Capstone Project : 3

Credit Card Default Prediction

By

Mayank Mishra

Points for Discussion

- Understanding Problem Statement
- Data Set Information
- Feature Summary
- Approach Overview
- Data Preprocessing
- Exploratory Data Analysis
- Implementing Algorithm
- Challenges
- Conclusion
- Q&A

Understanding Problem Statement

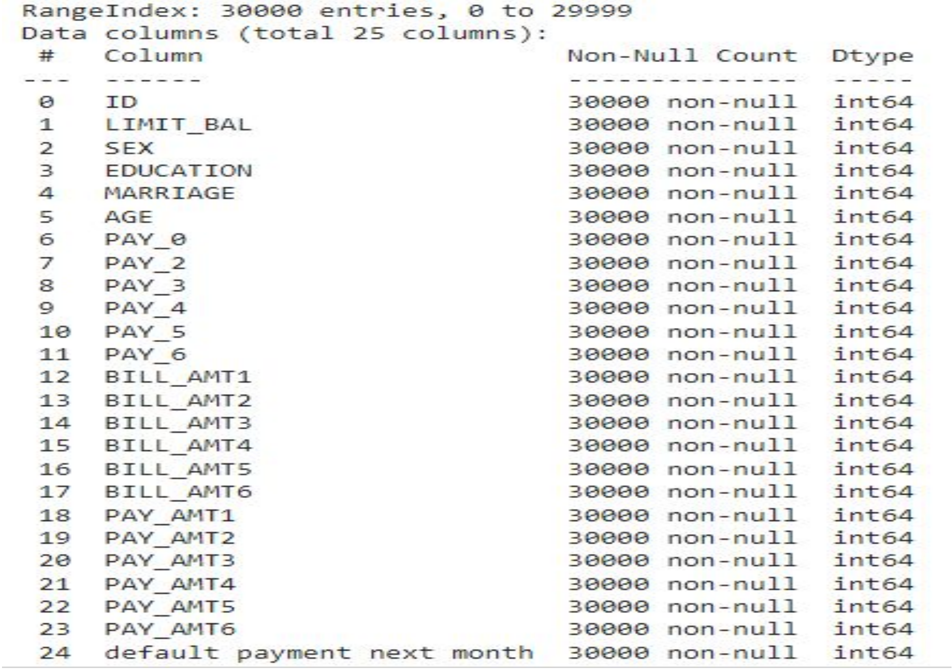
- Topic – “Credit Card Default Prediction”
- Problem Statement -- Explore and Analyze the the prediction whether a customer will default on his/her Credit Card.

- Financial threats are displaying a trend about the risk of commercial banks as the improvement in the financial industry has arisen. In this way one of the biggest threats faced by them is the risk prediction of credit card clients.

- The goal of this project is to predict the credit card defaulter comprising variables like AGE, SEX, EDUCATION, LIMIT BAL, MARRIAGE and others.

Data Set Information

- This dataset contains 30000 observations and 23 features of six months.
- There are nine categorical features in our dataset.
- This dataset is from the city of Taiwan and doesn’t have any null or duplicate values.



```
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          30000 non-null  int64
 1   LIMIT_BAL                   30000 non-null  int64
 2   SEX                         30000 non-null  int64
 3   EDUCATION                   30000 non-null  int64
 4   MARRIAGE                    30000 non-null  int64
 5   AGE                         30000 non-null  int64
 6   PAY_0                       30000 non-null  int64
 7   PAY_2                       30000 non-null  int64
 8   PAY_3                       30000 non-null  int64
 9   PAY_4                       30000 non-null  int64
 10  PAY_5                       30000 non-null  int64
 11  PAY_6                       30000 non-null  int64
 12  BILL_AMT1                   30000 non-null  int64
 13  BILL_AMT2                   30000 non-null  int64
 14  BILL_AMT3                   30000 non-null  int64
 15  BILL_AMT4                   30000 non-null  int64
 16  BILL_AMT5                   30000 non-null  int64
 17  BILL_AMT6                   30000 non-null  int64
 18  PAY_AMT1                    30000 non-null  int64
 19  PAY_AMT2                    30000 non-null  int64
 20  PAY_AMT3                    30000 non-null  int64
 21  PAY_AMT4                    30000 non-null  int64
 22  PAY_AMT5                    30000 non-null  int64
 23  PAY_AMT6                    30000 non-null  int64
 24  default payment next month  30000 non-null  int64
```

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)

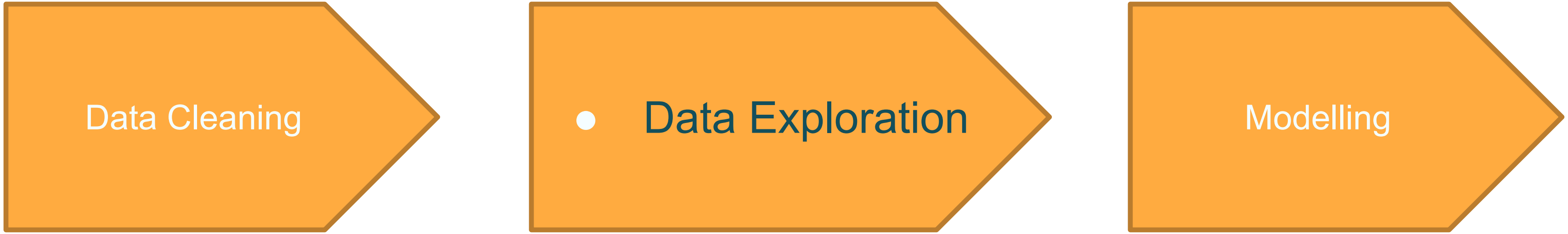
- X5: Age in year.

- 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 Se
- Y: Default payment

Approach Overview



Data Cleaning

- Data Exploration

Modelling

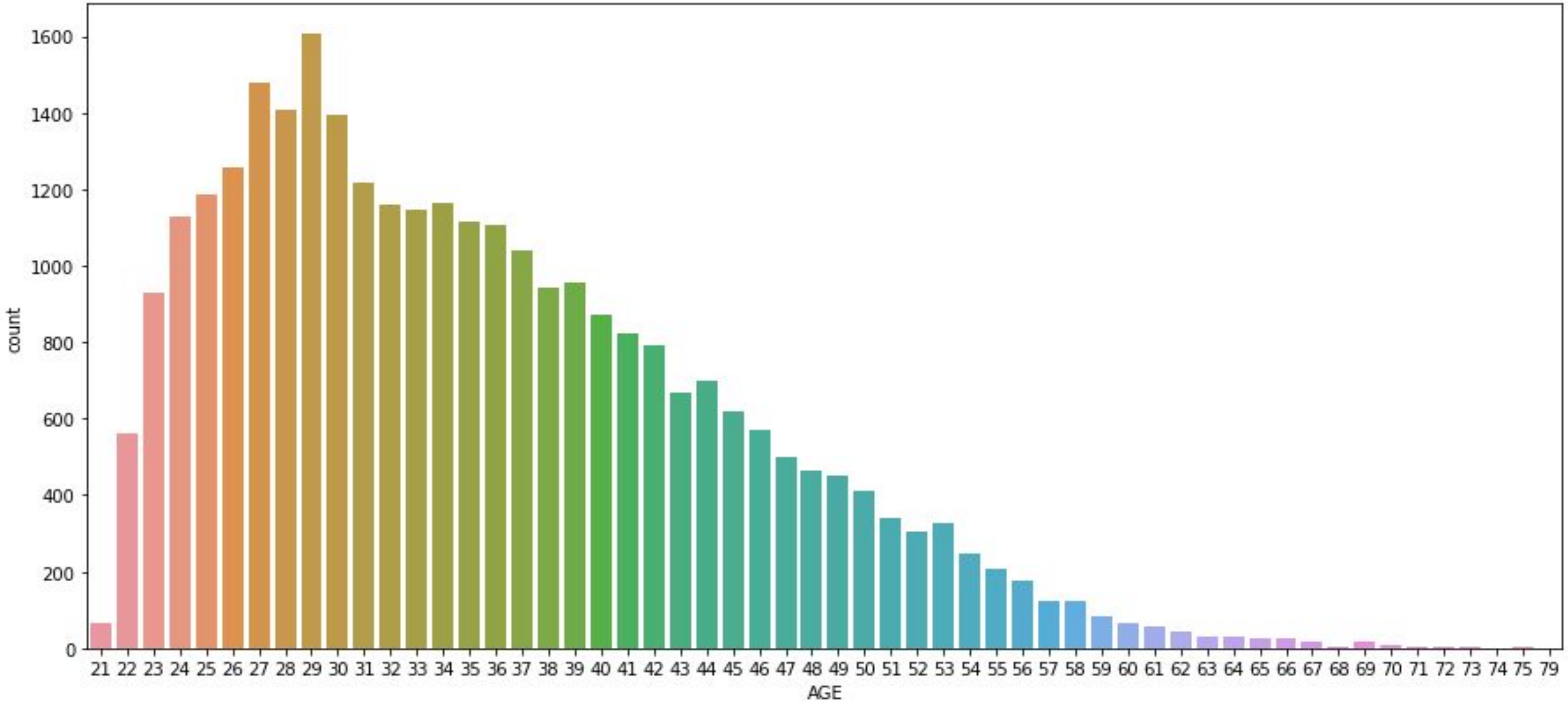
- Find information on documented columns values.
- Clean data for further analysis.

