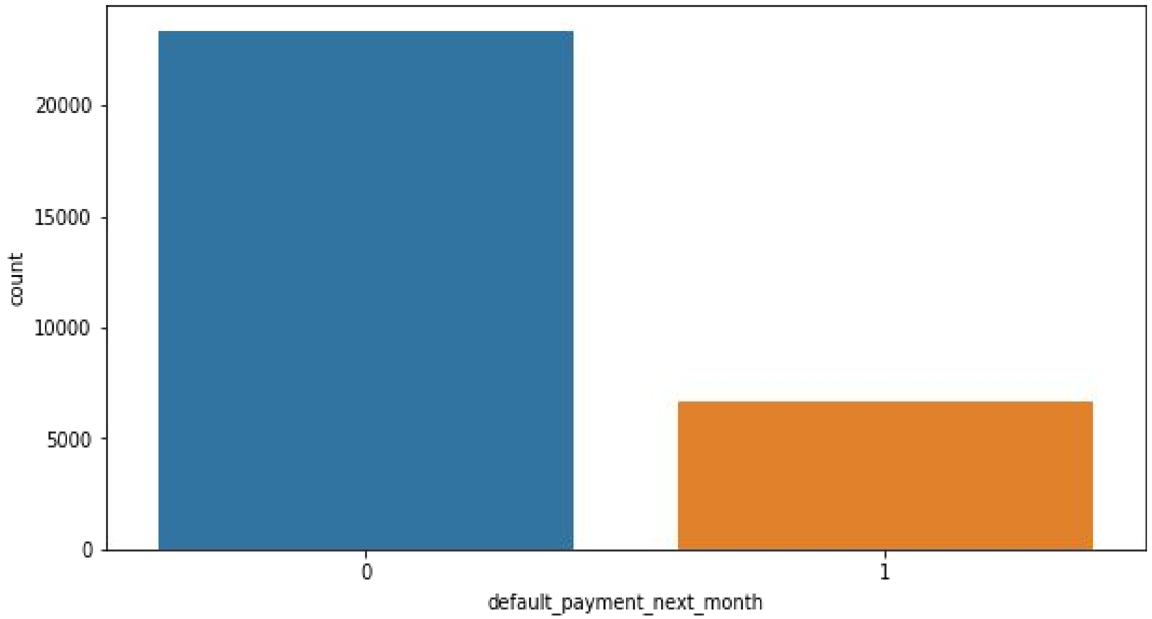
- X1-Amount of credit(includes individual as well as family credit)
- X2 -Gender
- X3 -Education
- X4 -Marital Status
- X5 -Age
- X6 to X11-History of past payments from April to September
- X12 to X17-Amount of bill statement from April to September
- X18 to X23 -Amount of previous payment from April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

INSIGHTS FROM OUR DATASET

- This Dataset is from Taiwan.
- In our data set there are 30000 rows, 25 columns
- There are No Missing Values present
- There are No Duplicate values present
- There are No null values.
- And finally we have 'default payment next month' variable which we need to predict for new observations
- All the features have integer datatype.
- The Columns are: - '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', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6', 'defaulters_payment_next_month'.

ANALYSIS OF DEPENDENT VARIABLE

- ➢ **As we can see from above graph that both classes are not in proportion and we have imbalanced dataset. we need to do normalize the data in next step.**

