CAPSTONE PROJECT : 3

CREDIT CARD DEFAULT PREDECTION

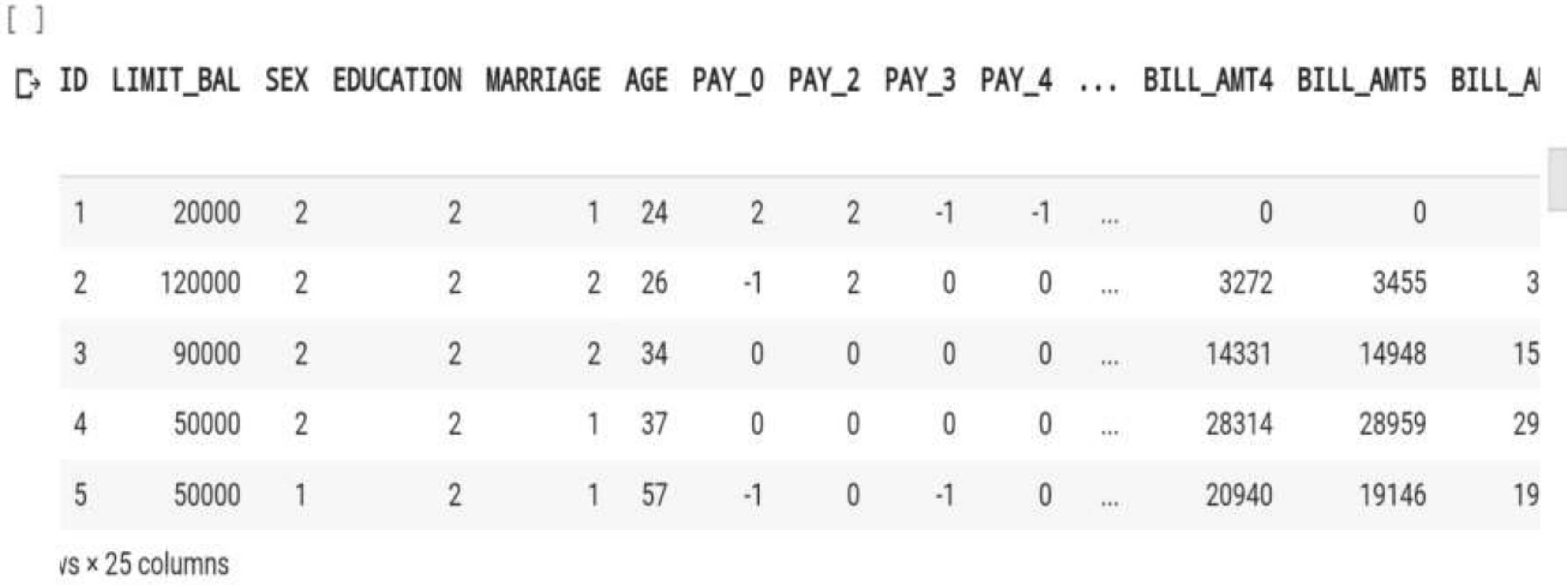
BY

ABHAY DEEP SINGH

PROBLEM DESCRIPTION:

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. We can use the K-S chart to evaluate which customers will default on their credit card payments.

DATA SUMMARY



[ ]

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_A
1	20000	2	2	1	24	2	2	-1	-1	...	0	0	
2	120000	2	2	2	26	-1	2	0	0	...	3272	3455	3
3	90000	2	2	2	34	0	0	0	0	...	14331	14948	15
4	50000	2	2	1	37	0	0	0	0	...	28314	28959	29
5	50000	1	2	1	57	-1	0	-1	0	...	20940	19146	19

vs × 25 columns

DATA SUMMARY

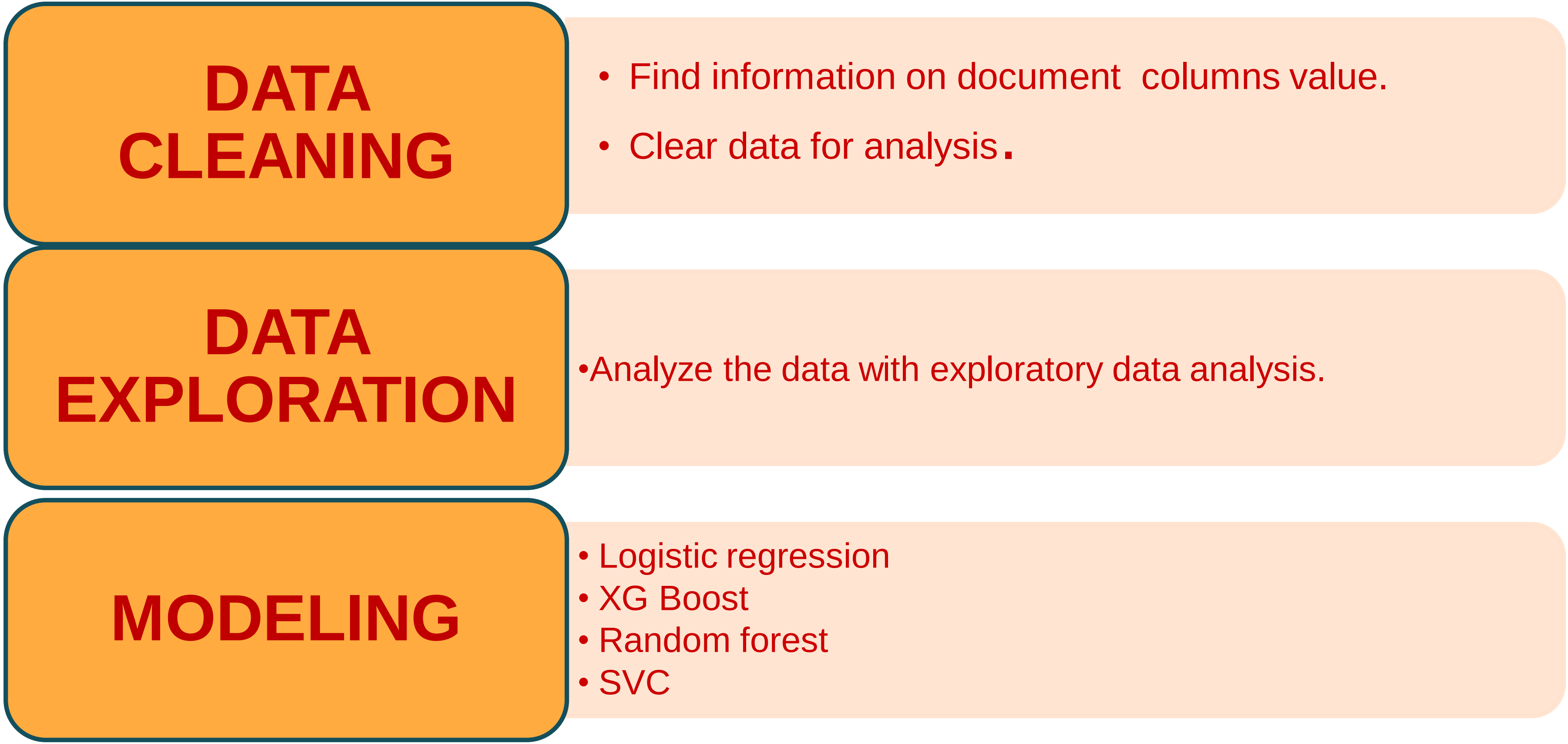
- This data set contains 30000 rows and 23 columns of six month.
- There are nine categorical features in our dataset.
- There are no missing value.
- There are no duplicate value.
- There are no null value.

Feature Summary

X1: Amount of the given credit, includes both individual and family credit.

- X2: Gender(1=Male and 2=Female)
- X3: Education(1=graduate, 2= university, 3= high school and 4= others)
- X4: Marital status (1= Married, 2 = single, 3= others)
- X5: Age in year.
- X6-X11: History of past payment from April to September
- X12-17: Amount of bill statement fro April to September
- X18-X23: Amount of previous payment from April to September
- Y: Default payment

Approach to analyze the dataset



DATA CLEANING

- Find information on document columns value.
- Clear data for analysis.