- Analyze the data with EDA

- Logistic Regression
- XGBoost
- Random Forest
- SVC

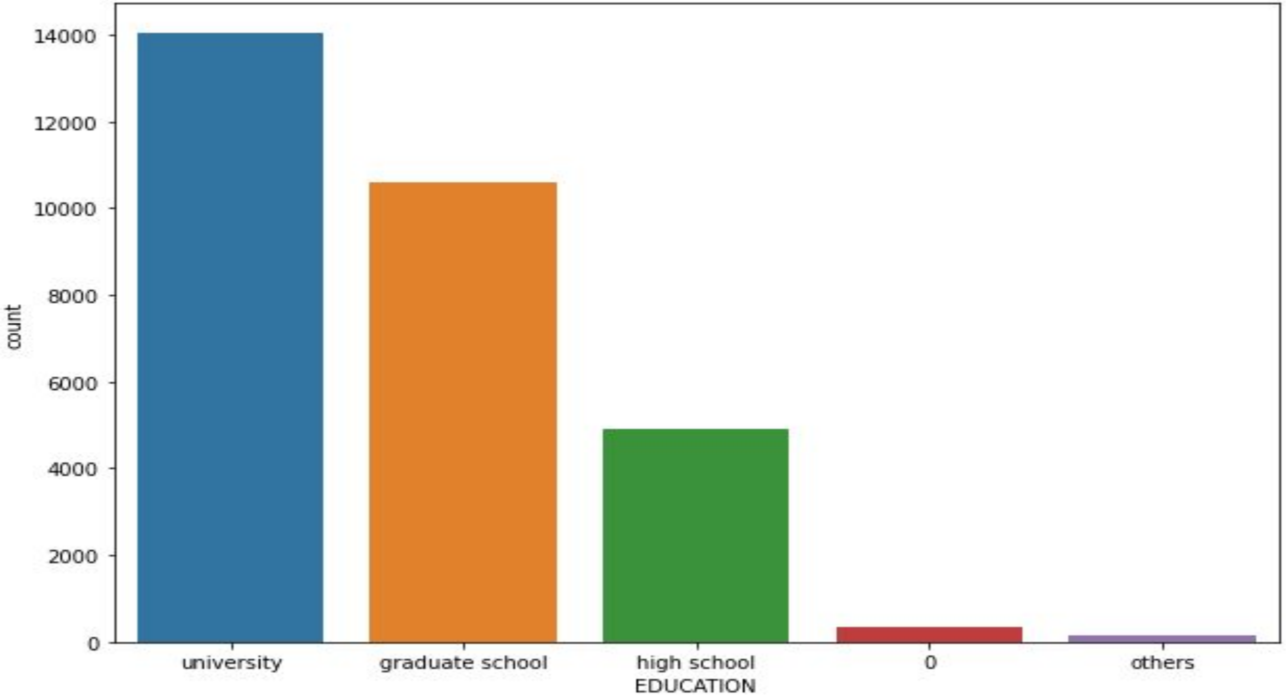
Count of credit card on basis of age

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



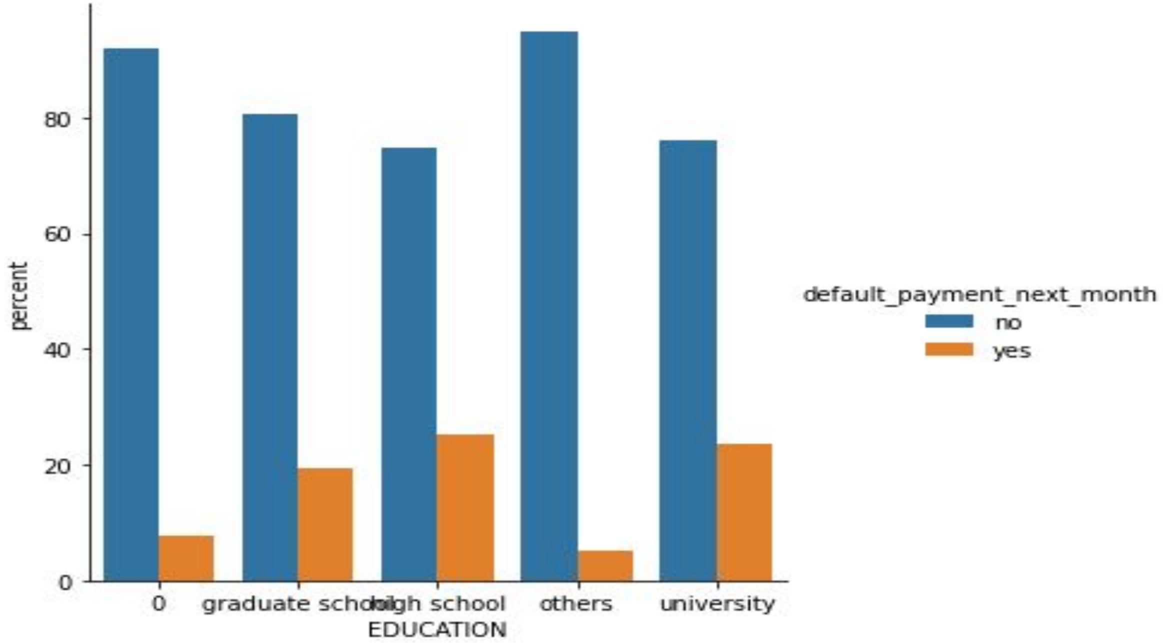
Count of credit card basis of eduation

- Most number of credit card are used by university students.

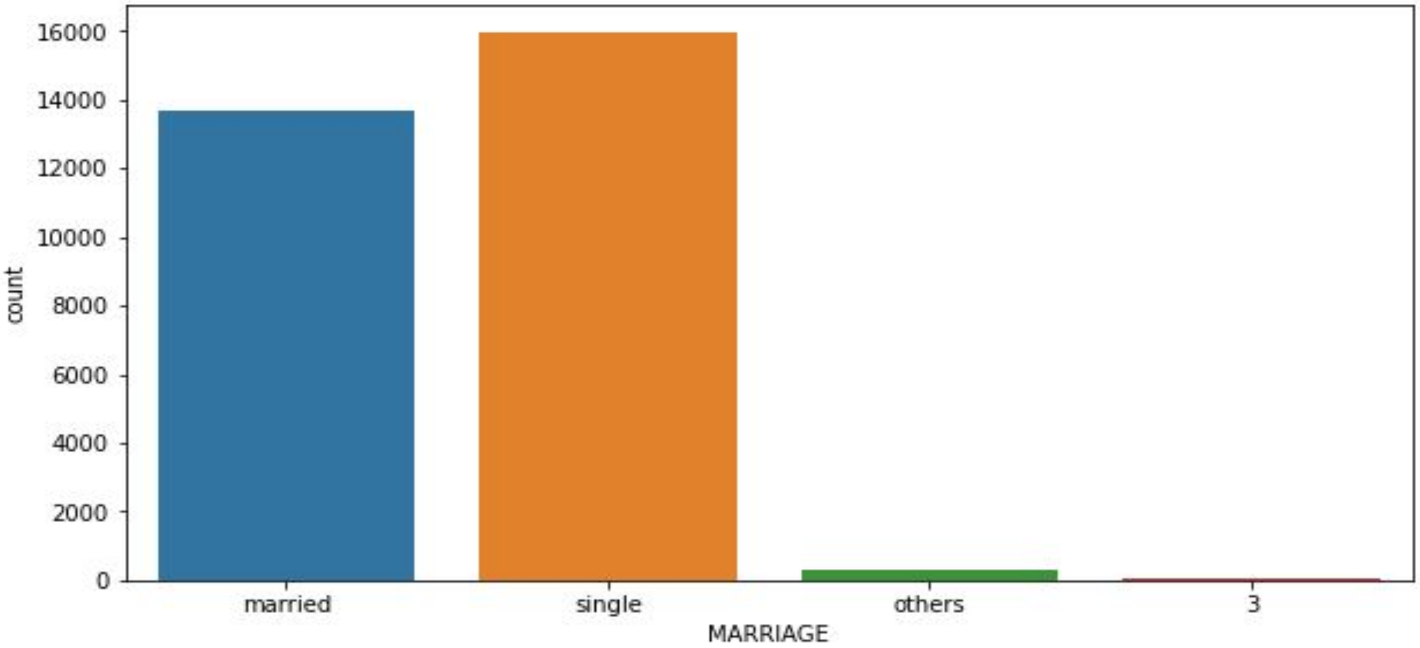


Credit card defaulter on basis of Education

- High school and university students has high default risk.



