

Capstone Project Credit Card Default Prediction

By
NIDHI PANDEY





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

Al

Content

- Problem Statement
- Data Pipeline
- Data Description
- Handling Null values
- Exploratory Data Analysis
- Hypothesis Testing
- Models Performed
- ☐ ROC AUC for all the models.
- Model Validation & Selection
- Evaluation Matrix of All the models
- Challenges
- Conclusion



Data Pipeline

- Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to see the trend.
- Data Processing: In this part we went through each attributes and encoded the categorical features.
- Model Creation: Finally in this part we created the various models.
 These various models are being analysed and we tried to study various models so as to get the best performing model for our project.



Data Description

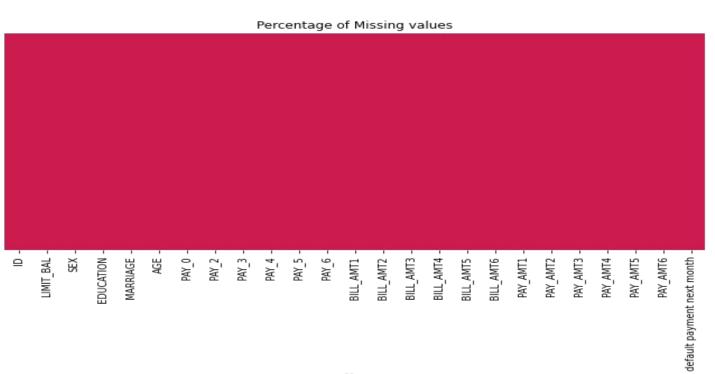
We have considered 23 variables as explanatory variables:

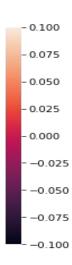
- X1: Amount of the given credit -it includes both the individual consumer credit and his/her family credit
- X2: Gender
- X3: Education
- X4: Marital status
- X5: Age (year)
- X6 -X11: History of past payment (from April to September, 2005)
- X12-X17: Amount of bill statement (from April to September, 2005)
- X18-X23: Amount of previous payment (from April to September, 2005)



Handling Missing / Null / Duplicate Values

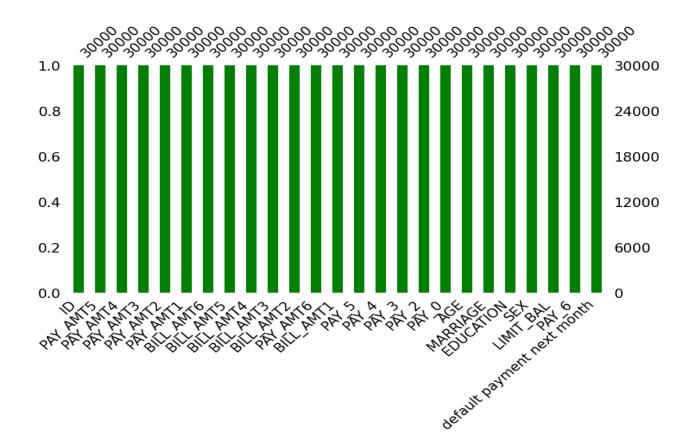
No missing or null values or duplicates are found in our dataset.





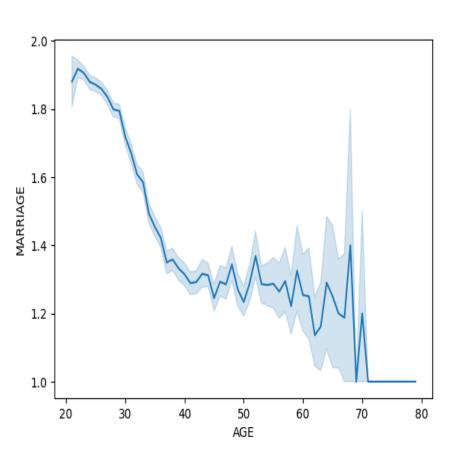
Handling Missing / Null / Duplicate Values Contd.