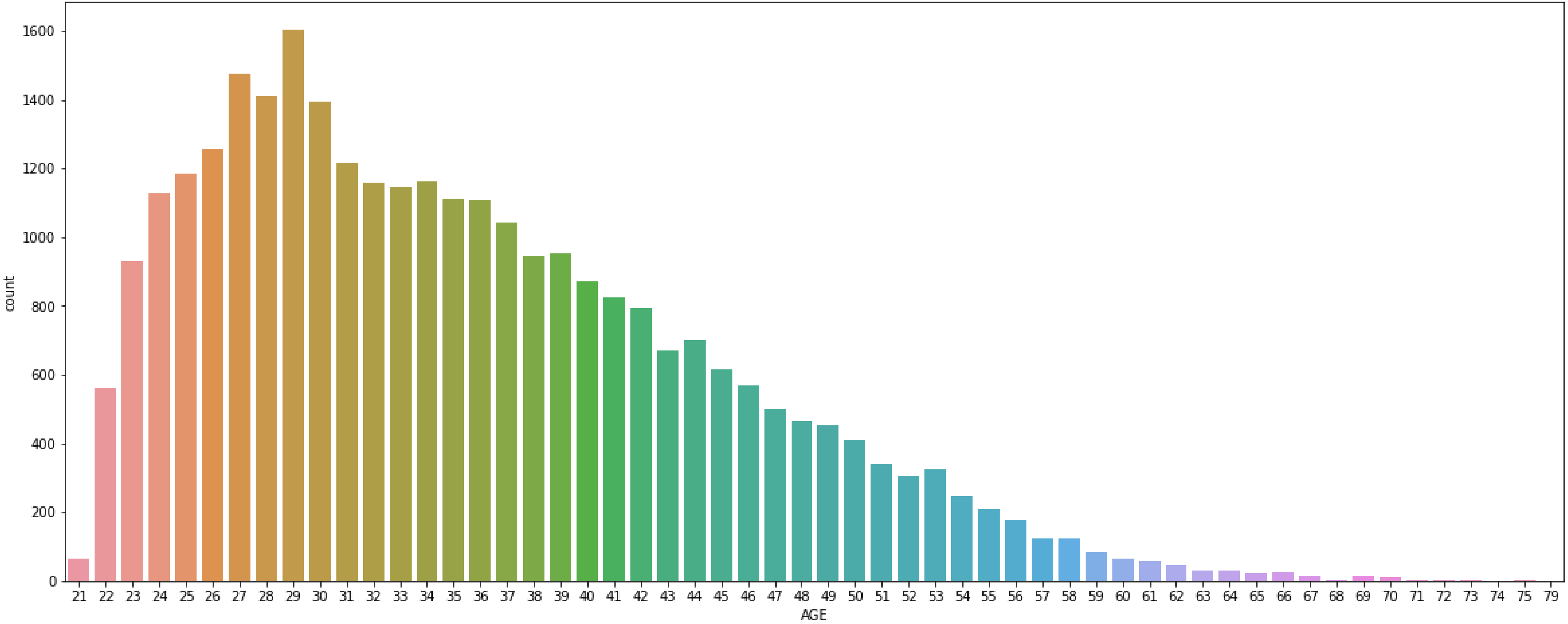
DATA EXPLORATION

- Analyze the data with exploratory data analysis.

MODELING

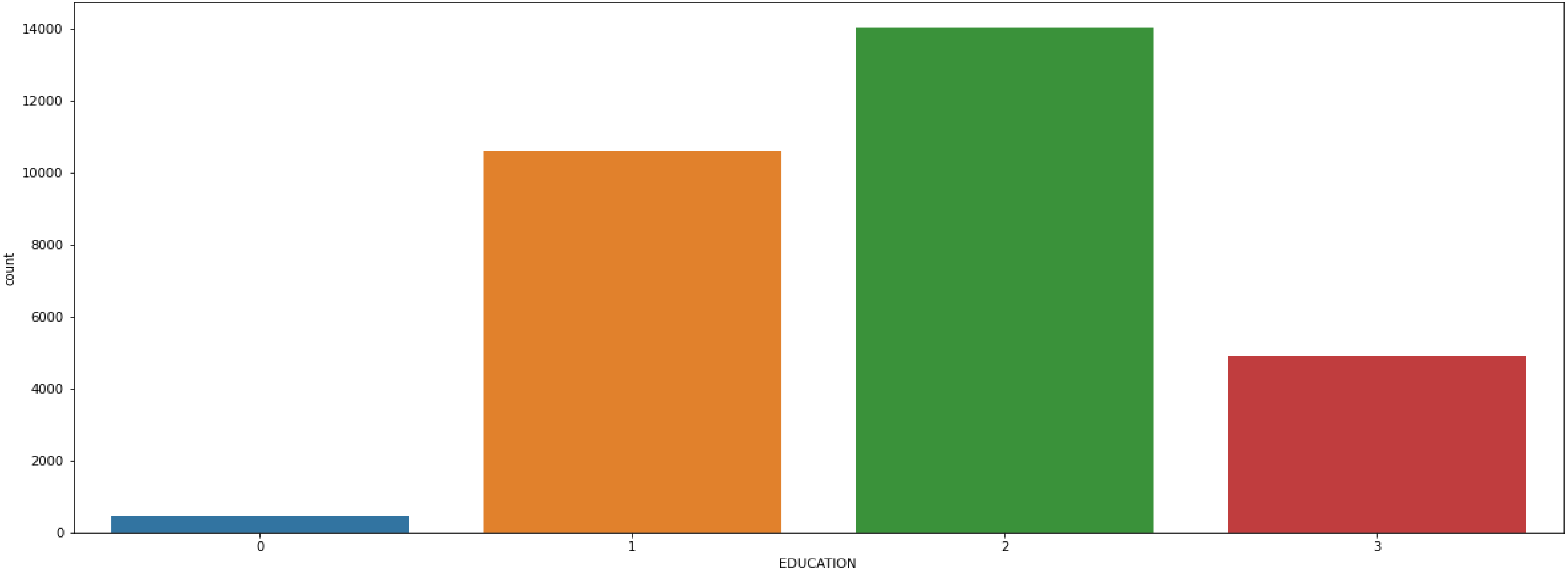
- Logistic regression
- XG Boost
- Random forest
- SVC

Count of credit card on basis of age



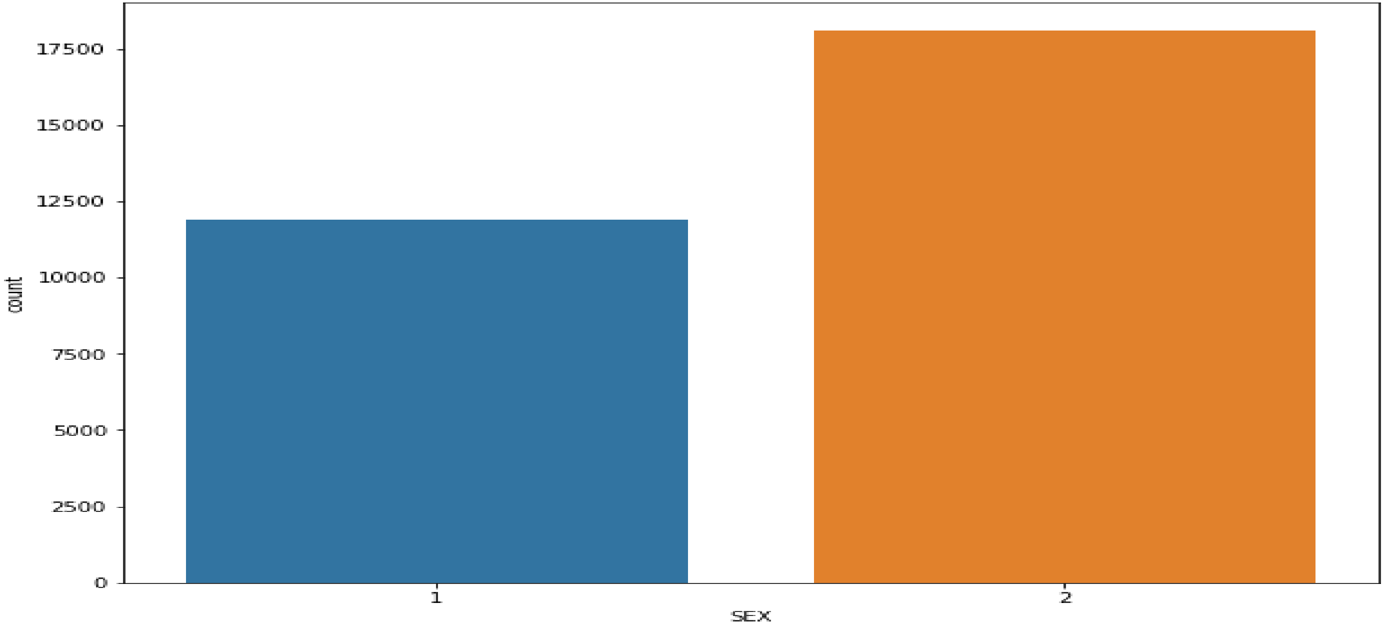
People from age 24 to 36 uses more credit and as the age increase the count decrease.