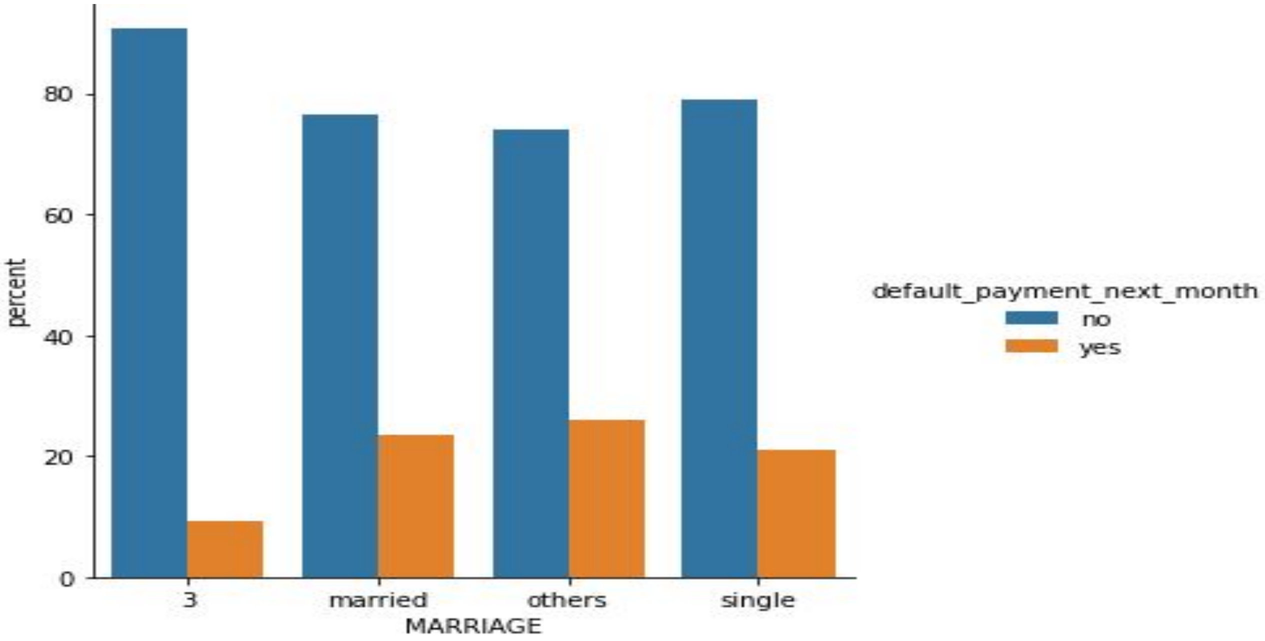
Count of credit card on basis of marital status



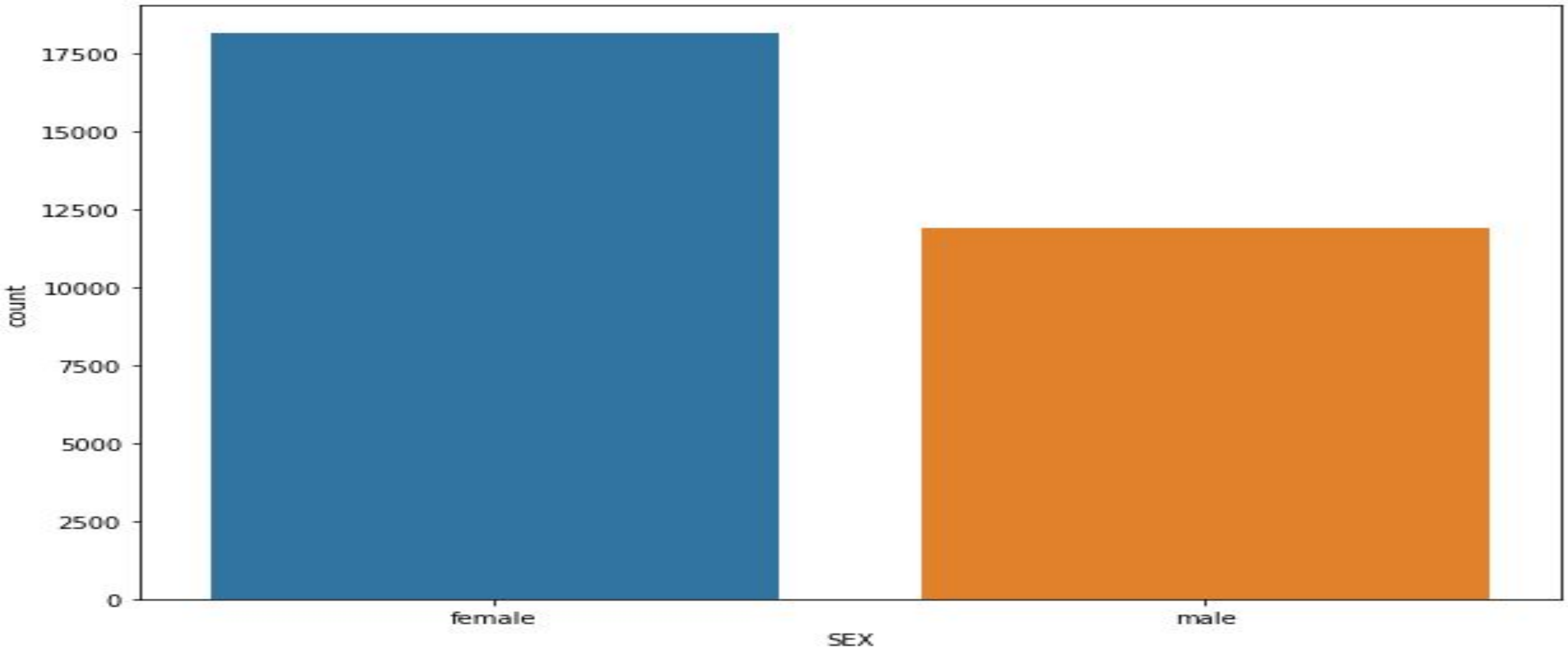
- Most number credit card are used by the people who are unmarried.

Credit card defaulter on basis of marital status

- No significant correlation of default risk and marital status.