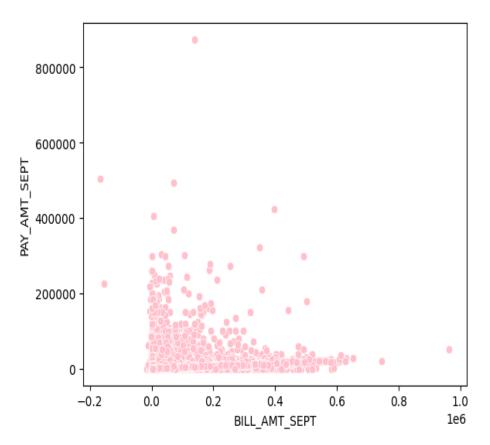




EDA- Dependent Variables

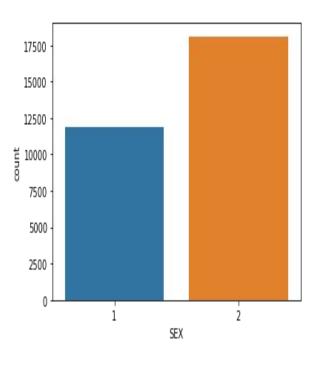




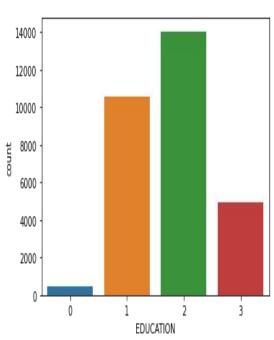


EDA- Independent Variables





1 : 'MALE'
2 : 'FEMALE'

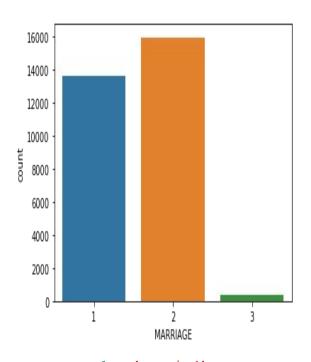


O : 'graduate school'

l : 'university'

2 : 'high school'

3 : 'others'



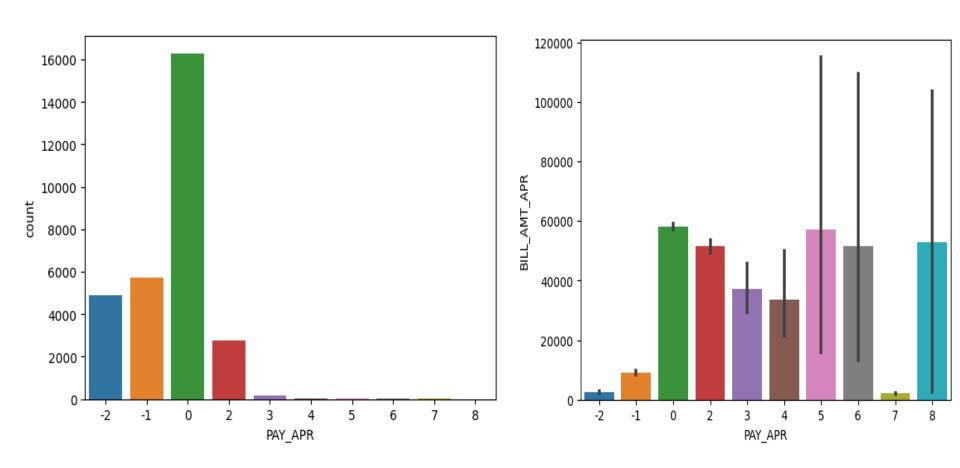
1 : 'married'

2 : 'single'

3 : 'others'