Count of credit card basis of education



- Most number of credit card are used by university students and university student shown by 2.

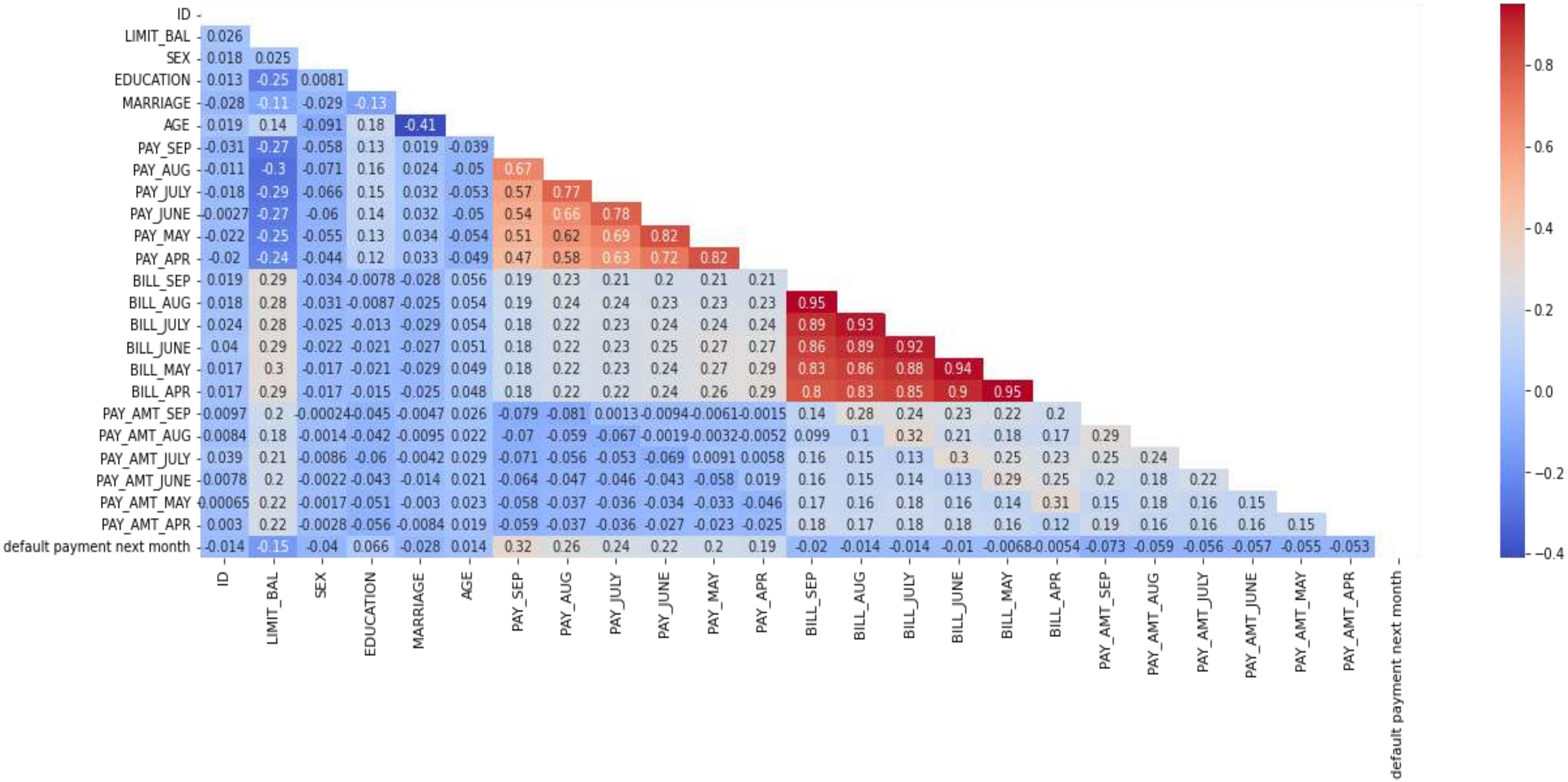
Count of credit card on basis of Gender



- From above plot we can see that number of female credit card holder are more than male.