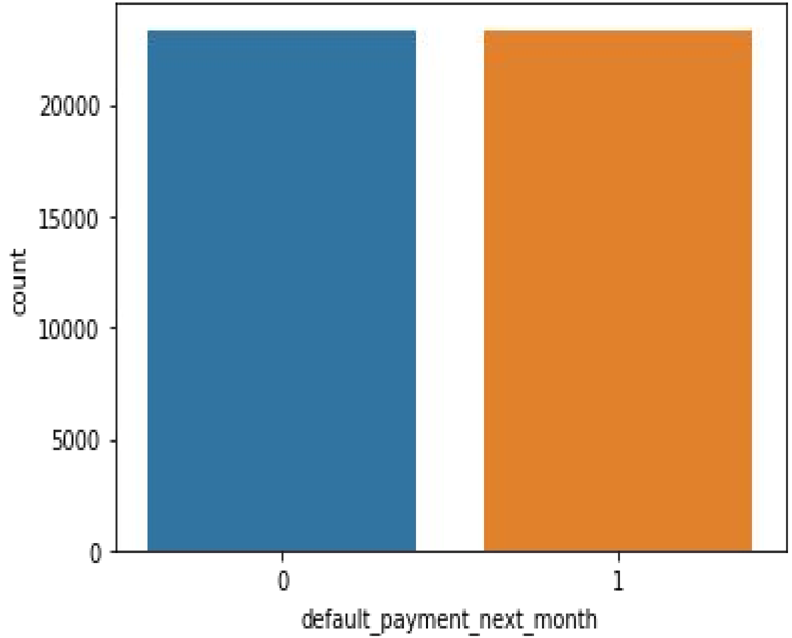


- **0 23364**
- **1 6636**

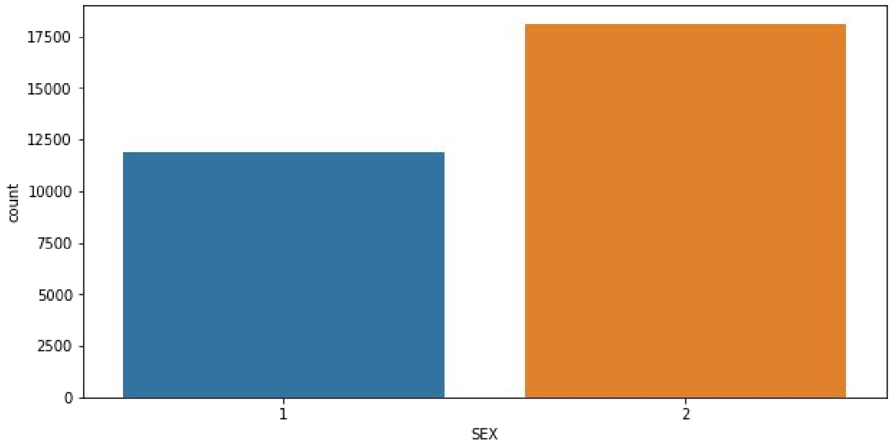
- **0 - Not Default**
- **1 - Default**

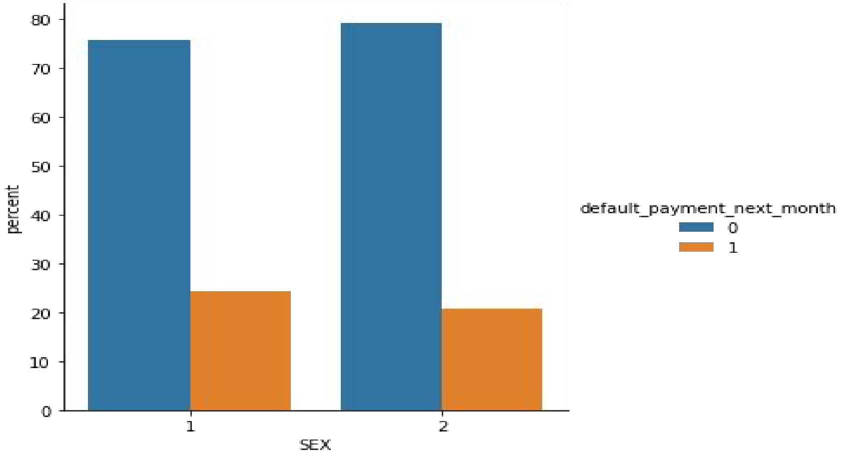
SMOTE

- **SMOTE (Synthetic Minority Oversampling Technique) – Oversampling is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them.**
- **After performing SMOTE operation we get this balance dataset**



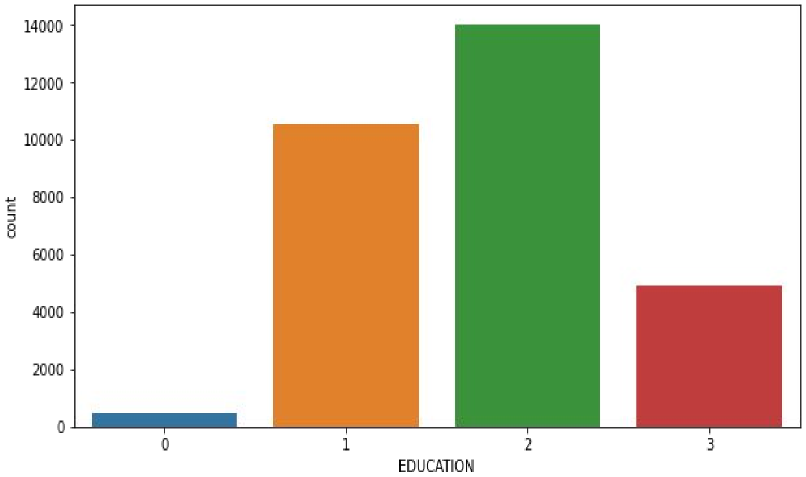
ANALYSIS OF SEX VARIABLE

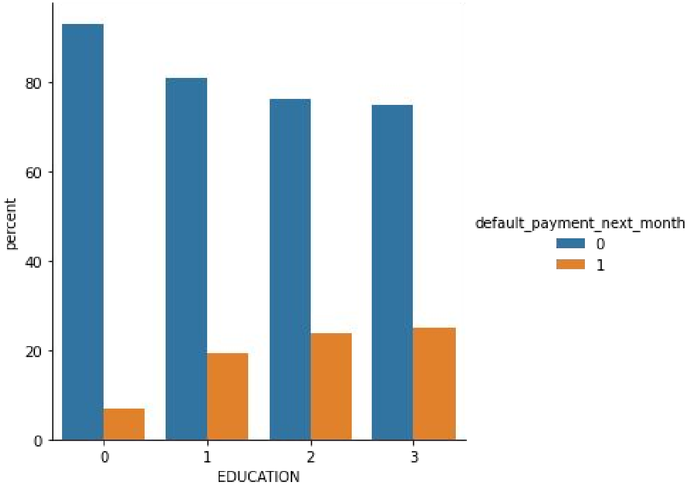




- 1- Male 2 - Female
- Number of Male credit holder is less than Female.
- It is evident from the above graph that the number of defaulter have high proportion of males

