## **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 becomeseverely delinquent on your credit card payments. Missing credit card payments once or twicedoes 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

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

default

#### **FEATURE SUMMARY**

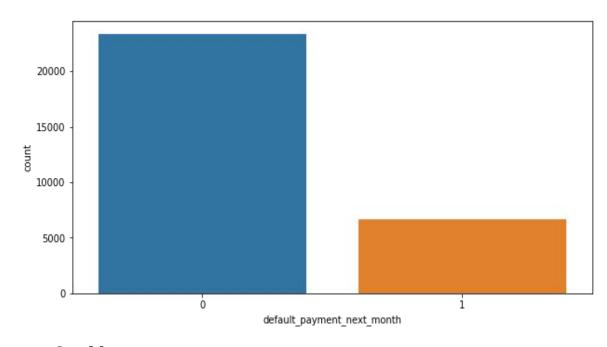
- > X1-Amount of credit(includes individual as well as family credit)
- X2 -Gender
- X3 -Education
- > X4 -Marital Status
- > X5 -Age
- > X6 to X11-History of past payments from April to September
- > X12to X17-Amount of bill statement from April to September
- X18 to X23 Amount of previous payment from 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\_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.

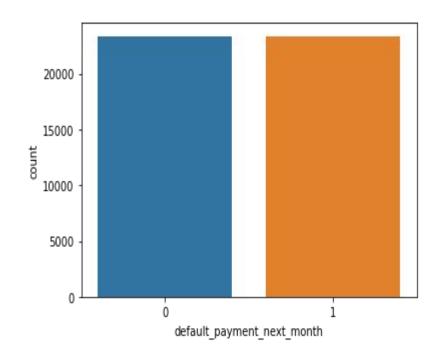


- 0 23364
- 1 6636

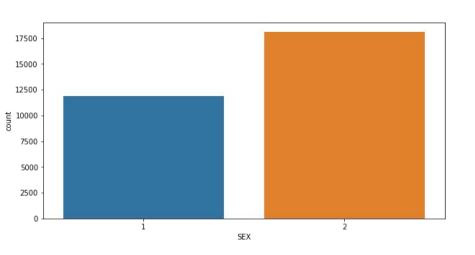
- 0 Not Default
- 1 Default

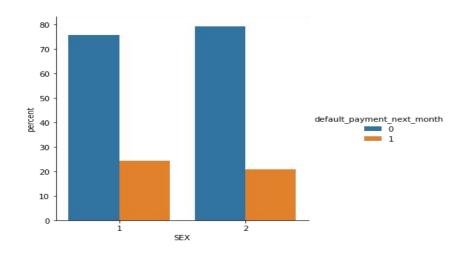
#### **SMOTE**

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



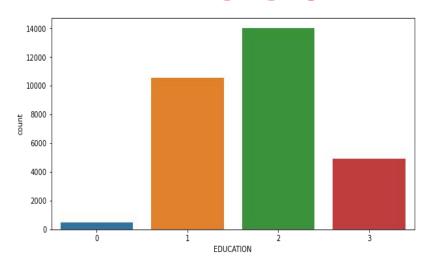
#### **ANALYSIS OF SEX VARIABLE**

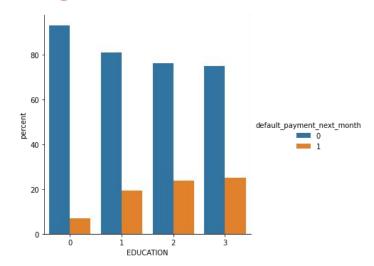




- > 1- Male 2 Female
- Number of Male credit holder is less than Female.
- It is evident from the above graph that the number of defaulter have high proportion of males

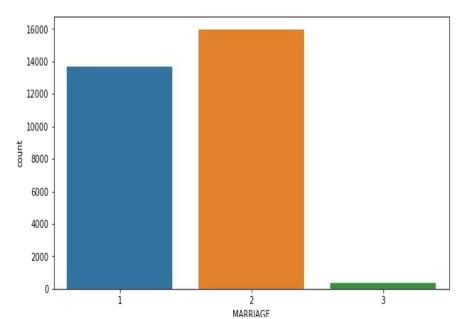
#### **ANALYSIS OF EDUCATION VARIABLE**

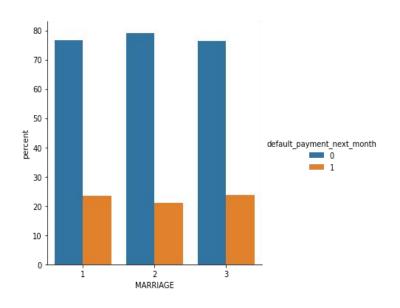




- > 1=graduate school, 2=university, 3=high school, 0=others
- > From the above left side plot we can say that
- More number of credit holders are university students followed by Graduates and then High school students.
- > From the right side plot it is clear that those people who are other students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