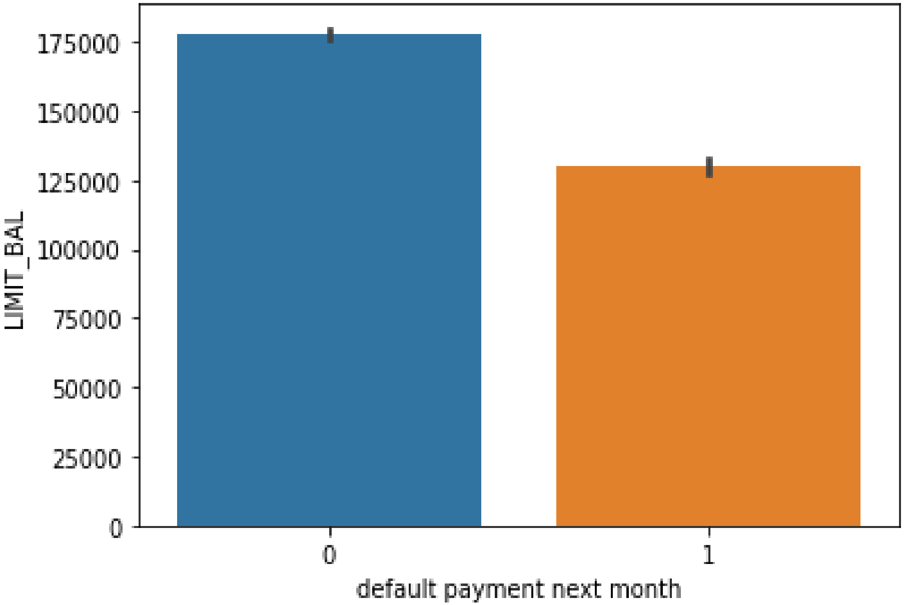
- **1: Male**
- **2: Female**

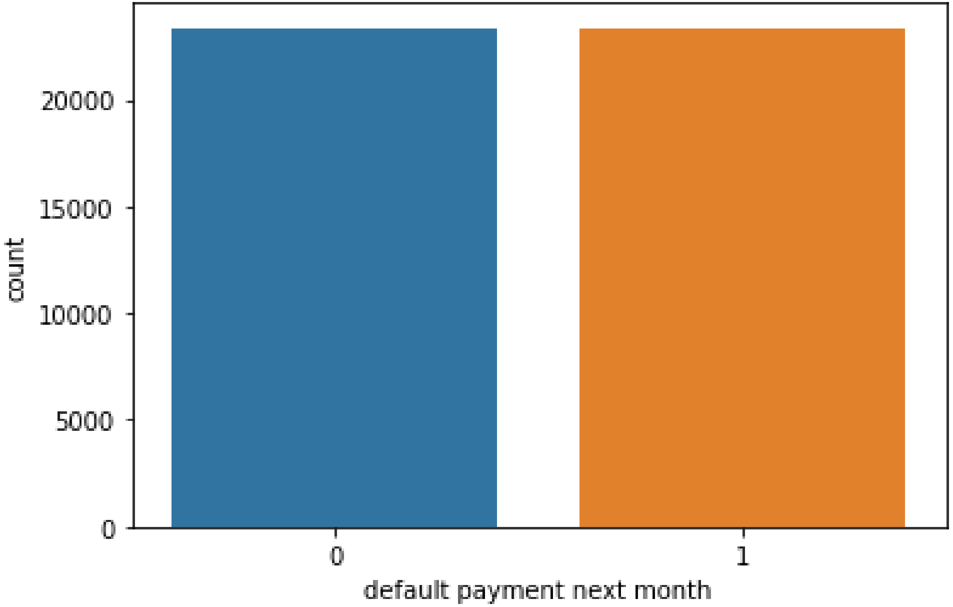
Correlation Matrix



	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_SEP	PAY_AUG	PAY_JULY	PAY_JUNE	PAY_MAY	PAY_APR	BILL_SEP	BILL_AUG	BILL_JULY	BILL_JUNE	BILL_MAY	BILL_APR	PAY_AMT_SEP	PAY_AMT_AUG	PAY_AMT_JULY	PAY_AMT_JUNE	PAY_AMT_MAY	PAY_AMT_APR	default payment next month
ID																									
LIMIT_BAL	0.026																								
SEX	0.018	0.025																							
EDUCATION	0.013	-0.25	0.0081																						
MARRIAGE	-0.028	-0.11	-0.029	-0.13																					
AGE	0.019	0.14	-0.091	0.18	-0.41																				
PAY_SEP	-0.031	-0.27	-0.058	0.13	0.019	-0.039																			
PAY_AUG	-0.011	-0.3	-0.071	0.16	0.024	-0.05	0.67																		
PAY_JULY	-0.018	-0.29	-0.066	0.15	0.032	-0.053	0.57	0.77																	
PAY_JUNE	-0.0027	-0.27	-0.06	0.14	0.032	-0.05	0.54	0.66	0.78																
PAY_MAY	-0.022	-0.25	-0.055	0.13	0.034	-0.054	0.51	0.62	0.69	0.82															
PAY_APR	-0.02	-0.24	-0.044	0.12	0.033	-0.049	0.47	0.58	0.63	0.72	0.82														
BILL_SEP	0.019	0.29	-0.034	-0.0078	-0.028	0.056	0.19	0.23	0.21	0.2	0.21	0.21													
BILL_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	0.95												
BILL_JULY	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	0.93											
BILL_JUNE	0.04	0.29	-0.022	-0.021	-0.027	0.051	0.18	0.22	0.23	0.25	0.27	0.27	0.86	0.89	0.92										
BILL_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	0.83	0.86	0.88	0.94									
BILL_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	0.8	0.83	0.85	0.9	0.95								
PAY_AMT_SEP	0.0097	0.2	-0.00024	-0.045	-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							
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						
PAY_AMT_JULY	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					
PAY_AMT_JUNE	0.0078	0.2	-0.0022	-0.043	-0.014	0.021	-0.064	-0.047	-0.046	-0.043	-0.058	0.019	0.16	0.15	0.14	0.13	0.29	0.25	0.2	0.18	0.22				
PAY_AMT_MAY	0.00065	0.22	-0.0017	-0.051	-0.003	0.023	-0.058	-0.037	-0.036	-0.034	-0.033	-0.046	0.17	0.16	0.18	0.16	0.14	0.31	0.15	0.18	0.16	0.15			
PAY_AMT_APR	0.003	0.22	-0.0028	-0.056	-0.0084	0.019	-0.059	-0.037	-0.036	-0.027	-0.023	-0.025	0.18	0.17	0.18	0.18	0.16	0.12	0.19	0.16	0.16	0.16	0.15		
default payment next month	-0.014	-0.15	-0.04	0.066	-0.028	0.014	0.32	0.26	0.24	0.22	0.2	0.19	-0.02	-0.014	-0.014	-0.01	-0.0068	-0.0054	-0.073	-0.059	-0.056	-0.057	-0.055	-0.053	

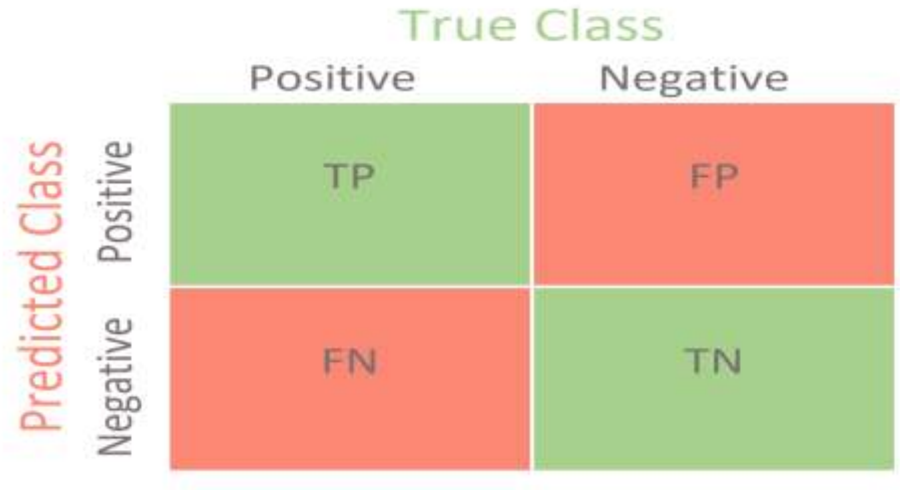
SMOTE(Synthetic Minority Oversampling Technique)





- Defaulters are less than non defaulters.
- We have solve the imbalance by SMOTE.
- We can see we have imbalanced dataset

Confusion Matrix



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

- A confusion matrix is a table that is used to define the performance of a classification algorithm.

- It is a table with 4 different combinations of predicted and actual values.

Modeling Overview

Models Used

- Logistic regression
- XG Boost
- Random forest
- SVC

Modeling Approach

DATA PREPROCESSING

- Feature selection
- Feature engineering
- Train test split
- SMOTE over sampeling

DATA FITTING AND TUNING

- Start with default parameter
- Hyperparameter tuning
- Measure RUC and AUC on training data