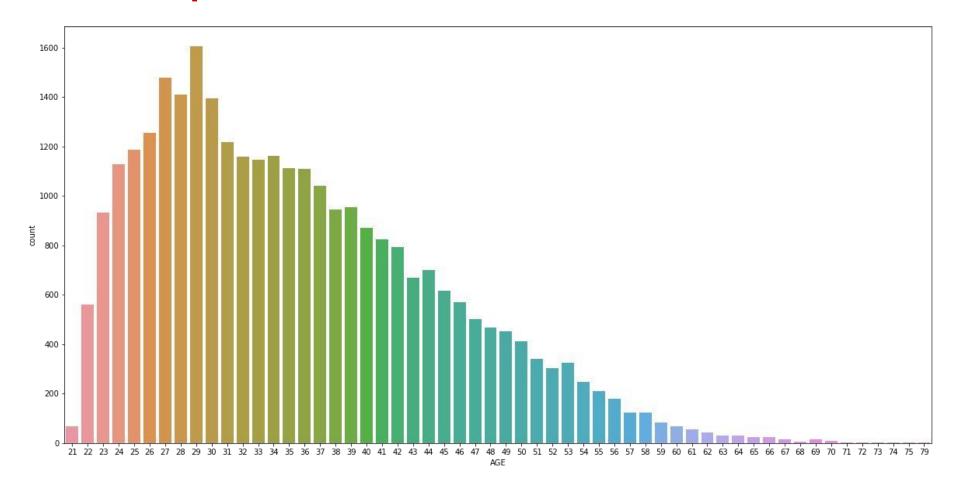
Visualization between BILL_AMT_APR and PAY_APR





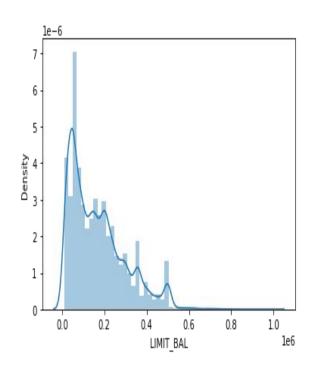
EDA- Independent Variables contd.