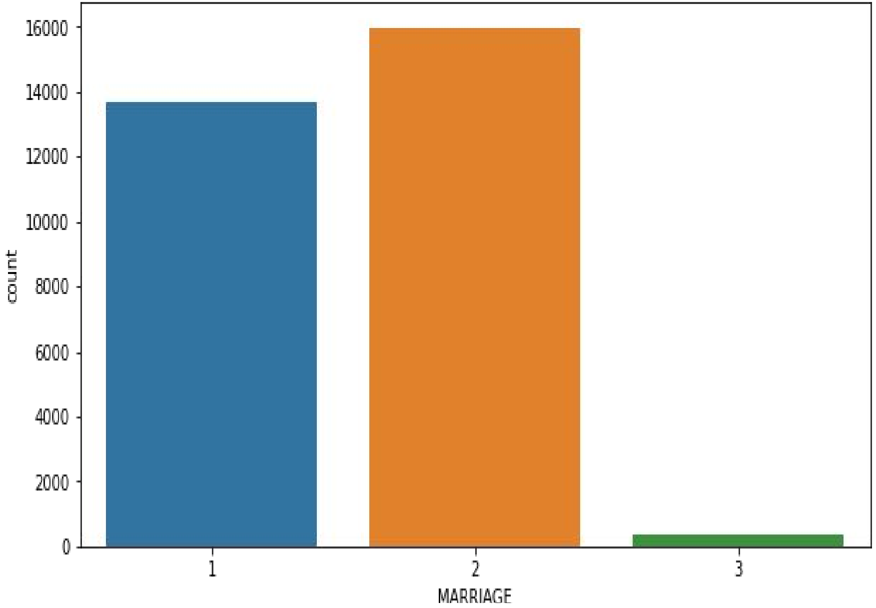
ANALYSIS OF EDUCATION VARIABLE

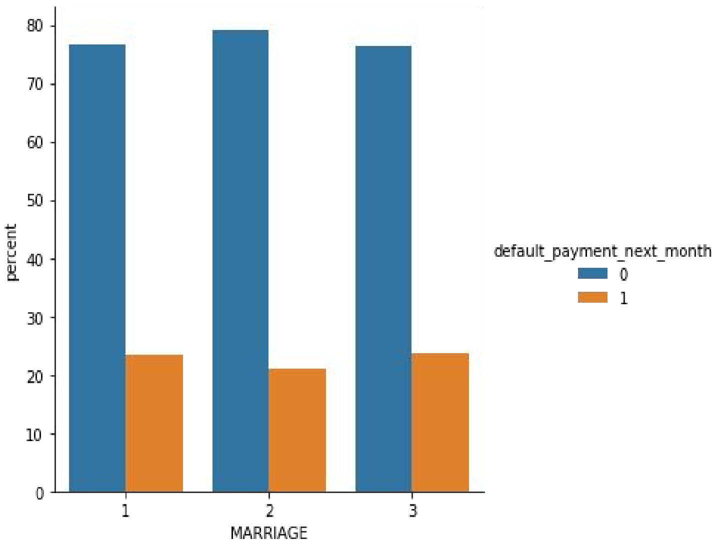




- ➢ 1=graduate school, 2=university, 3=high school, 0=others
- ➢ From the above left side plot we can say that
- ➢ More number of credit holders are university students followed by Graduates and then High school students.
- ➢ From the right side plot it is clear that those people who are other students have higher default payment wrt graduates and university people

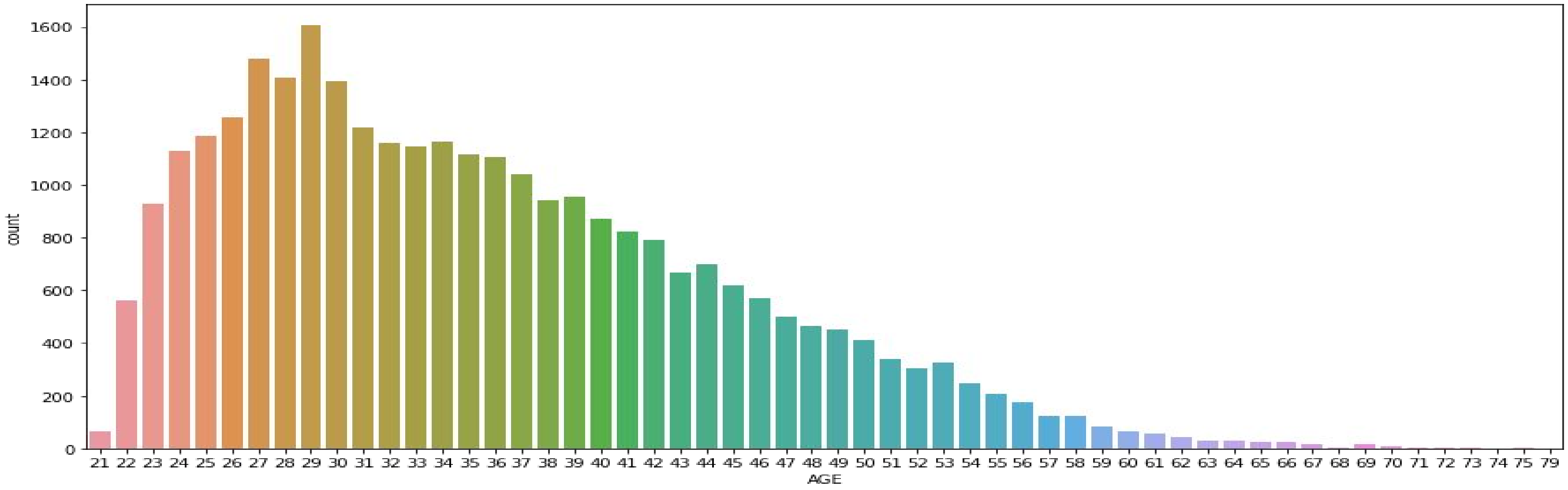
ANALYSIS OF MARRIAGE VARIABLE





- 1- married 2 - single 3 - others
- From the above data analysis we can say that
- More number of credit cards holder are Single.
- High defaulter rate when it comes to others