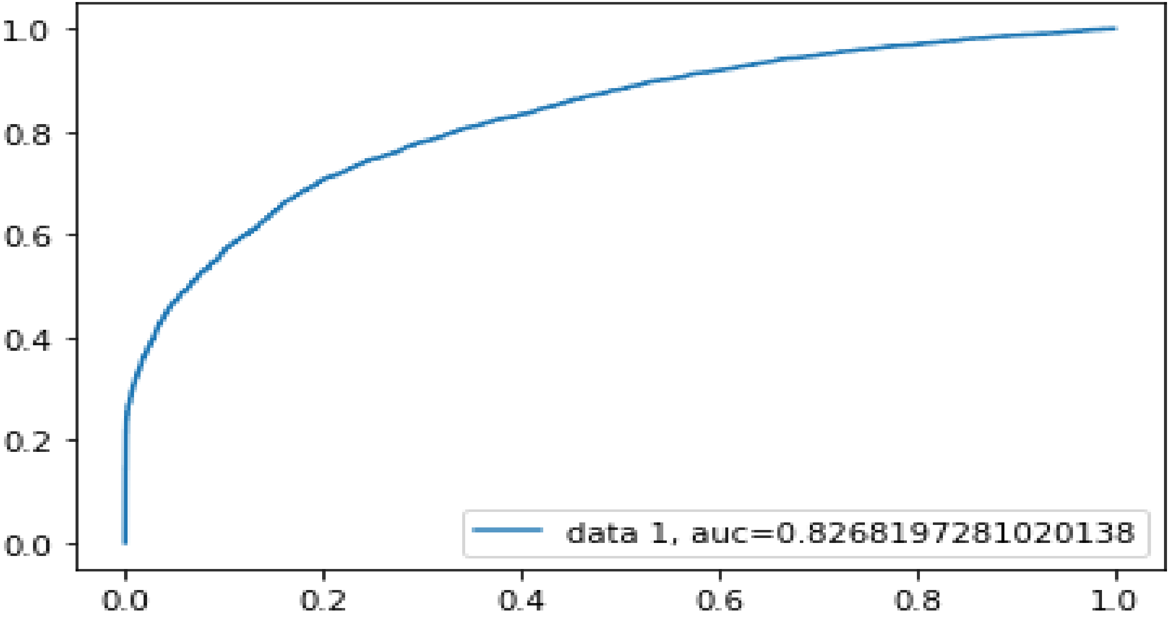
MODEL EVALUATION

- Model testing
- Precision score
- Recall score
- Model evaluation

Logistic Regression

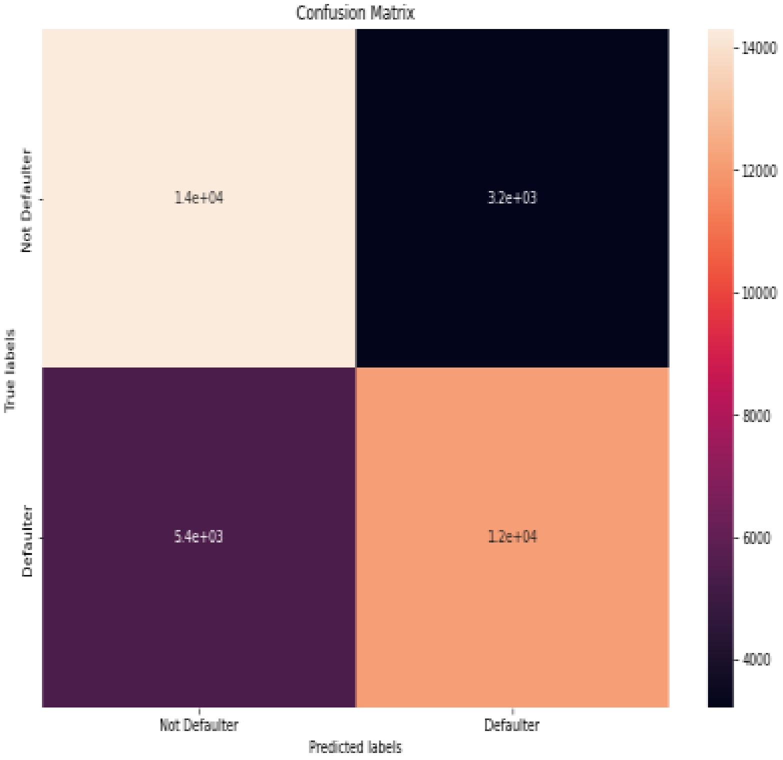
- Parameters
- C = 0.01
- Penalty = L2

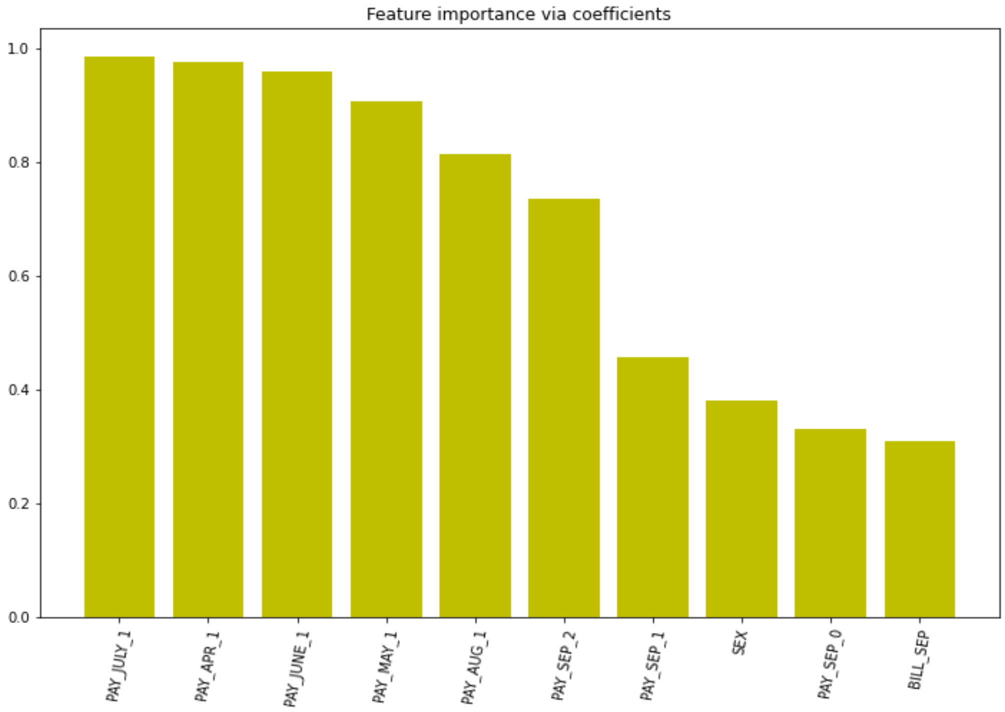
We have implemented logistic regression and we getting f1_sore approx 73.4%. As we have imbalanced dataset, F1 score is better parameter.



- The accuracy on test data is 0.751.
- The precision on test data is 0.687.
- The recall on test data is 0.788.
- The f1 score on test data is 0.734.
- The roc score on test data is 0.755.

Logistic Regression(Cont.)





- [[14305 3218]
- [5415 12108]]

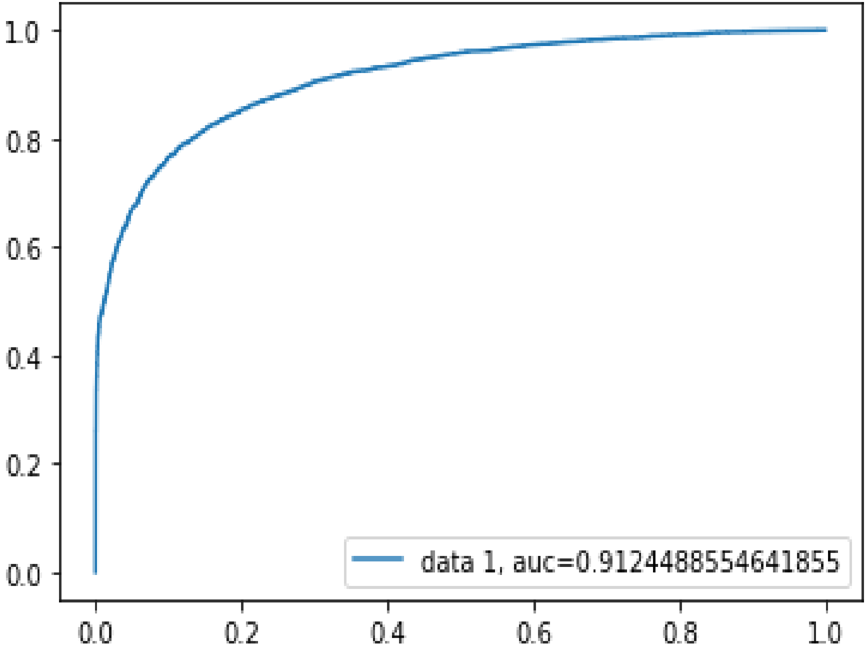
XG Boost Classifier

- **Parameters**

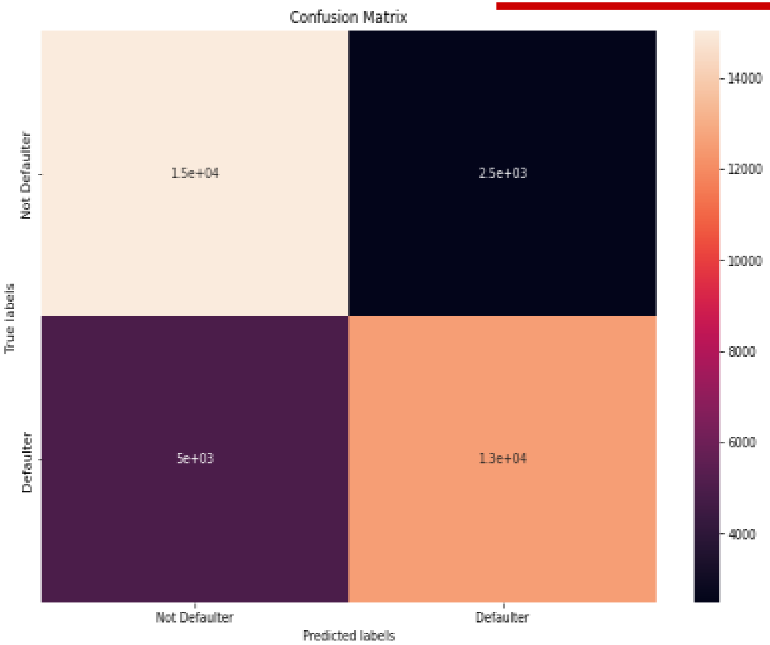
max_depth : 7

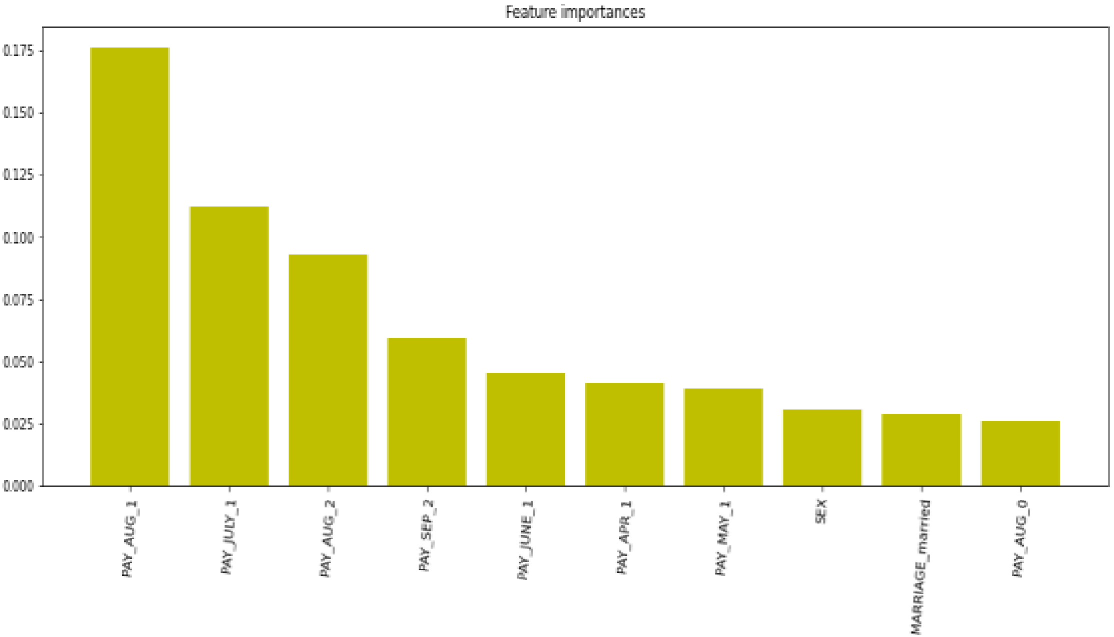
XG Boost is an implementation of gradient boosted decision trees
designed for speed and performance that is dominative competitive machine learning.