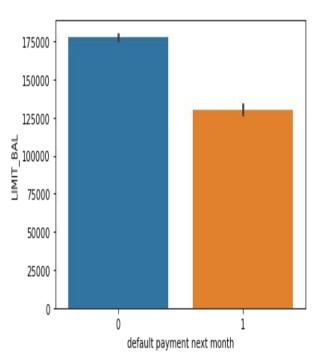


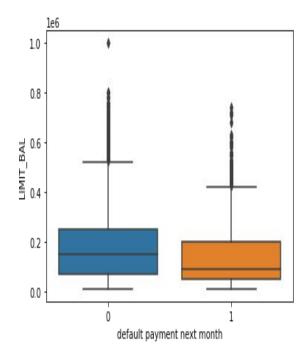




Limit Balance

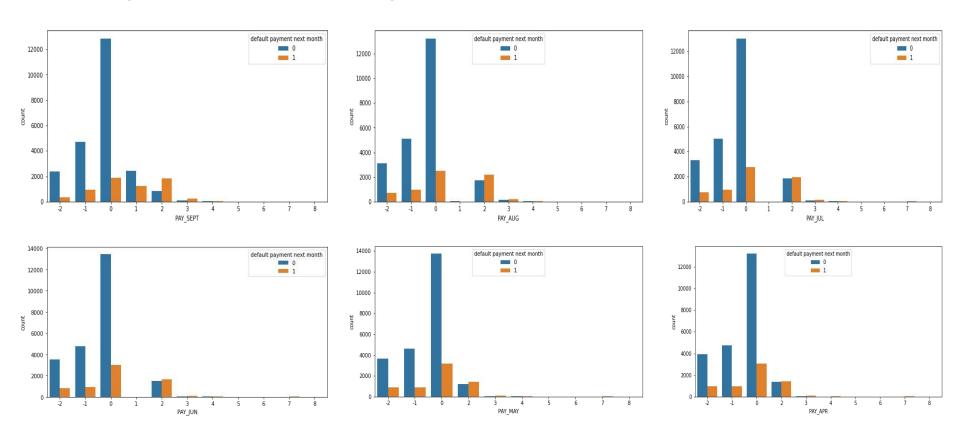






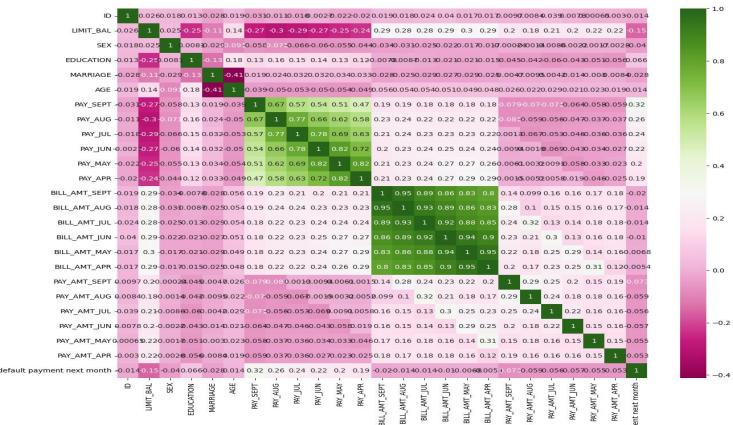
History / Previous Payments





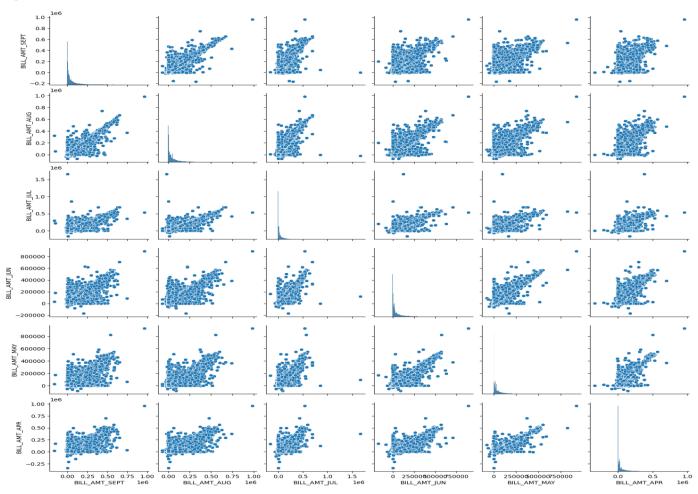
Checking Correlations

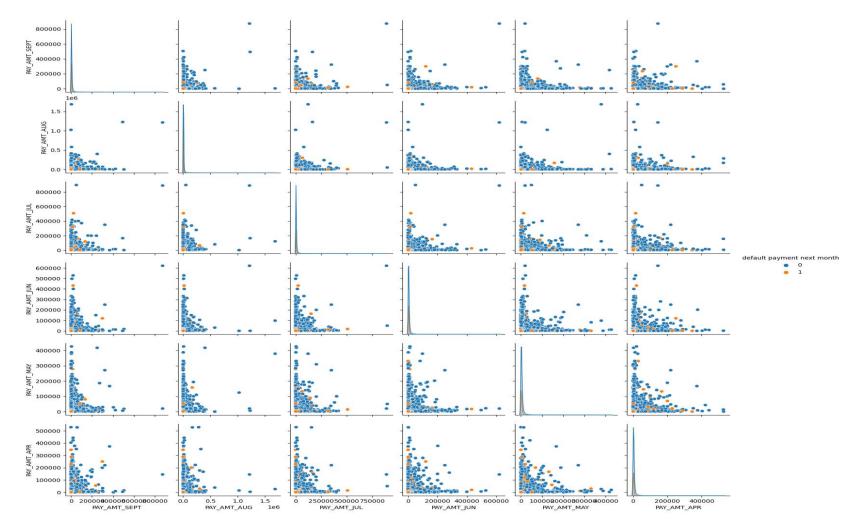




PAIR PLOT



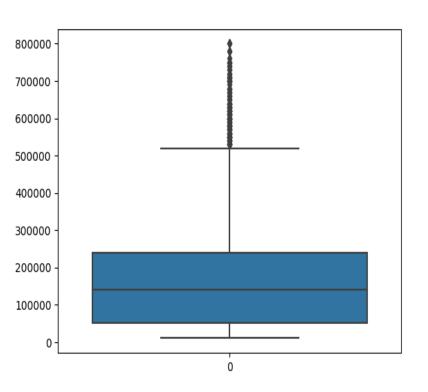


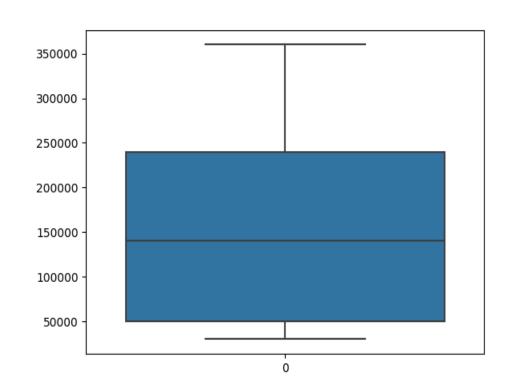


HANDLING OUTLIERS



I have used Removing/deleting the outliers, Replacing them with mean/median, Quantile based flooring and capping.





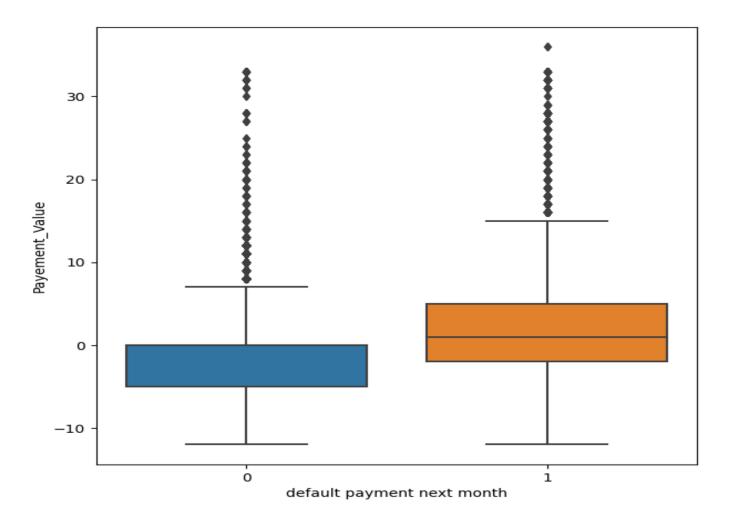
FEATURE MANIPULATION AND SELECTION



- 0.6

LIMIT_BAL -	0.027																								
SEX		1	0.025	-0.25	-0.11		-0.27	-0.3	-0.29	-0.27	-0.25	-0.24								0.18					-0.15
	0.018	0.025	1	0.0081	-0.029	-0.091	-0.058	-0.071	-0.066	-0.06	-0.055	-0.044	-0.034	-0.031	-0.025	-0.022	-0.017	-0.017-	0.00024	: 0.001∠	-0.0086	0.0022	0.0017	0.0028	-0.04
EDUCATION	0.013	-0.25	0.0081	1	-0.13								-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	-0.11	-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	-0.091	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	0.67	0.57	0.54	0.51	0.47	0.19						-0.079	-0.07	-0.071	-0.064	-0.058	-0.059	0.32
PAY_AUG	-0.011	-0.3	-0.071	0.16	0.024	-0.05	0.67	1	0.77										-0.081	-0.059	-0.056	-0.047	-0.037	-0.037	0.26
PAY_JUL -	-0.018	-0.29	-0.066	0.15	0.032	-0.053		0.77	1	0.78	0.69								0.0013	-0.067	-0.053	-0.046	-0.036	-0.036	0.24