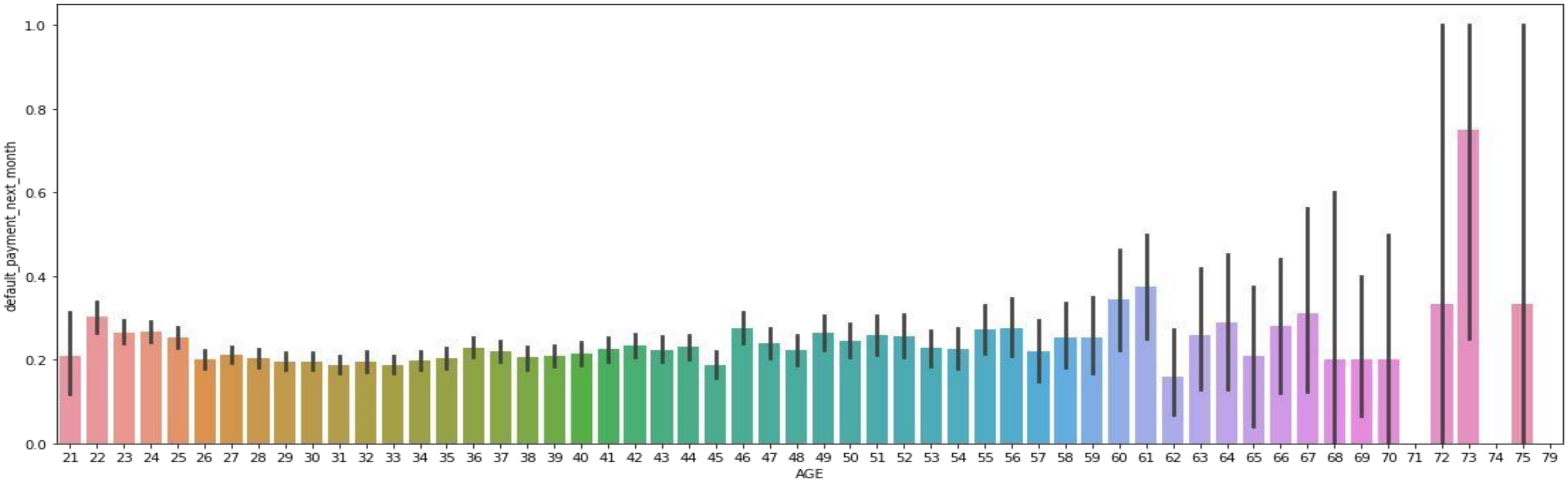
ANALYSIS OF AGE VARIABLE



From the above count plot analysis we can say that

- **We can see more number of credit cards holder age are between 26-30 years old.**
- **Age above 60 years old rarely uses the credit card.**

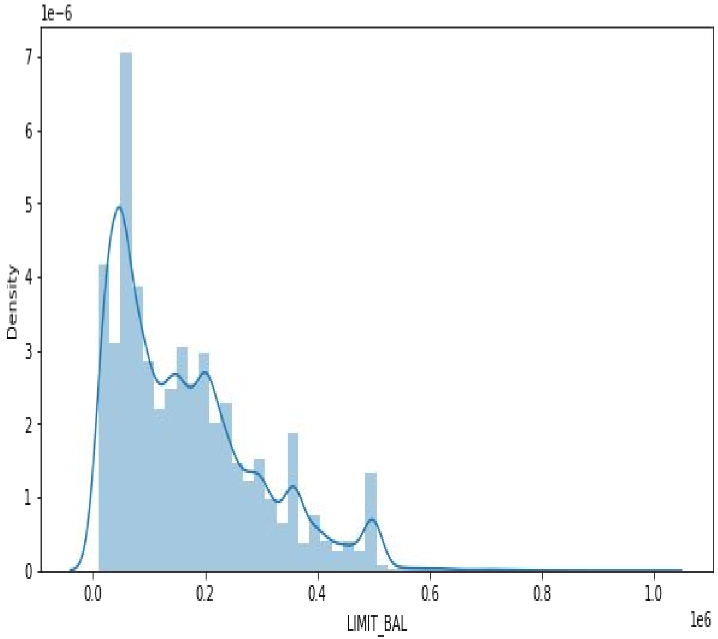
ANALYSIS OF AGE VARIABLE

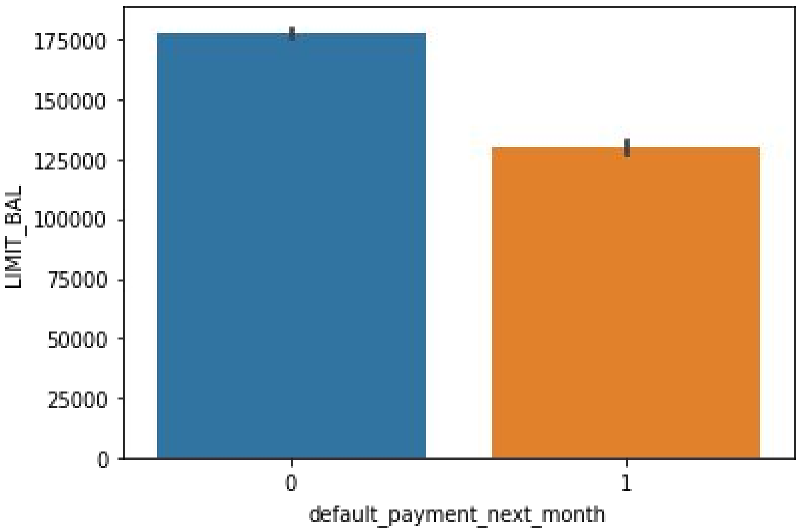


From the above bar plot which shows the relationship between age and defaulter, we can say that

- **Those who default are 60 years and older, that may be they don’t use their card frequently**