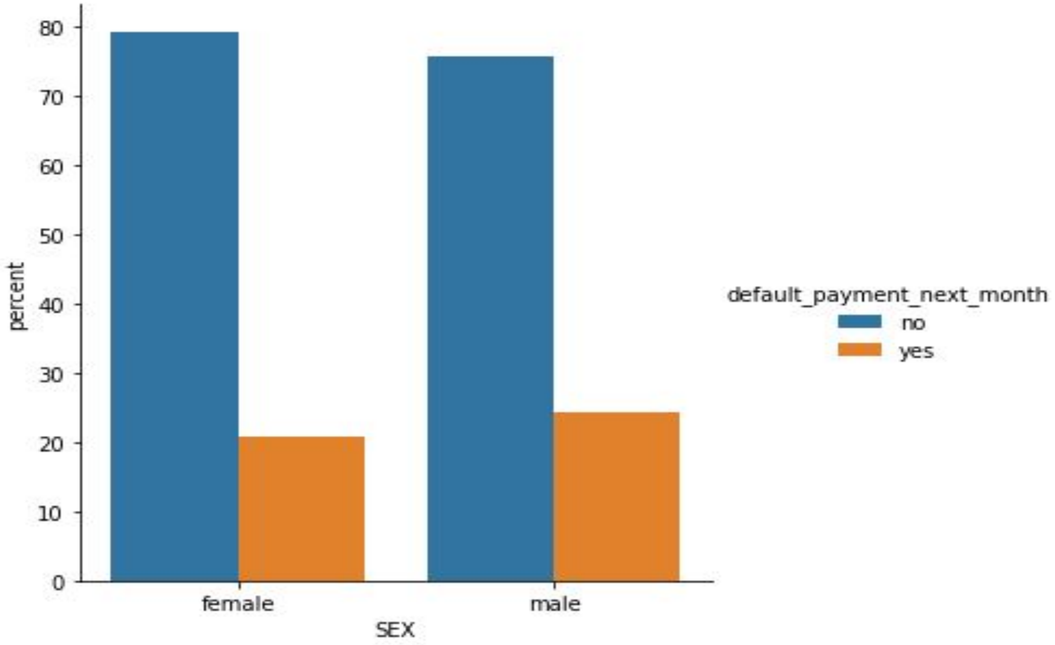


Count of credit card in basis of Gender



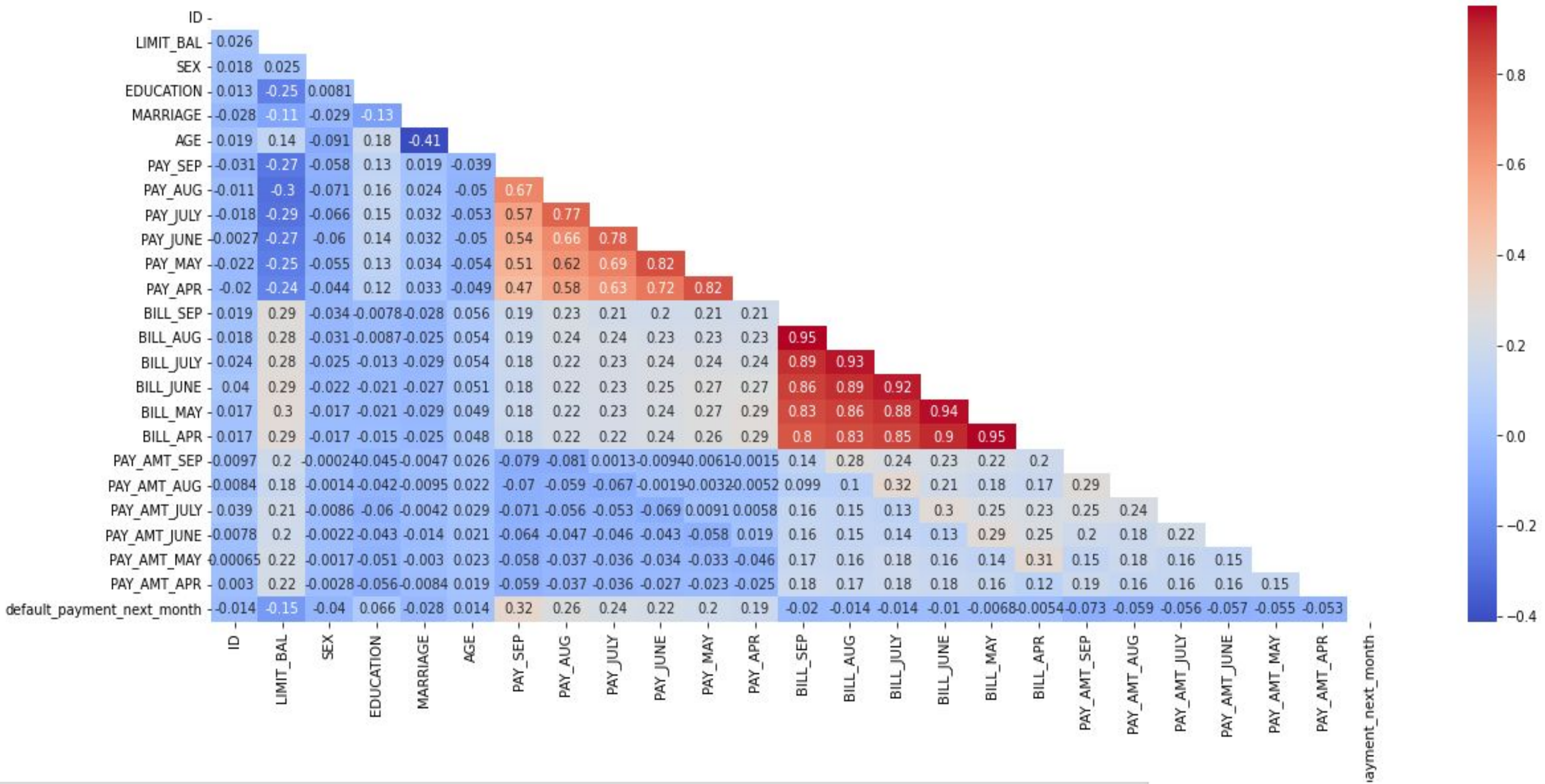
- No of female credit card holder are more as compared to male.

Credit card defaulter on basis of Gender



- Males credit card holder has more default as compared to females
- Males defaulter are around 25% and females defaulter are around 20%

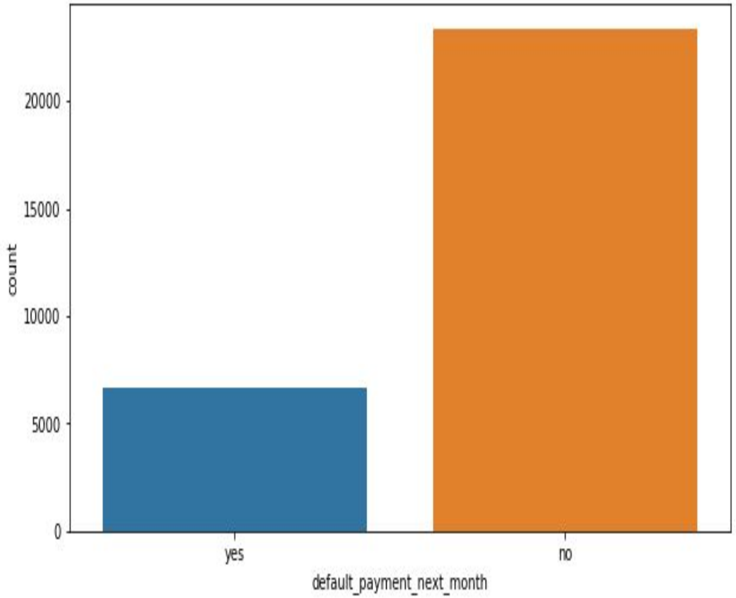
Correlation Matrix

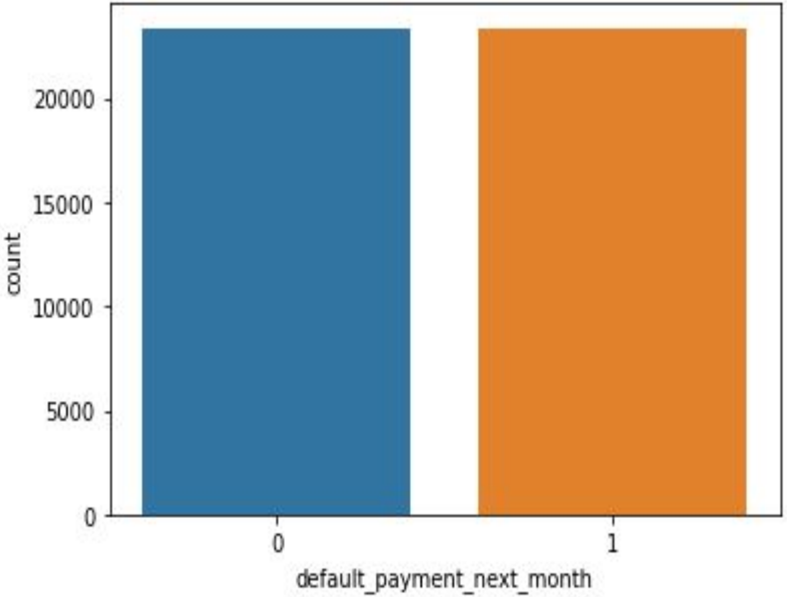


	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_SEP	PAY_AUG	PAY_JULY	PAY_JUNE	PAY_MAY	PAY_APR	BILL_SEP	BILL_AUG	BILL_JULY	BILL_JUNE	BILL_MAY	BILL_APR	PAY_AMT_SEP	PAY_AMT_AUG	PAY_AMT_JULY	PAY_AMT_JUNE	PAY_AMT_MAY	PAY_AMT_APR	ayment_next_month
ID																									
LIMIT_BAL	0.026																								
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)

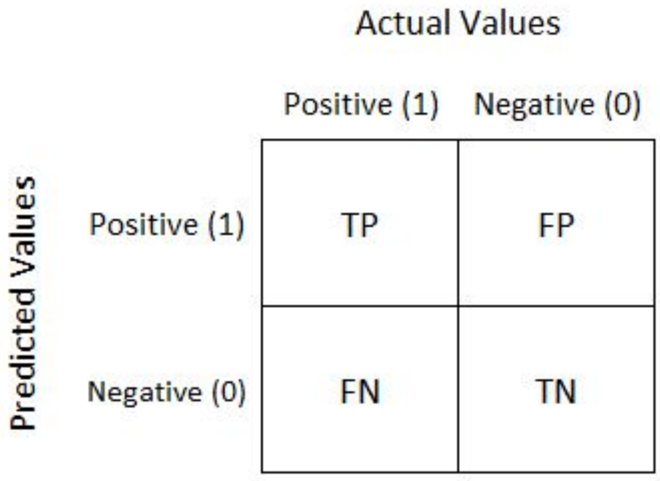
- We can see we have imbalanced dataset.
- Defaulters are less than non defaulters.
- We have solve the imbalance by SMOTE.





Confusion Matrix

- Confusion matrix is a performance measurement for machine learning classification problem where output can be two or more classes.
- It is a table with 4 different combinations of predicted and actual values.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

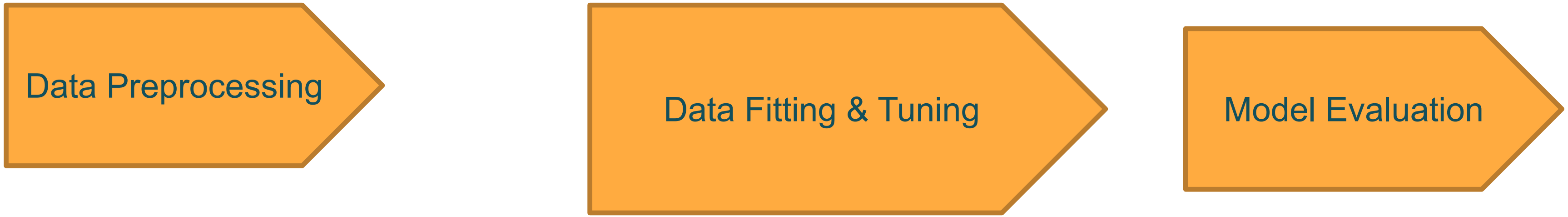
Modeling Overview

- Supervised learning / Binary classification
- Imbalance data with 78% non defaulters and 22% defaulters.