PAY_JUN	-0.0027	-0.27	-0.06	0.14	0.032	-0.05			0.78	1	0.82	0.72							0.0094	0.0019	-0.069	-0.043	-0.034	-0.027	0.22
PAY_MAY	-0.022	-0.25	-0.055		0.034	-0.054			0.69	0.82	1	0.82							0.0061	-0.0032	0.0091	-0.058	-0.033	-0.023	0.2
PAY_APR	-0.02	-0.24	-0.044	0.12	0.033	-0.049		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		-0.034	0.0078	-0.028	0.056							1	0.95	0.89	0.86	0.83	0.8		0.099					-0.02
BILL_AMT_AUG	0.018		-0.031	0.0087	-0.025	0.054							0.95	1	0.93	0.89	0.86	0.83			0.15				-0.014
BILL_AMT_JUL	0.024		-0.025	-0.013	-0.029	0.054							0.89	0.93	1	0.92	0.88	0.85			0.13	0.14			-0.014
BILL_AMT_JUN	0.04		-0.022	-0.021	-0.027	0.051							0.86	0.89	0.92	1	0.94	0.9							-0.01
BILL_AMT_MAY	0.017		-0.017	-0.021	-0.029	0.049							0.83	0.86	0.88	0.94	1	0.95		0.18					0.0068
BILL_AMT_APR	0.017		-0.017	-0.015	-0.025	0.048							0.8	0.83	0.85	0.9	0.95	1	0.2						0.0054
PAY_AMT_SEPT	0.0097		0.00024	1 -0.045	-0.0047	0.026	-0.079	-0.081	0.0013-	0.0094	F0.0061	0.0015	0.14						1	0.29	0.25				-0.073
PAY_AMT_AUG																			0.29	1	0.24	0.18			-0.059
PAY_AMT_JUL																					1	0.22			-0.056
PAY_AMT_JUN																				0.18		1	0.15		-0.057
PAY_AMT_MAY																			0.15		0.16	ļ	1	0.15	
PAY_AMT_APR															0.18		0.16		0.19						-0.053
default payment next month	_			0.066			0.32		_	-,-		-,-	-0.02				-	-0.0054	-0.073		-0.056			,	1
	Q	UMIT_BAL	XX	EDUCATION	MARRIAGE	AGE	PAY_SEPT	PAY_AUG	PAYJUL	PAYJUN	PAY_MAY	PAY_APR	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	fault payment next month