ANALYSIS OF LIMIT BALANCE VARIABLE

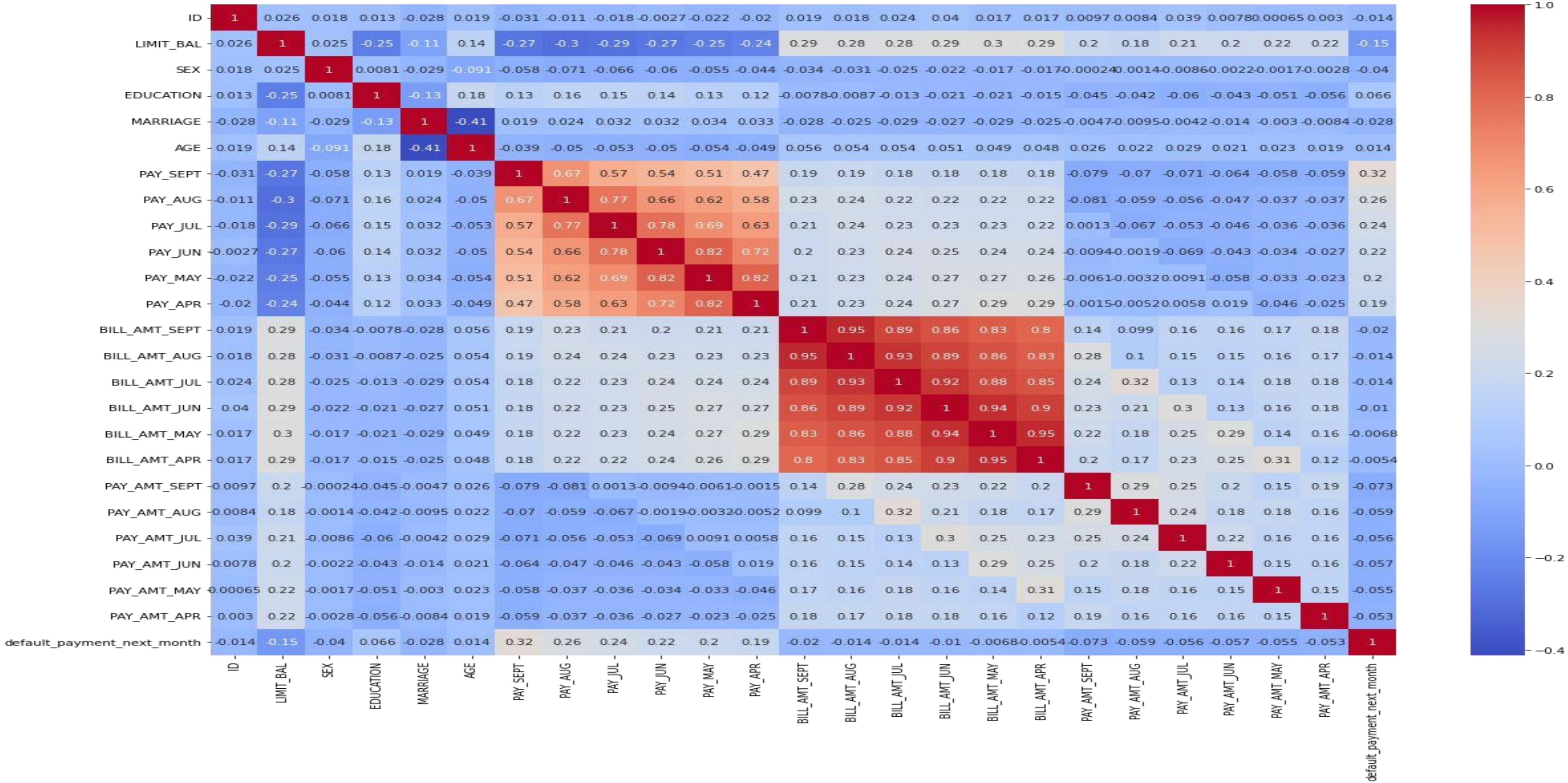




From the above plots analysis we can say that

- **Maximum amount of given credit in NT dollars is 50,000 followed by 30,000 and 20,000.**

CHECKING OF CORRELATION



ONE HOT ENCODING

- **One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.**
- **here we perform one hot encoding on 'EDUCATION','MARRIAGE','PAY_SEPT', 'PAY_AUG', 'PAY_JUL', 'PAY_JUN', 'PAY_MAY', 'PAY_APR'**
- **and label encoding for 'SEX'**
- **After this we get these features in our dataset:**

(['LIMIT_BAL', 'SEX', 'AGE', 'BILL_AMT_SEPT', 'BILL_AMT_AUG','BILL_AMT_JUL', 'BILL_AMT_JUN', 'BILL_AMT_MAY', 'BILL_AMT_APR','PAY_AMT_SEPT', 'PAY_AMT_AUG', 'PAY_AMT_JUL', 'PAY_AMT_JUN','PAY_AMT_MAY','PAY_AMT_APR', 'default_payment_next_month','total_Payement_Value', 'Dues', 'EDUCATION_graduate school','EDUCATION_high school', 'EDUCATION_others', 'EDUCATION_university','MARRIAGE_married', 'MARRIAGE_others','MARRIAGE_single', 'PAY_SEPT_-1','PAY_SEPT_0', 'PAY_SEPT_1', 'PAY_SEPT_2', 'PAY_SEPT_3', 'PAY_SEPT_4','PAY_SEPT_5', 'PAY_SEPT_6', 'PAY_SEPT_7', 'PAY_SEPT_8', 'PAY_AUG_-1','PAY_AUG_0', 'PAY_AUG_1', 'PAY_AUG_2', 'PAY_AUG_3', 'PAY_AUG_4','PAY_AUG_5', 'PAY_AUG_6','PAY_AUG_7', 'PAY_AUG_8', 'PAY_JUL_-1','PAY_JUL_0', 'PAY_JUL_1', 'PAY_JUL_2', 'PAY_JUL_3', 'PAY_JUL_4','PAY_JUL_5', 'PAY_JUL_6', 'PAY_JUL_7', 'PAY_JUL_8', 'PAY_JUN_-1','PAY_JUN_0', 'PAY_JUN_1', 'PAY_JUN_2', 'PAY_JUN_3', 'PAY_JUN_4','PAY_JUN_5', 'PAY_JUN_6', 'PAY_JUN_7', 'PAY_JUN_8', 'PAY_MAY_-1', 'PAY_MAY_0', 'PAY_MAY_1', 'PAY_MAY_2', 'PAY_MAY_3', 'PAY_MAY_4','PAY_MAY_5', 'PAY_MAY_6', 'PAY_MAY_7', 'PAY_MAY_8', 'PAY_APR_-1','PAY_APR_0', 'PAY_APR_1', 'PAY_APR_2', 'PAY_APR_3', 'PAY_APR_4','PAY_APR_5', 'PAY_APR_6', 'PAY_APR_7', 'PAY_APR_8'],