Accuracy on test data after hyperparameter tuning is 0.83.
Precision on test data after hyperparameter tuning is 0.795.
recall on test data after hyperparameter tuning is 0.86.
f1 score on test data after hyperparameter tuning is 0.82.
roc score on test data after hyperparameter tuning is 0.83.



XGBoost Classifier(Cont.)



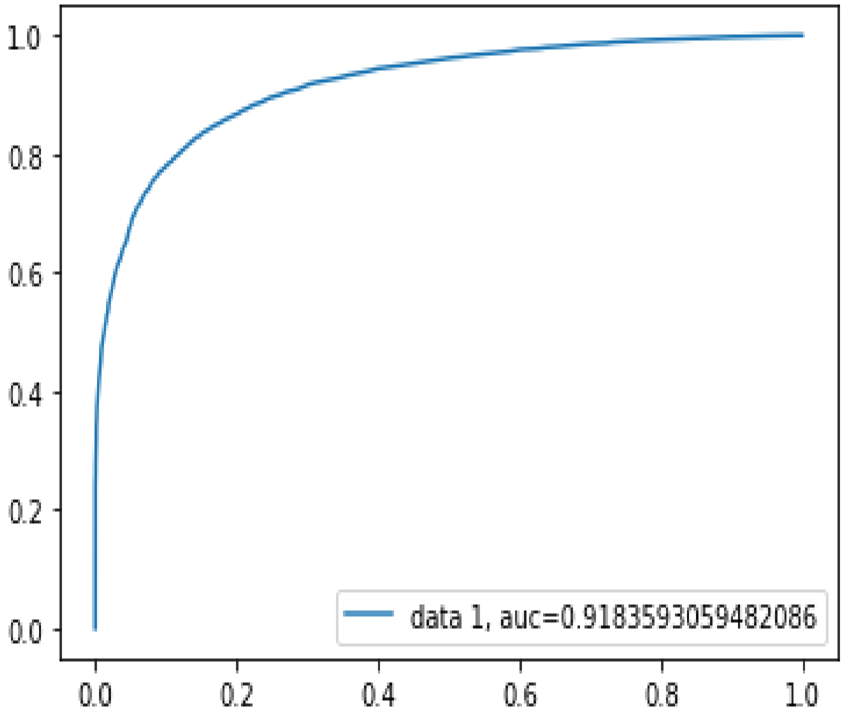


```
[[15037 2486]
[ 4963 12560]]
```

Random Forest

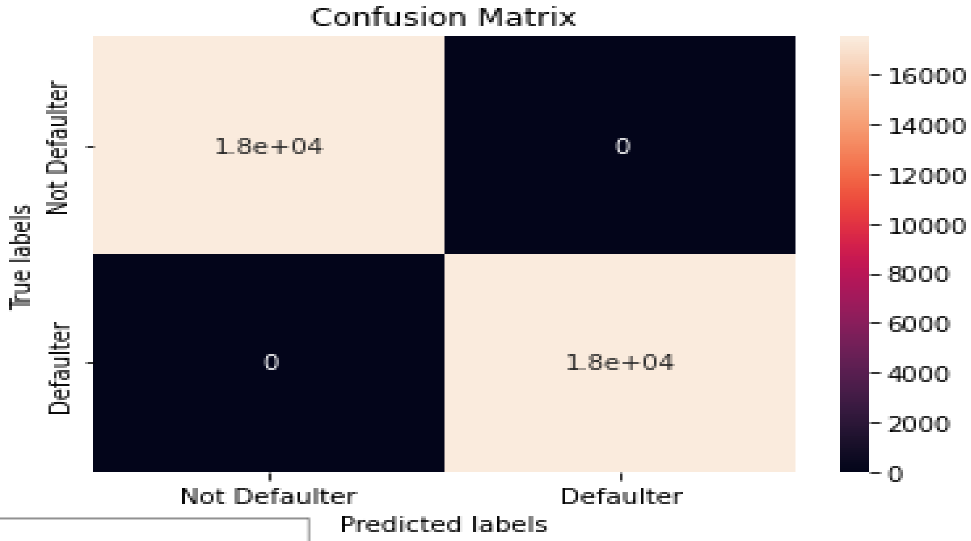
- **Parameters**
- Max_depth = 20
- N_estimators = 200

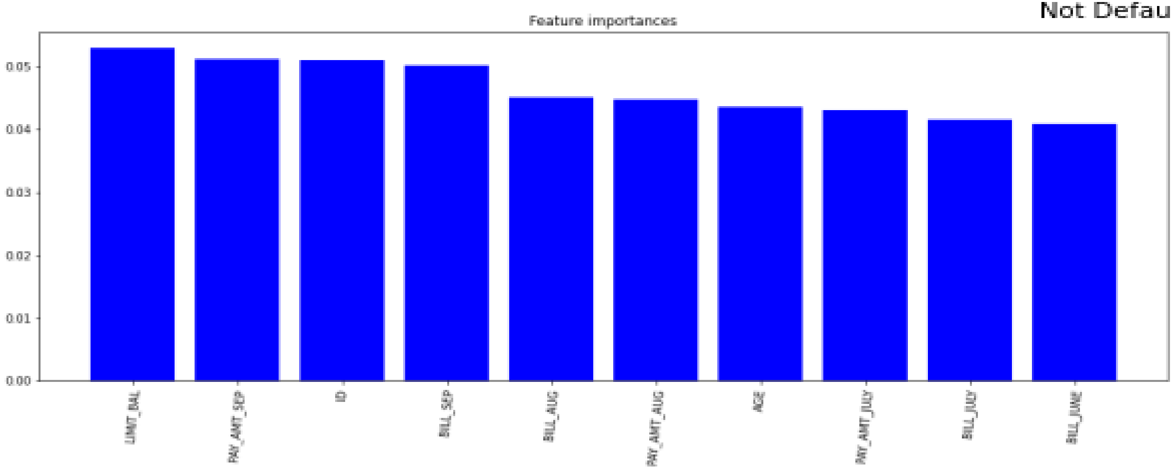
- The accuracy after hyperparameter tuning is 0.84.
- The precision after hyperparameter tuning is 0.81.
- The recall after hyperparameter tuning is 0.86.
- The f1 score after hyperparameter tuning is 0.83.
- The roc score after hyperparameter tuning is 0.84.



Random Forest(Cont.)

```
[[17523 0]
[ 0 17523]]
```





Support Vector Classifier

- **Parameters**
- C = 5
- Kernel = rbf

- A support vector machine (SVM) is a type of deep learning algorithm that performs supervised learning for classification or regression of data groups. In AI and machine learning, supervised learning systems provide both input and desired output data, which are labeled for classification.