- Models Used

- Logistic Regression
- XGBoost CLF
- Random Forest CLF
- Support Vector Classifier (SVC)

Modelling Approach



Data Preprocessing

Data Fitting & Tuning

Model Evaluation

- Feature Selection
- Feature engineering
- Train Test split
- SMOTE Oversampling

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

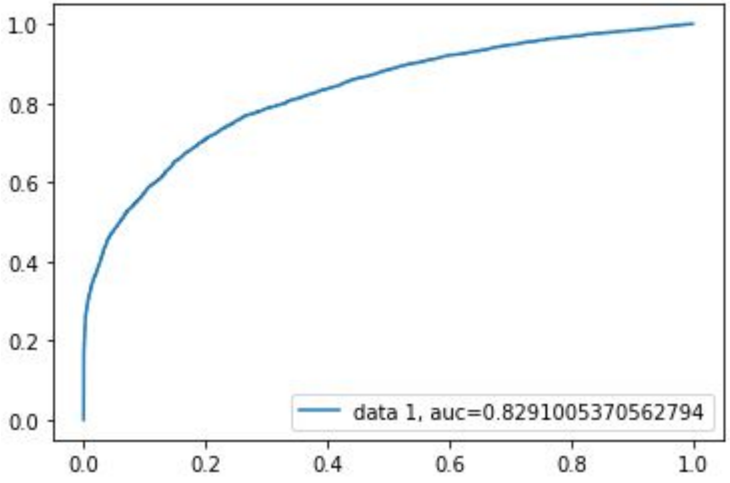
- Model testing
- Precision score
- Recall score
- Model Evaluation

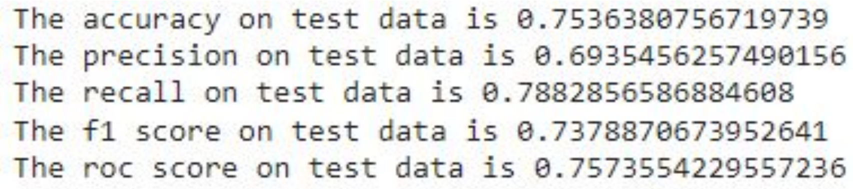
Logistic Regression

- **Parameters**

- C = 0.01
- Penalty = L2

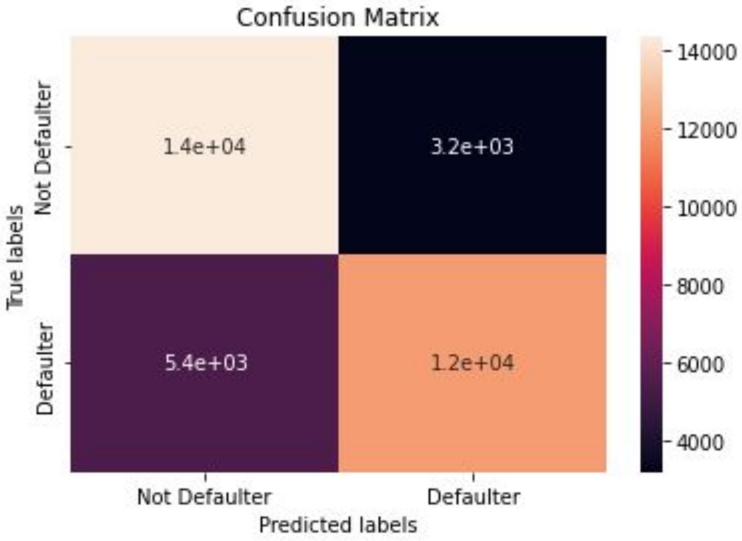
- We have implemented logistic regression.
- we getting f1_sore approx. 73%. As we have imbalanced dataset, F1- score is better parameter

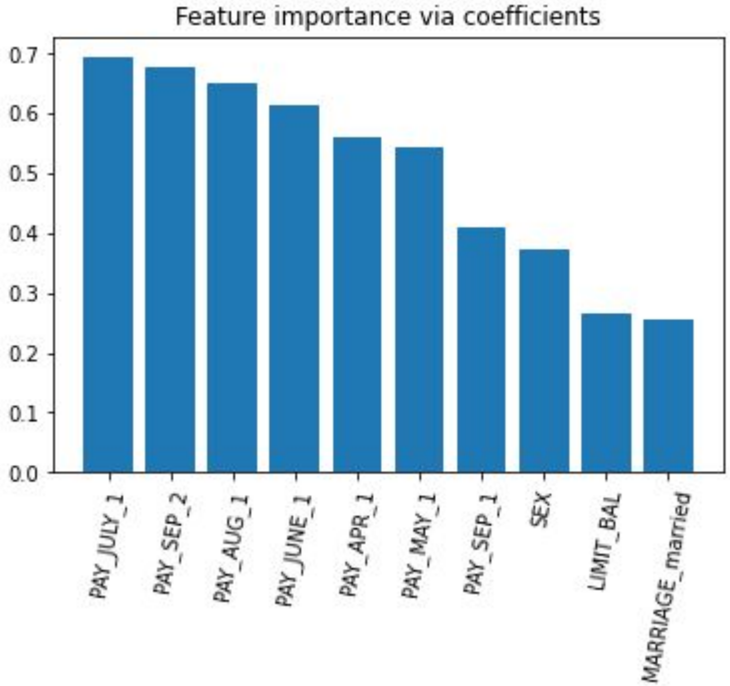


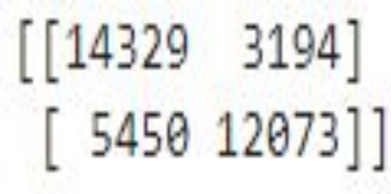


```
The accuracy on test data is 0.7536380756719739
The precision on test data is 0.6935456257490156
The recall on test data is 0.7882856586884608
The f1 score on test data is 0.7378870673952641
The roc score on test data is 0.7573554229557236
```

Logistic Regression(Cont.)







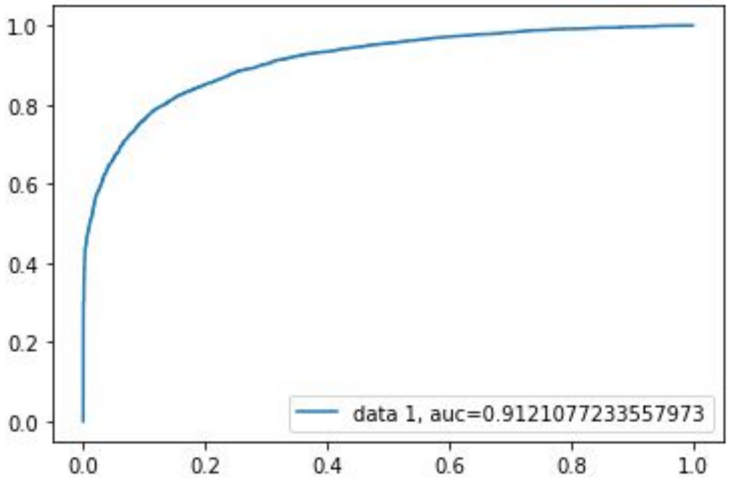
```
[[14329  3194]
 [ 5450 12073]]
```

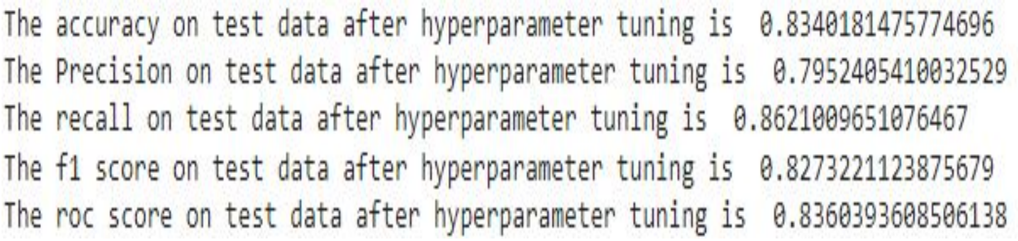
XGBoost Classifier

- **Parameters**

  - max_depth = 10
  - min_child_weight =6

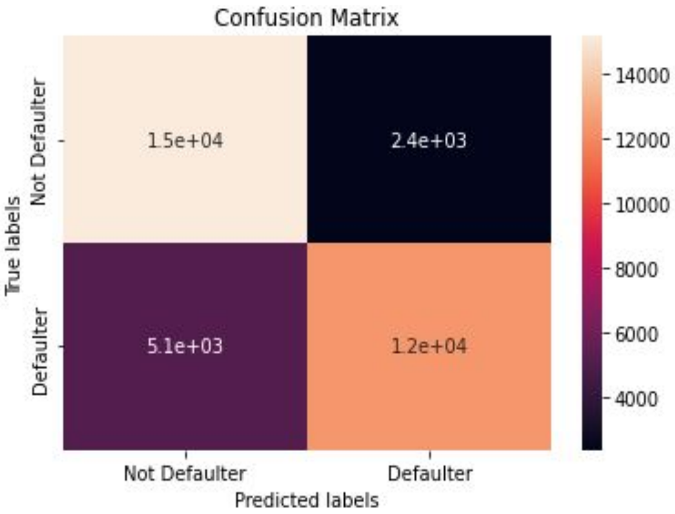
- The XGBoost model for classification is called XGBClassifier. We can create and fit it to our training dataset. Models are fit using the scikit-learn API and the model.fit() function.