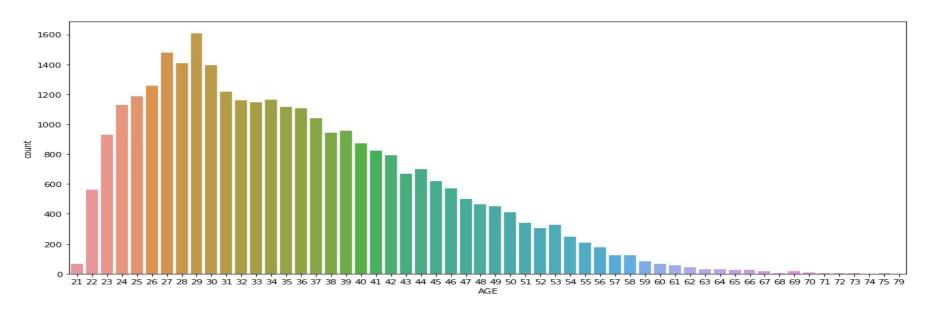
#### **ANALYSIS OF MARRIAGE VARIABLE**





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

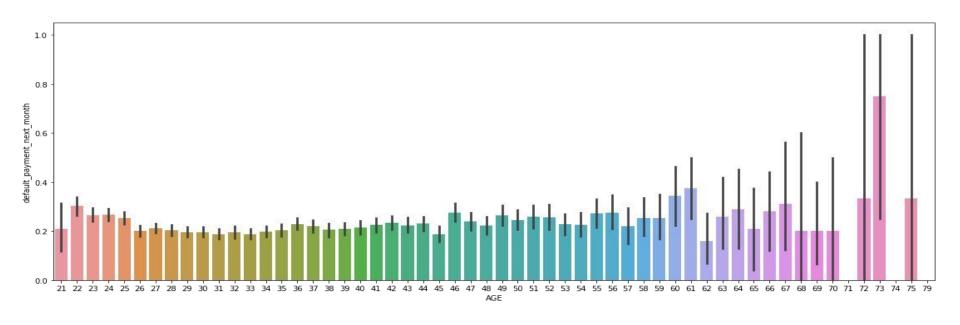
#### **ANALYSIS OF AGE VARIABLE**



#### 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.

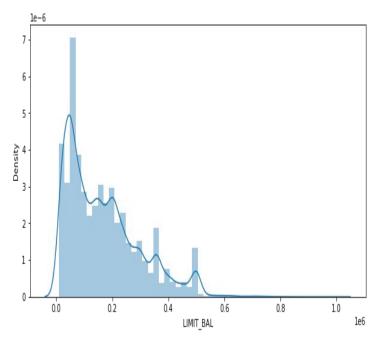
#### **ANALYSIS OF AGE VARIABLE**

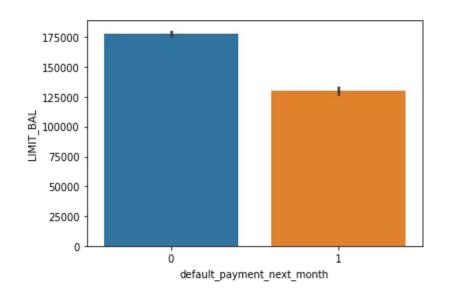


From the above bar plot which shows the relationship between age and defaulter, we can say that