MODEL BUILDING

- LOGISTIC REGRESSION
- RANDOM FOREST
- SVC
- XGBOOST

LOGISTIC REGRESSION

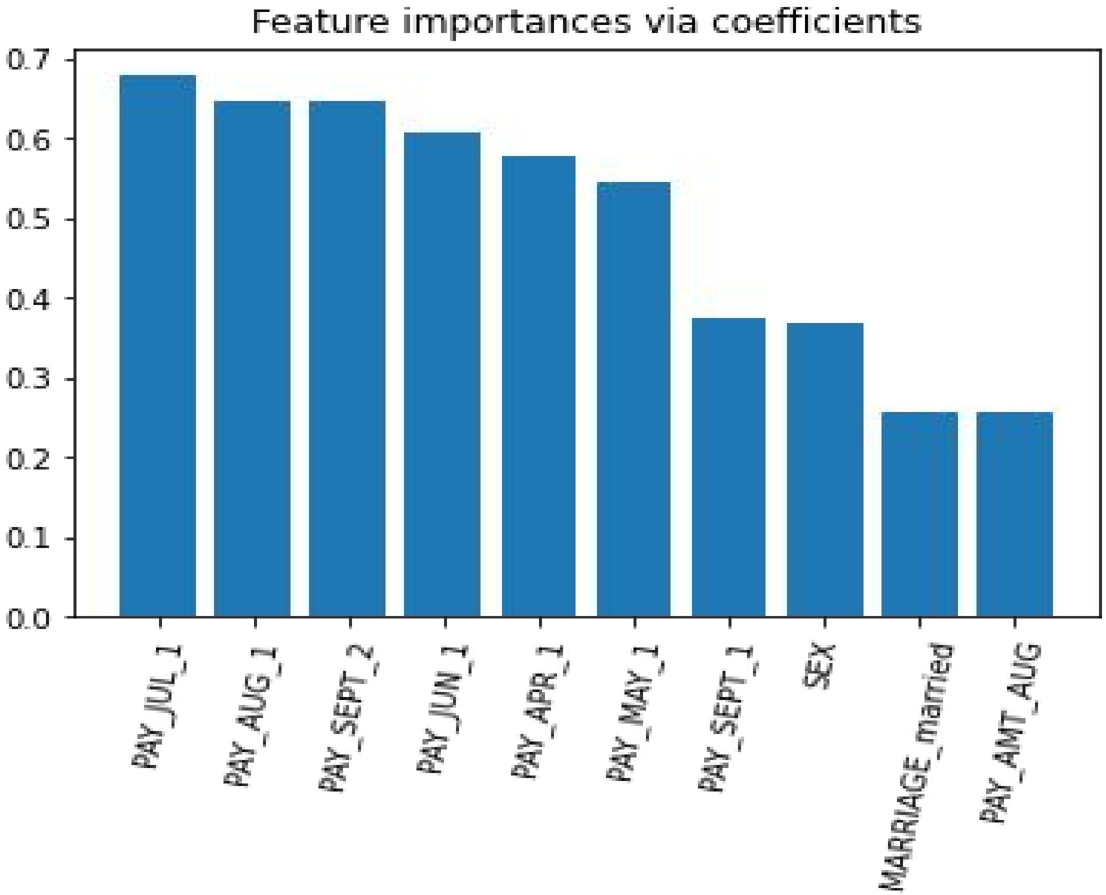
PARAMETERS :

{'C': 0.01, 'penalty': 'l2'}

from this regression model we get the results as below

- The accuracy on test data is **0.7553984825886778**
- The precision on test data is **0.6936446173800259**
- The recall on test data is **0.7913583900562297**
- The f1 on test data is **0.7392867016864806**
- The roc_score on test data is **0.7593522874903104**