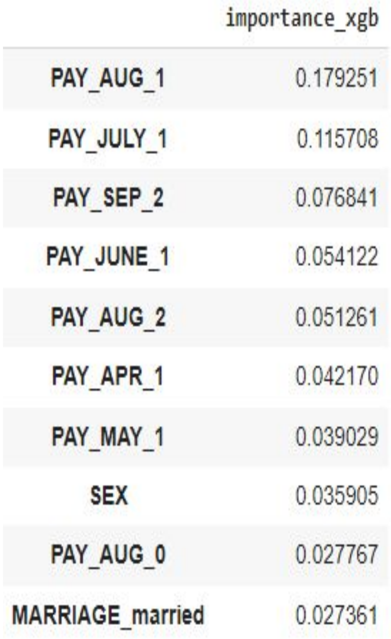




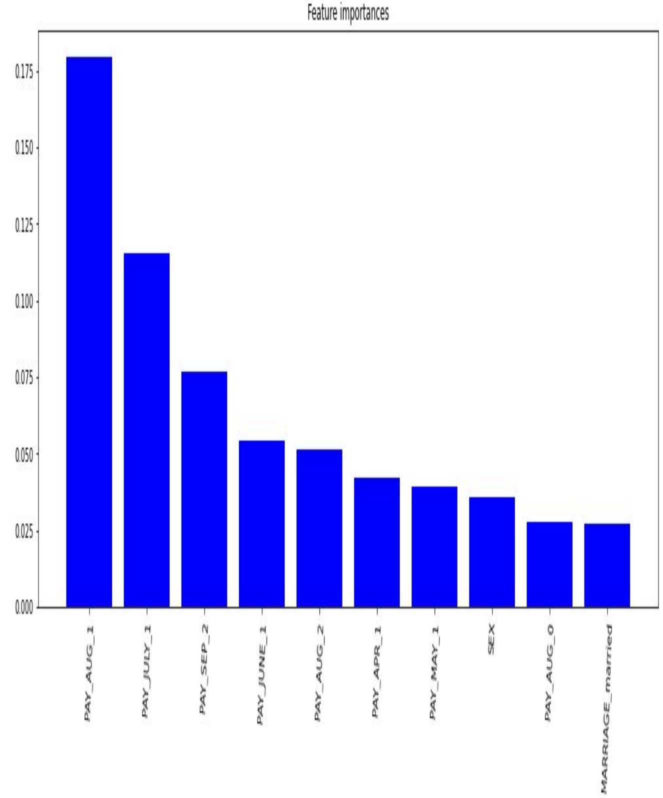
```
The accuracy on test data after hyperparameter tuning is  0.8340181475774696
The Precision on test data after hyperparameter tuning is  0.7952405410032529
The recall on test data after hyperparameter tuning is  0.8621009651076467
The f1 score on test data after hyperparameter tuning is  0.8273221123875679
The roc score on test data after hyperparameter tuning is  0.8360393608506138
```

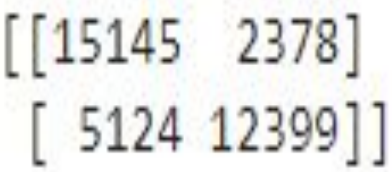
XGBoost Classifier(Cont.)





	importance_xgb
PAY_AUG_1	0.179251
PAY_JULY_1	0.115708
PAY_SEP_2	0.076841
PAY_JUNE_1	0.054122
PAY_AUG_2	0.051261
PAY_APR_1	0.042170
PAY_MAY_1	0.039029
SEX	0.035905
PAY_AUG_0	0.027767
MARRIAGE_married	0.027361



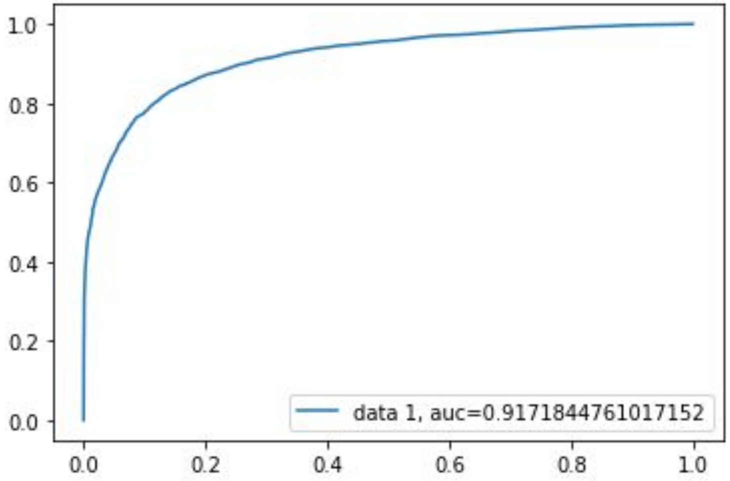


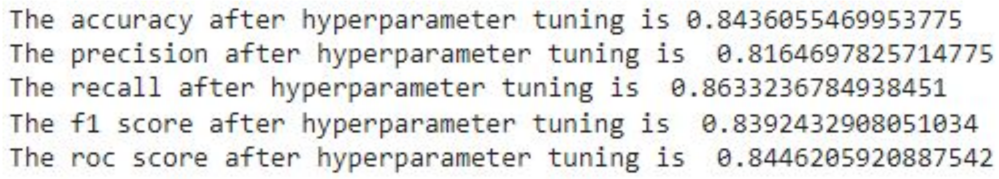
```
[[15145  2378]
 [ 5124 12399]]
```

Random Forest

- **Parameters**

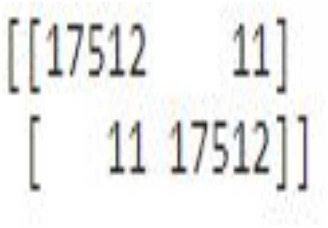
- Max_depth = 40
- N_estimators = 250



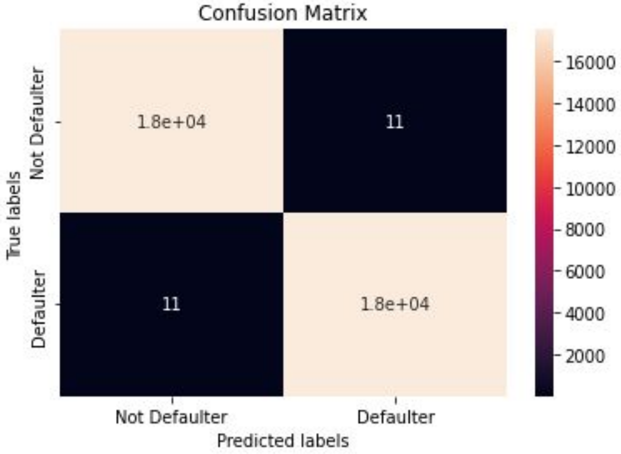


```
The accuracy after hyperparameter tuning is 0.8436055469953775
The precision after hyperparameter tuning is  0.8164697825714775
The recall after hyperparameter tuning is  0.8633236784938451
The f1 score after hyperparameter tuning is  0.8392432908051034
The roc score after hyperparameter tuning is  0.8446205920887542
```