 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 saythat

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

#### **CHECKING OF CORRELATION**

ID -	1	0.026	0.018	0.013	-0.028	0.019	-0.031	-0.011	-0.018	-0.0027	-0.022	-0.02	0.019	0.018	0.024	0.04	0.017	0.017	0.0097	0.0084	0.039	0.00780	0.00065	0.003	-0.014
LIMIT_BAL	0.026	1	0.025			0.14						-0.24	0.29	0.28	0.28	0.29	0.3	0.29	0.2	0.18	0.21	0.2	0.22	0.22	-0.15
SEX -	0.018	0.025	1	0.0081	-0.029		-0.058	-0.071	-0.066	-0.06	-0.055	-0.044	-0.034	-0.031	-0.025	-0.022	-0.017	-0.0174	0.00024	10.0014	0.0086	0.0022	0.0017-	0.0028	-0.04
EDUCATION -	0.013		0.0081		-0.13	0.18	0.13	0.16	0.15	0.14	0.13	0.12	0.0078	0.0087	-0.013	-0.021	-0.021	-0.015	-0.045	-0.042	-0.06	-0.043	-0.051	0.056	0.066
MARRIAGE -	-0.028		-0.029	-0.13	1	-0.41	0.019	0.024	0.032	0.032	0.034	0.033	-0.028	-0.025	-0.029	-0.027	-0.029	-0.025	0.0047	-0.0095	0.0042	-0.014	-0.003 -	0.0084	-0.028
AGE -	0.019	0.14		0.18	-0.41	1	-0.039	-0.05	-0.053	-0.05	-0.054	-0.049	0.056	0.054	0.054	0.051	0.049	0.048	0.026	0.022	0.029	0.021	0.023	0.019	0.014
PAY_SEPT	-0.031	-0.27	-0.058	0.13	0.019	-0.039	1		0.57	0.54	0.51	0.47	0.19	0.19	0.18	0.18	0.18	0.18	-0.079	-0.07	-0.071	-0.064	-0.058	-0.059	0.32
PAY_AUG -	-0.011		-0.071	0.16	0.024	-0.05	0.67	1	0.77	0.66	0.62	0.58	0.23	0.24	0.22	0.22	0.22	0.22	-0.081	-0.059	-0.056	-0.047	-0.037	-0.037	0.26
PAY_JUL -	-0.018		-0.066	0.15	0.032	-0.053	0.57	0.77	1			0.63	0.21	0.24	0.23	0.23	0.23	0.22	0.0013	-0.067	-0.053	-0.046	-0.036	-0.036	0.24
PAY_JUN -	0.0027		-0.06	0.14	0.032	-0.05	0.54	0.66	0.78	1	0.82	0.72	0.2	0.23	0.24	0.25	0.24	0.24	0.0094	0.0019	-0.069	-0.043	-0.034	-0.027	0.22
PAY_MAY -	-0.022		-0.055	0.13	0.034	-0.054	0.51	0.62	0.69	0.82	1	0.82	0.21	0.23	0.24	0.27	0.27	0.26	0.0061	-0.0032	0.0091	-0.058	-0.033	-0.023	0.2
PAY_APR ·	-0.02		-0.044	0.12	0.033	-0.049	0.47	0.58	0.63	0.72	0.82	1	0.21	0.23	0.24	0.27	0.29	0.29	0.0015	0.0052	0.0058	0.019	-0.046	-0.025	0.19