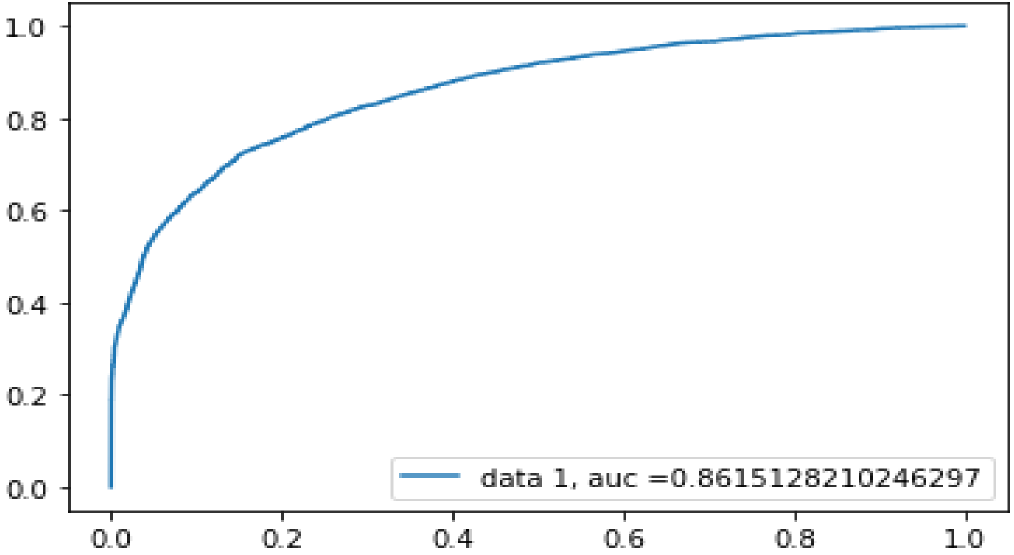
The accuracy score on test data is 0.78.

The precision score on test data is 0.71.

The recall on test data is 0.82.

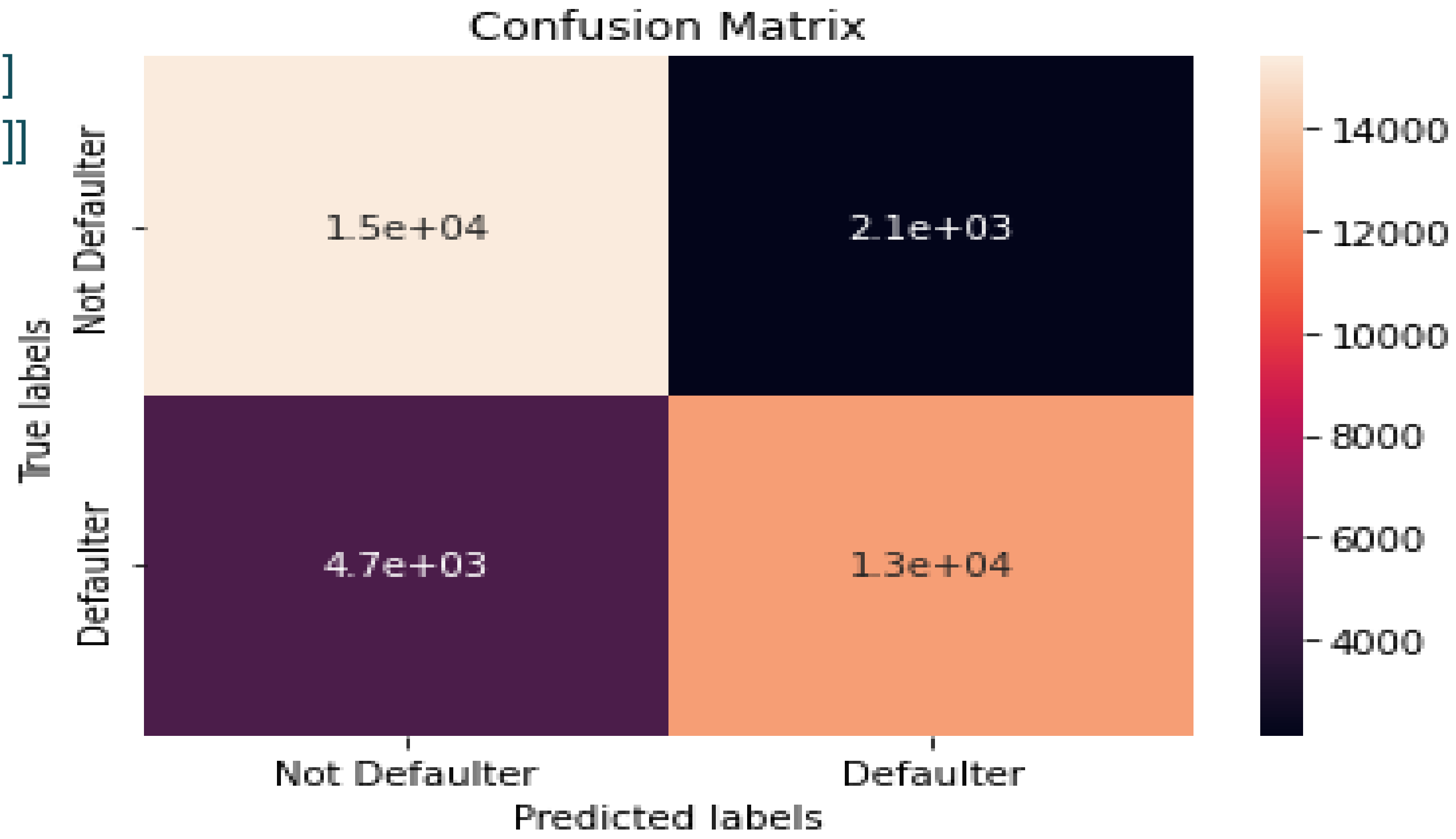
The f1 score on test data is 0.76.

The roc score on tes data is 0.78.

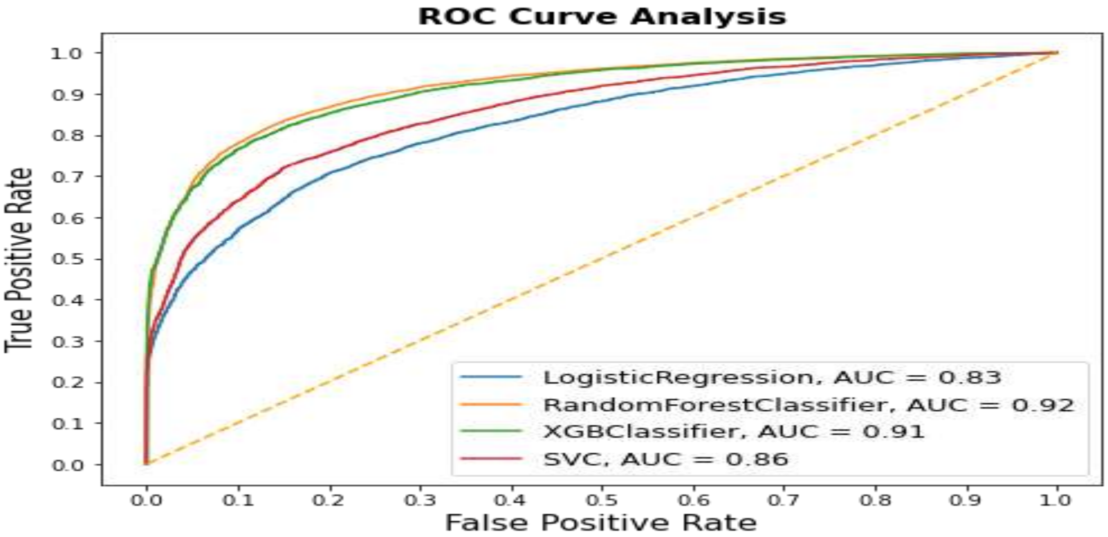


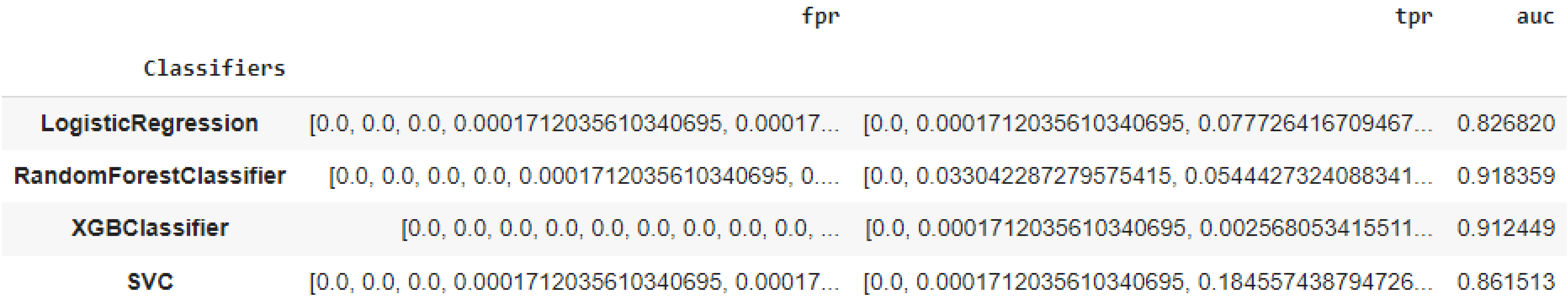
Support Vector Classifier(Cont.)

```
[[15386 2137]
[ 4731 12792]]
```