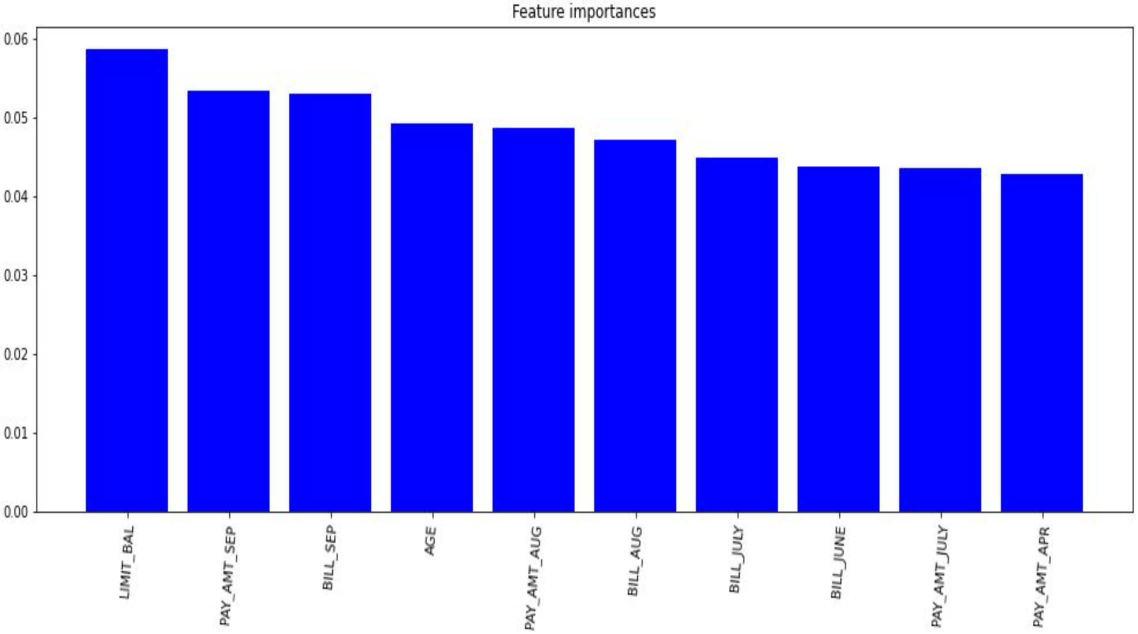
Random Forest(Cont.)



```
[[17512    11]
 [   11 17512]]
```



Confusion Matrix



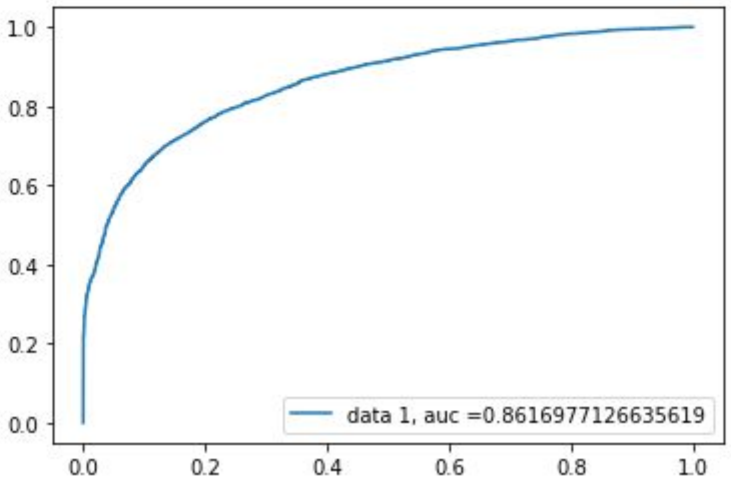
Feature importances

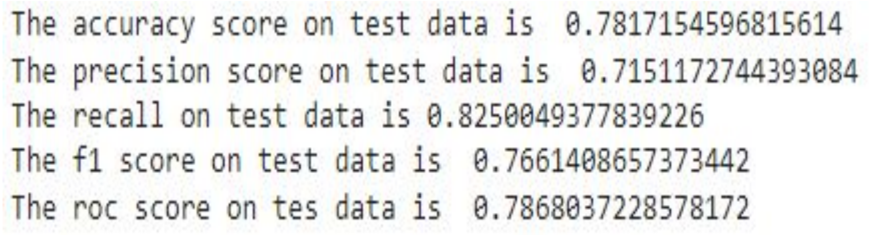
Support Vector Classifier

- **Parameters**

□ **C = 5**

□ **Kernel = ‘rbf’**

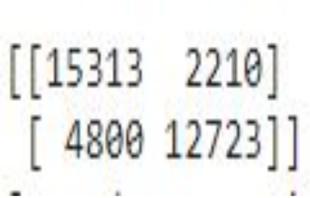




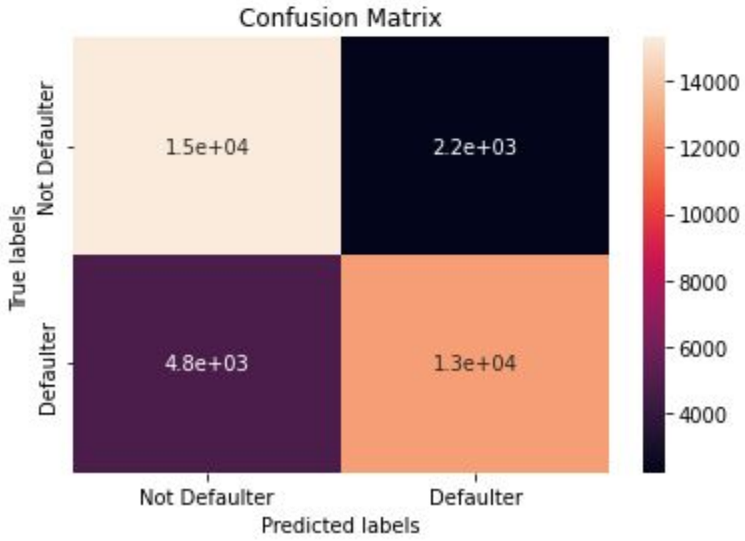
```
The accuracy score on test data is  0.7817154596815614
The precision score on test data is  0.7151172744393084
The recall on test data is 0.8250049377839226
The f1 score on test data is  0.7661408657373442
The roc score on tes data is  0.7868037228578172
```

Support Vector Classifier(Cont.)

- **Confusion Matrix**

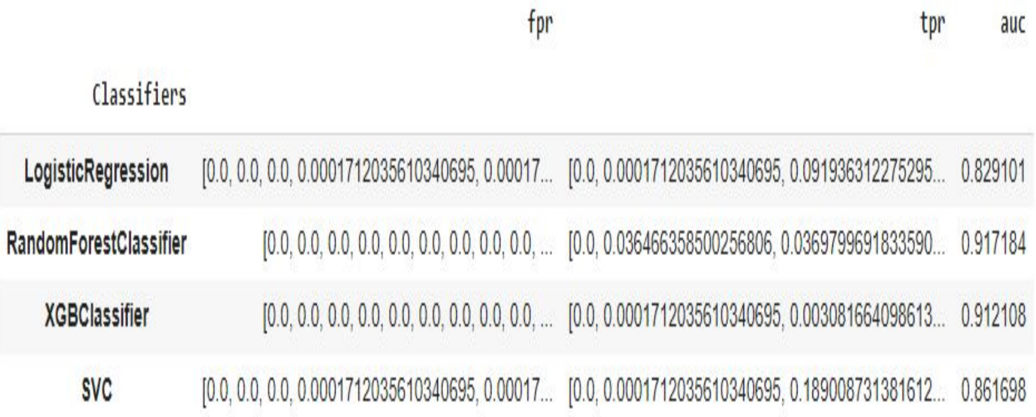


```
[[15313  2210]
 [ 4800 12723]]
```

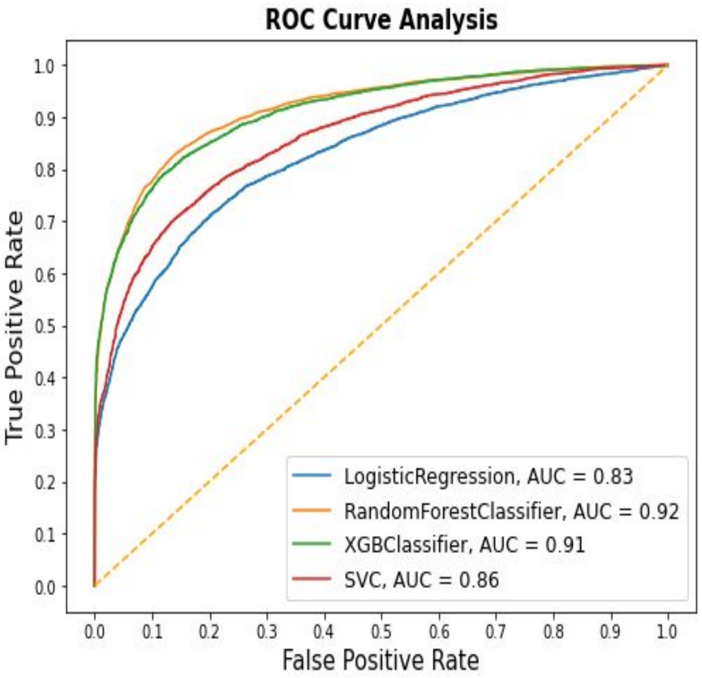


Plotting ROC AUC for all the models

- Curve Analysis of all model.