FEATURE IMPORTANCES



RANDOM FOREST

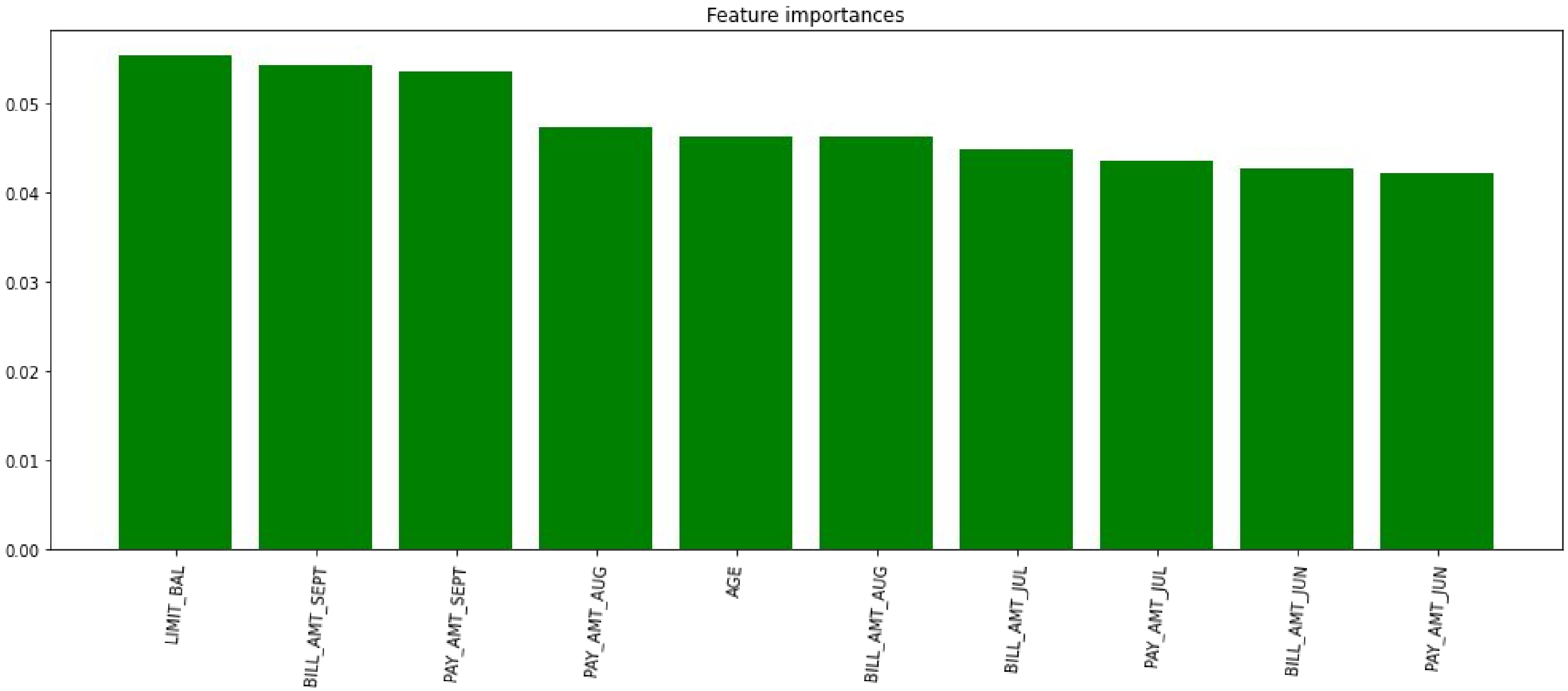
PARAMETERS :

{'max_depth': 30, 'n_estimators': 150}

from the regression model we get the results as below

- The accuracy on test data is **0.8337332209324947**
- The precision on test data is **0.8020752269779508**
- The recall on test data is **0.856272500692329**
- The f1 on test data is **0.8282882400214305**
- The roc_score on test data is **0.8350761210621055**

FEATURE IMPORTANCES



SUPPORT VECTOR CLASSIFIER (SVC)

PARAMETERS :

{'C': 10, 'kernel': 'rbf'}

from the regression model
we get the results as below

- The accuracy on test data is **0.766746644186499**
- The precision on test data is **0.6900129701686122**
- The recall on test data is **0.8150758388233492**
- The f1 on test data is **0.7473484582426073**
- The roc_score on test data is **0.7731776765513193**

XGBOOST

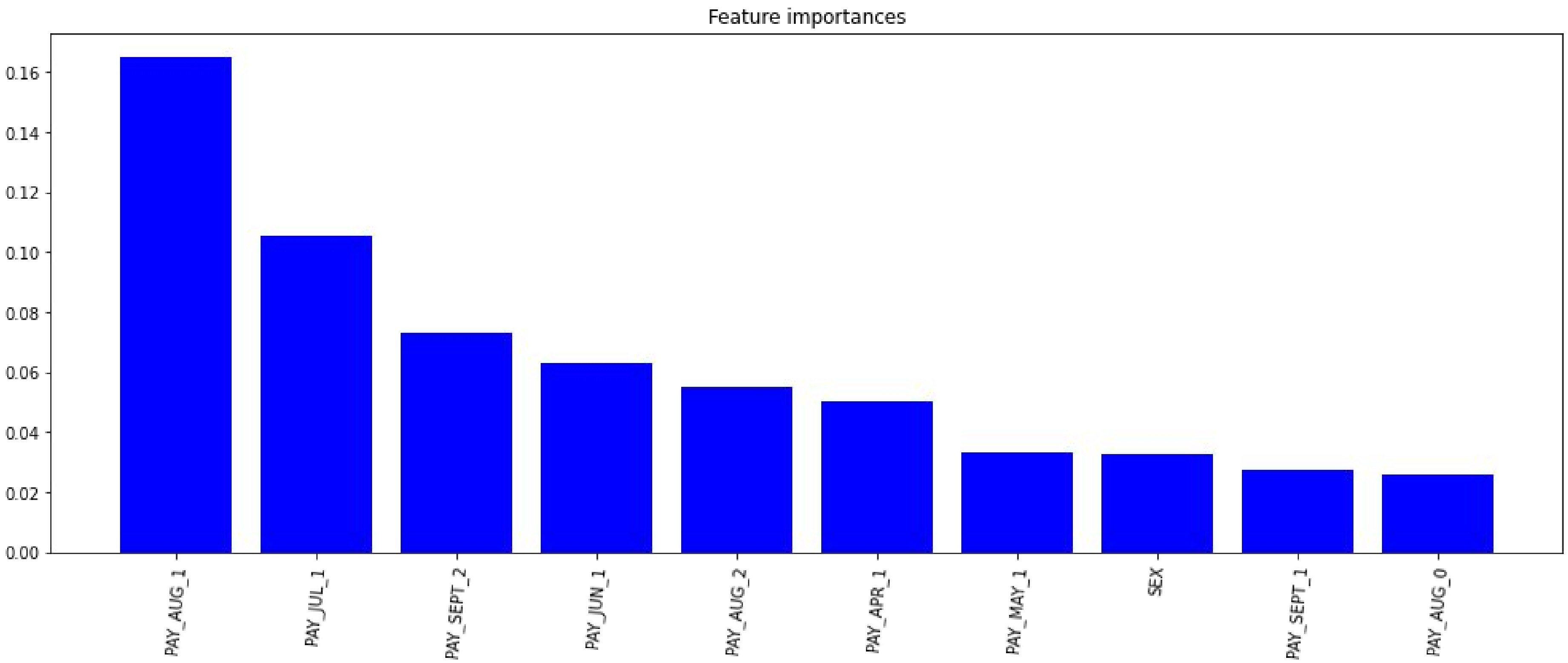
PARAMETERS :