BILL_AMT_SEPT -	0.019	0.29	-0.034	-0.0078	-0.028	0.056	0.19	0.23	0.21	0.2	0.21	0.21	1	0.95	0.89	0.86	0.83	0.8	0.14	0.099	0.16	0.16	0.17	0.18	-0.02
BILL_AMT_AUG -	0.018	0.28	-0.031	-0.0087	-0.025	0.054	0.19	0.24	0.24	0.23	0.23	0.23		1				0.83	0.28	0.1	0.15	0.15	0.16	0.17	-0.014
BILL AMT JUL	0.024	0.28	-0.025	-0.013	-0.029	0.054	0.18	0.22	0.23	0.24	0.24	0.24	0.89		1			0.85	0.24	0.32	0.13	0.14	0.18	0.18	-0.014
BILL_AMT_JUN -	0.04	0.29	-0.022	-0.021	-0.027	0.051	0.18	0.22	0.23	0.25	0.27	0.27		0.89		1	0.94	0.9	0.23	0.21	0.3	0.13	0.16	0.18	-0.01
BILL AMT MAY	0.017	0.3	-0.017	-0.021	-0.029	0.049	0.18	0.22	0.23	0.24	0.27	0.29			0.88	0.94		0.95	0.22	0.18	0.25	0.29	0.14	0.16	-0.0068
BILL AMT APR	0.017	0.29	-0.017	-0.015	-0.025	0.048	0.18	0.22	0.22	0.24	0.26	0.29					0.95	1	0.2	0.17	0.23	0.25	0.31	0.12	0.0054
PAY AMT SEPT	0.0097	0.2 -	0.00024	10.045	0.0047	0.026	-0.079	-0.081	0.0013	-0.0094	0.0061	0.0015	0.14	0.28	0.24	0.23	0.22	0.2	1	0.29	0.25	0.2	0.15	0.19	-0.073
PAY AMT AUG	0.0084	0.18	-0.0014	-0.042	0.0095	0.022	-0.07	-0.059	-0.067	-0.0019	0.0032	0.0052	0.099	0.1	0.32	0.21	0.18	0.17	0.29	1	0.24	0.18	0.18	0.16	-0.059
PAY AMT JUL -	0.039	0.21	-0 0086	-0.06	0 0042	0.029	-0 071	-0.056	-0 053	-0.069	0 0091	0.0058	0.16	0.15	0.13	0.3	0.25	0.23	0.25	0.24	1	0.22	0.16	0.16	-0.056
PAY AMT JUN																0.13		0.25	0.2	0.18	0.22	1	0.15		
PAY AMT MAY																	0.14	0.31	0.15	0.18	0.16	0.15	1	0.15	
PAY AMT APR																						0.16	0.15		-0.053
default payment next month										0.22	0.2		-0.02												1
deladic_payment_next_month	_	-	_			-	,				,	-	-			,			,	1		,			-
	=	LIMIT_BAL	æ	EDUCATION	MARRIAGE	AGE	PAY_SEPT	PAY_AUG	PAY JUL	PAY JUN	PAY_MAY	PAY_APR	_AMT_SEPT	_AMT_AUG	AMT JUL	BILL_AMT_JUN	AMT_MAY	_AMT_APR	AMT_SEPT	_AMT_AUG	_AMT_JUL	AMT_JUN	AMT_MAY	PAY_AMT_APR	/ment_next_month
		=		ā	W/		~	-				_	MAM	BILLA		BILL	BILL_A	BILL_A	PAY_AI	PAY_AI	PAY	PAY A	PAY A	PAY A	nt_next
													ш				_		_						ymer