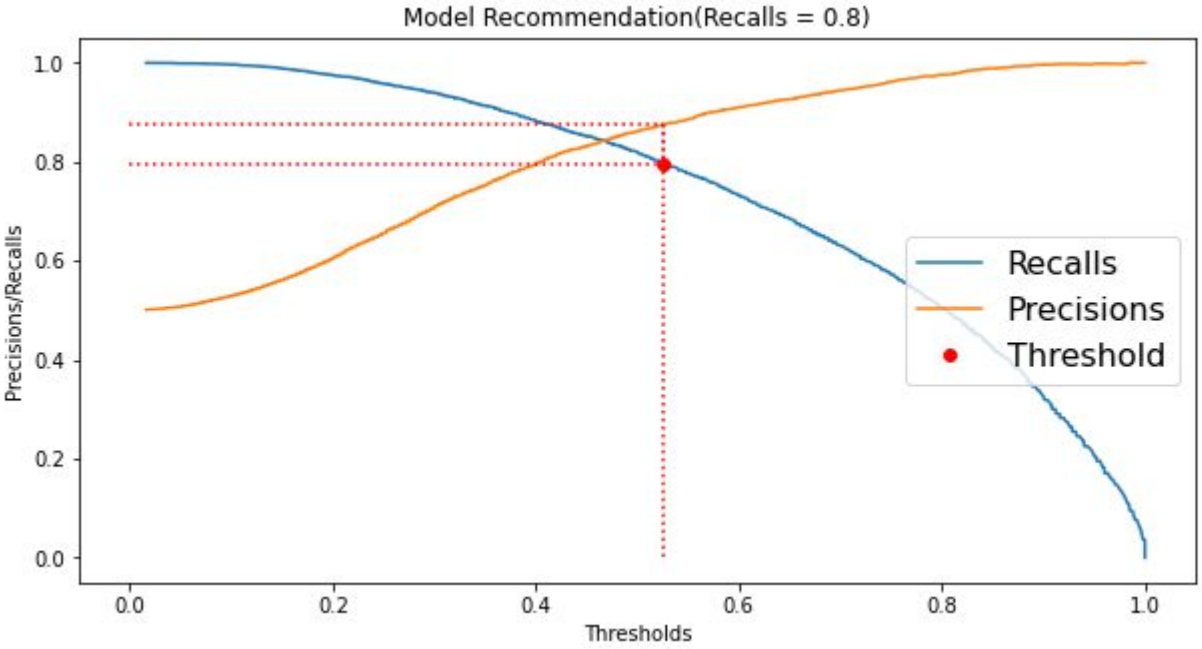


	fpr	tpr	auc
Classifiers			
LogisticRegression	[0.0, 0.0, 0.0, 0.0001712035610340695, 0.00017...	[0.0, 0.0001712035610340695, 0.091936312275295...	0.829101
RandomForestClassifier	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[0.0, 0.036466358500256806, 0.0369799691833590...	0.917184
XGBClassifier	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[0.0, 0.0001712035610340695, 0.003081664098613...	0.912108
SVC	[0.0, 0.0, 0.0, 0.0001712035610340695, 0.00017...	[0.0, 0.0001712035610340695, 0.189008731381612...	0.861698



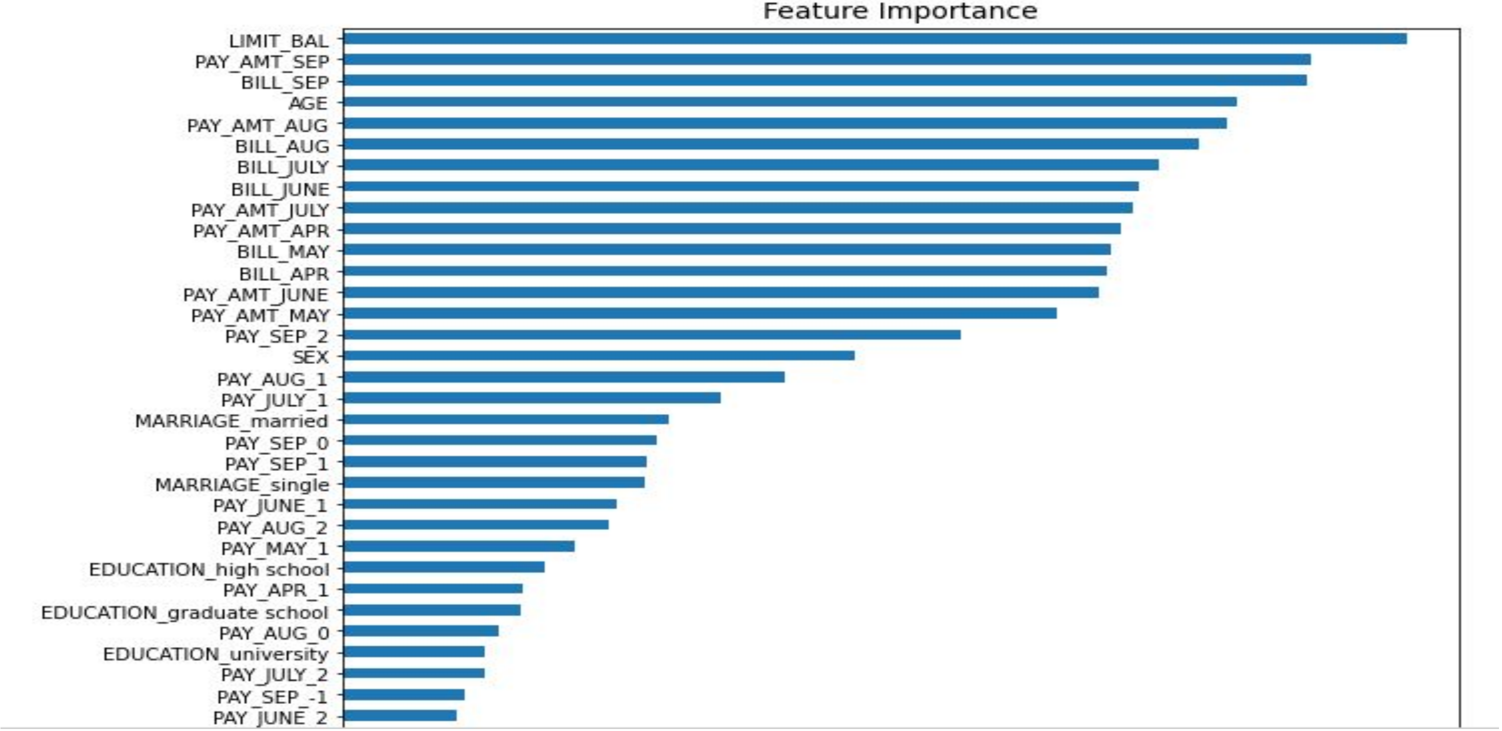
Model Recommendation

- We recommend recall = 0.8, however, the threshold can be adjusted to reach higher recall.



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.

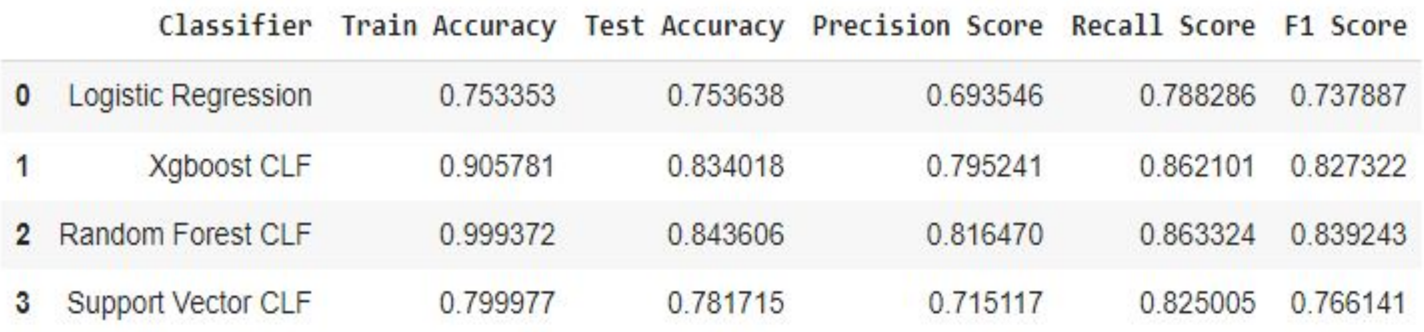


Challenges

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

Overall Conclusion

- Random Forest model and XGBoost 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	0.753353	0.753638	0.693546	0.788286	0.737887
1	Xgboost CLF	0.905781	0.834018	0.795241	0.862101	0.827322
2	Random Forest CLF	0.999372	0.843606	0.816470	0.863324	0.839243
3	Support Vector CLF	0.799977	0.781715	0.715117	0.825005	0.766141

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 the 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.
- In general, all models have comparable accuracy. Nevertheless, because the classes are imbalanced (the proportion of non-default credit cards is higher than default) this metric is misleading. Also, accuracy does not consider the rate of false positives (non-default credits cards that were predicted as default) and false negatives (default credit cards that were incorrectly predicted as non-default). Both cases have negative impact on the bank, since false positives leads to unsatisfied customers and false negatives leads to financial loss.
- 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 = 83% 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