{'max_depth': 15
'min_child_weight': 8}

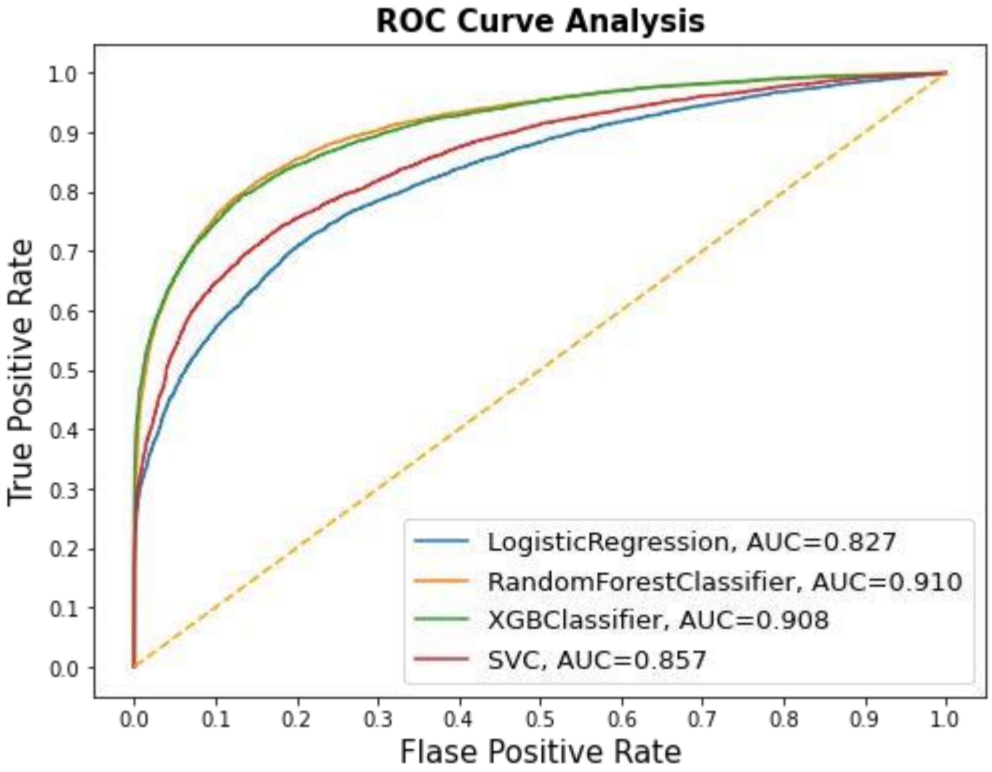
from the regression model
we get the results as below

- The accuracy on test data is **0.787562414888788**
- The precision on test data is **0.731647211413748**3
- The recall on test data is **0.8237441588785047**
- The f1 on test data is **0.7749690891605989**
- The roc_score on test data is **0.7912025355223038**

FEATURE IMPORTANCES



AUC-ROC CURVE COMPARISON



EVALUATING THE MODELS

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.753601	0.752091	0.687808	0.789254	0.735047
1	SVC	0.810713	0.779457	0.716732	0.819517	0.764686
2	Random Forest CLf	0.998722	0.832112	0.800389	0.854591	0.826602
3	Xgboost Clf	0.912448	0.829194	0.788197	0.858576	0.821883

CHALLENGES

- Large dataset to handle
- Feature Analysis
- Need to Remove outliers
- Feature engineering
- Getting a higher accuracy on the models.
- Carefully handled feature imbalanced data
- Carefully tuned Hyperparameters .

CONCLUSION

- Recent 2 months payment status and credit limit are the strongest default predictors.
- Recall is the best accuracy metrics here, because if the algorithm will not detect the defaulters, that will encounter more loss to the bank
- XGBoost provided us the best results giving us a recall of 85% percent(meaning out of 100 defaulters 85 will be correctly caught by XGBoost)
- Random Forest also had good score as well but leads to overfit the data.
- Logistic regression being the least accurate with a recall of 79%.
- Higher recall can be achieved if low precision is acceptable.
- This Model can only be served as an aid in decision making instead of replacing human decision.
- Model can be improved with more data and computational resources.

THANK YOU