Plotting ROC AUC for all the models

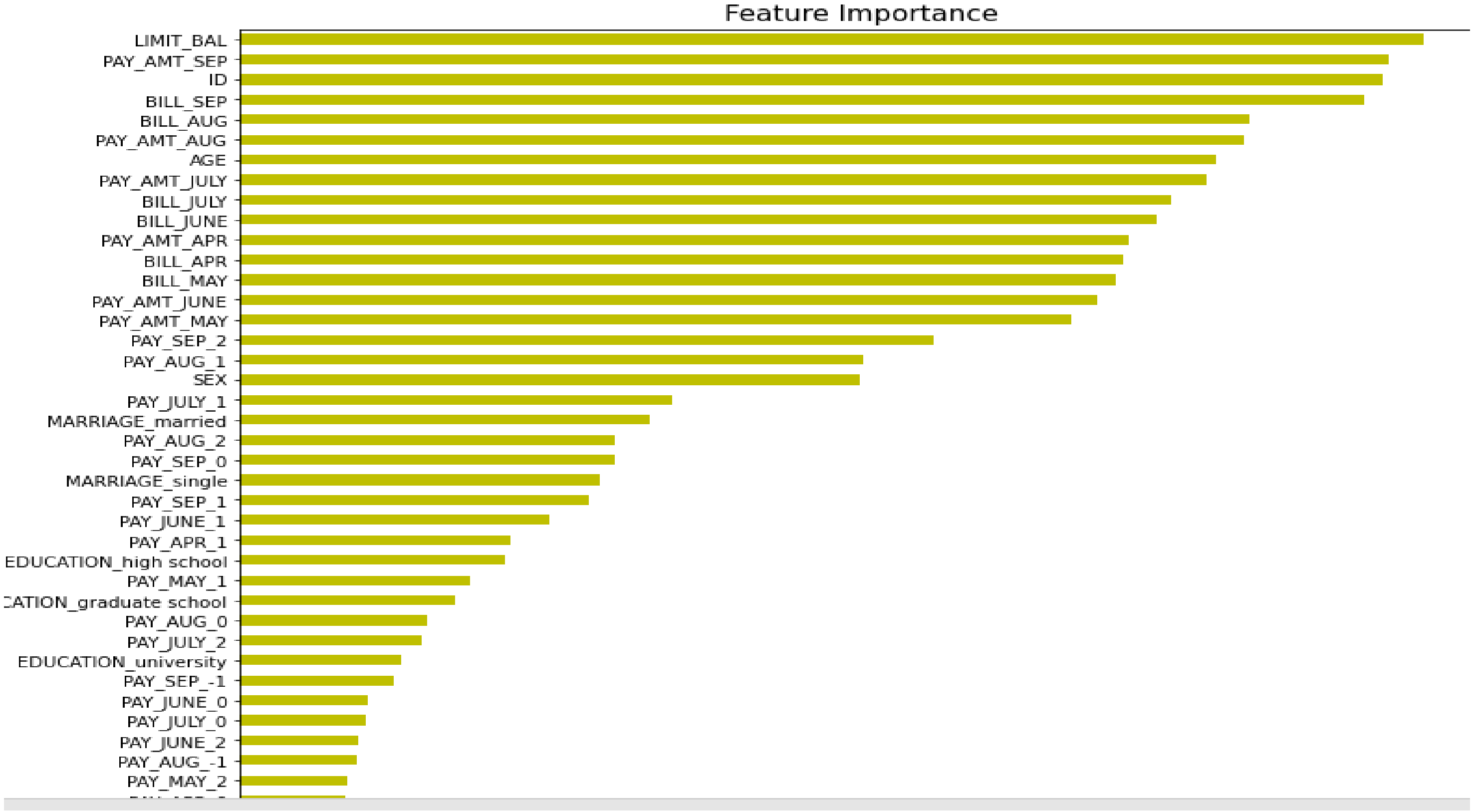




	fpr	tpr	auc
Classifiers			
LogisticRegression	[0.0, 0.0, 0.0, 0.0001712035610340695, 0.00017...	[0.0, 0.0001712035610340695, 0.077726416709467...	0.826820
RandomForestClassifier	[0.0, 0.0, 0.0, 0.0, 0.0001712035610340695, 0....	[0.0, 0.033042287279575415, 0.0544427324088341...	0.918359
XGBClassifier	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[0.0, 0.0001712035610340695, 0.002568053415511...	0.912449
SVC	[0.0, 0.0, 0.0, 0.0001712035610340695, 0.00017...	[0.0, 0.0001712035610340695, 0.184557438794726...	0.861513

Feature Importance for recommended model

- "LIMT_BAL","BILL_SEP" AND "PAY_AMT_SEP" are the most recent 2 months' payment status and they are the strongest predictors of future payment default risk.

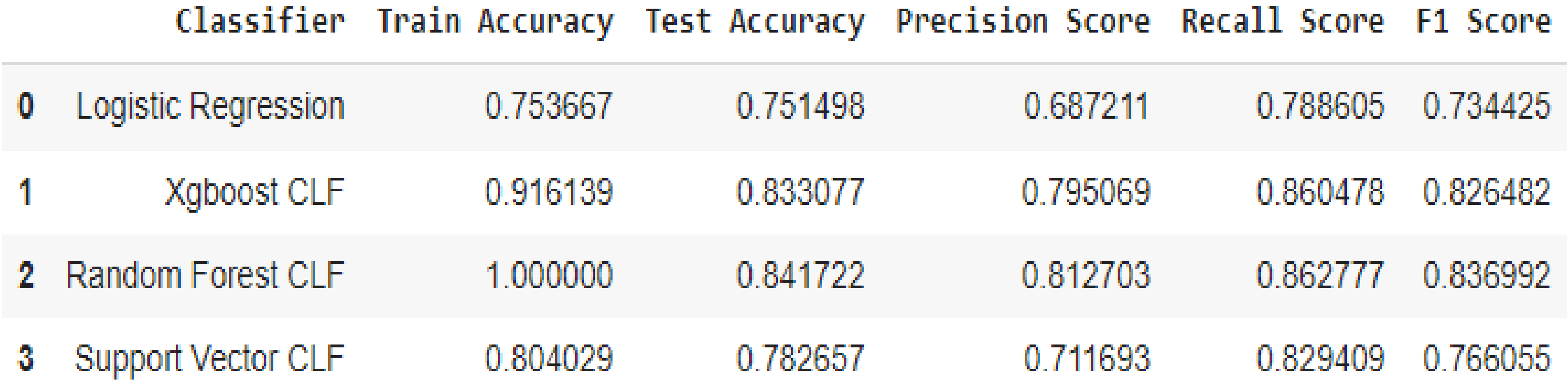


Challenges

- Data Cleaning
- Data mining
- Feature Engineering
- Feature Selection
- Model optimization
- Hyperparameter Tuning
- Deciding the flow of presentation

Overall Conclusion

Random Forest model and XG Boost model both has same recall, so if the business cares recall the most than both of this model are best candidate. If the balance of recall and precision is most important metric than Random Forest is the ideal model. Random Forest has recall and precision both higher than the other model applied. Hence, I would recommend Random Forest for this dataset.



	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.753667	0.751498	0.687211	0.788605	0.734425
1	Xgboost CLF	0.916139	0.833077	0.795069	0.860478	0.826482
2	Random Forest CLF	1.000000	0.841722	0.812703	0.862777	0.836992
3	Support Vector CLF	0.804029	0.782657	0.711693	0.829409	0.766055

Overall Conclusion

There were not huge gap but female clients tended to default the most.

- Labels of the data were imbalanced and had a significant difference.
- Gradient boost gave the highest accuracy of 82% on test dataset.
- Repayment in the month of September tended to be most important feature for our machine learning model.
- The best accuracy is obtained for the Random forest and XGBoost classifier.
- Data categorical variables had minority classes which were added to their closest majority class.
- From above table we can see that XGBoost Classifier having Recall = 86%, F1-score = 82%, and ROC Score= 83% and Random forest Classifier having Recall =86%, F1-score = 82% and ROC Score = 84%.
- XGBoost Classifier and Random Forest Classifier are giving us the best Recall, F1-score, and ROC Score among other algorithms. We can conclude that these two algorithms are the best to predict whether the credit card is default or not default according to our analysis on this dataset.

THANK YOU