#### 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_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_5', 'PAY_AUG_7', 'PAY_AUG_8', 'PAY_JUL_-1', '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_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_2', 'PAY_APR_3', 'PAY_APR_5', 'PAY_APR_5', 'PAY_APR_6', 'PAY_APR_7', 'PAY_APR_8'],
```

#### **MODEL BUILDING**

- > LOGISTIC REGRESSION
- > RANDOM FOREST
- > SVC
- > XGBOOST

#### **LOGISTIC REGRESSION**

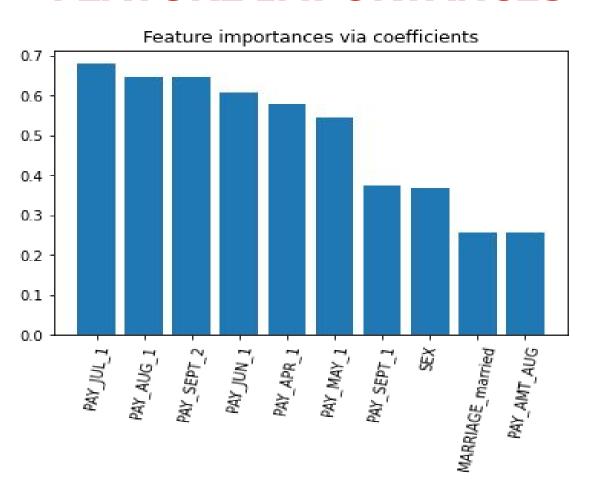
#### **PARAMETERS:**

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

## from this regression model we get the results as below

- The accuracy on test data is0.7553984825886778
- The precision on test data is
   0.6936446173800259
- The recall on test data is0.7913583900562297
- The f1 on test data is0.7392867016864806
- The roc\_score on test data is0.7593522874903104

#### **FEATURE IMPORTANCES**



#### RANDOM FOREST

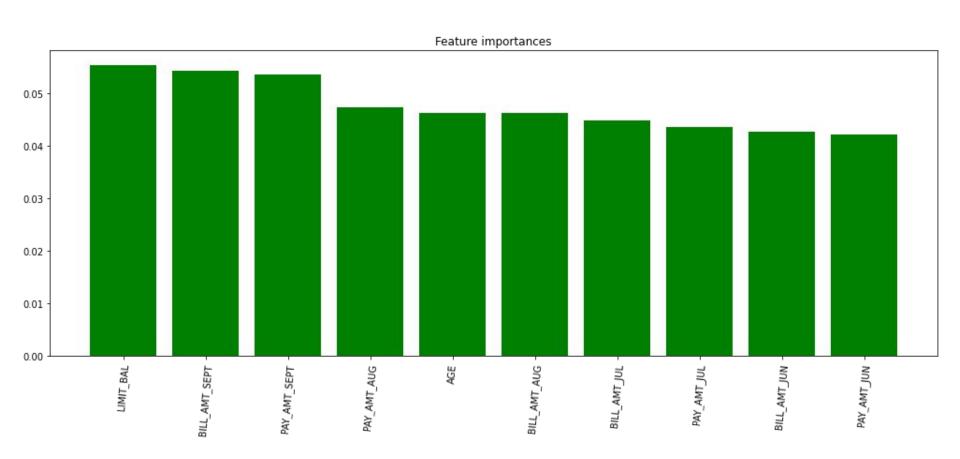
#### **PARAMETERS:**

{'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 is0.856272500692329
- The f1 on test data is0.8282882400214305
- 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 is0.8150758388233492
- The f1 on test data is0.7473484582426073
- The roc\_score on test data
   is 0.7731776765513193

#### **XGBOOST**

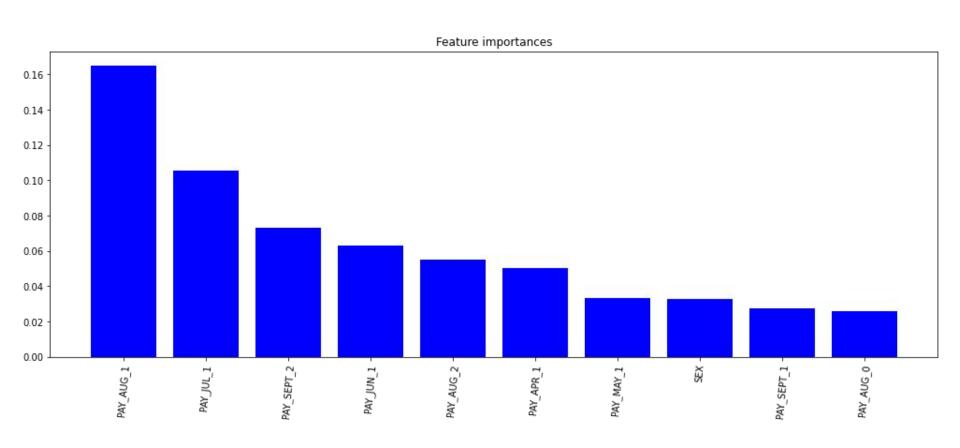
#### **PARAMETERS:**

{'max\_depth': 15
'min\_child\_weight': 8}

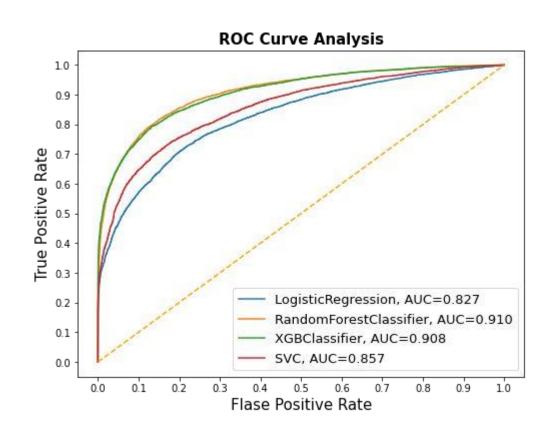
## from the regression model we get the results as below

- The accuracy on test data is 0.787562414888788
- The precision on test data
   is 0.7316472114137483
- The recall on test data is0.8237441588785047
- The f1 on test data is0.7749690891605989
- 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