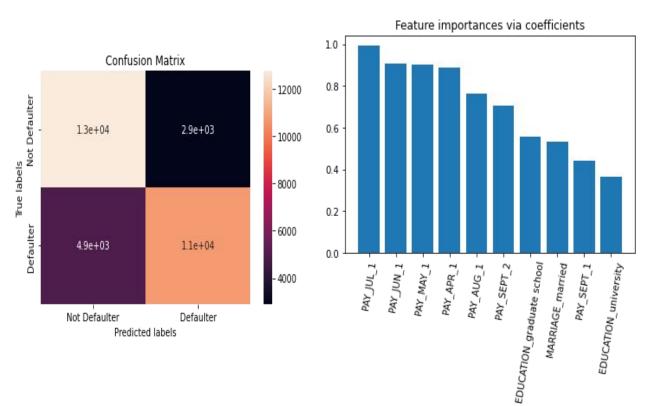
Model's Performed

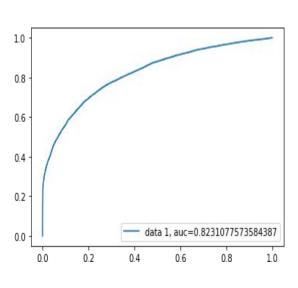


- Logistic Regression
- Decision tree
- Random forest
- eXtreme Gradient Boost

Logistic Regression

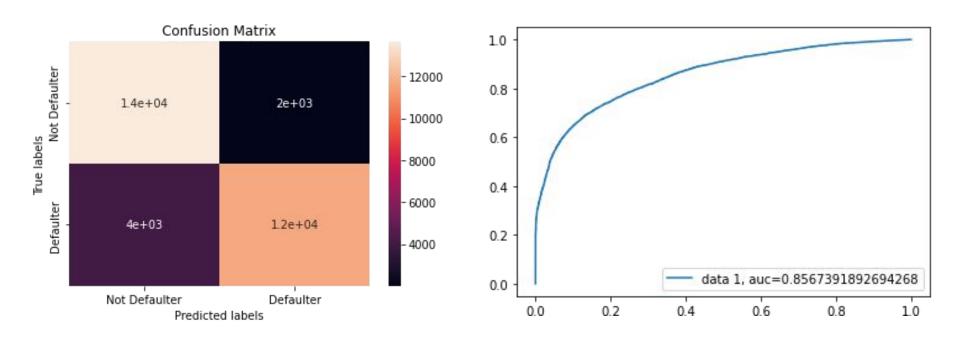






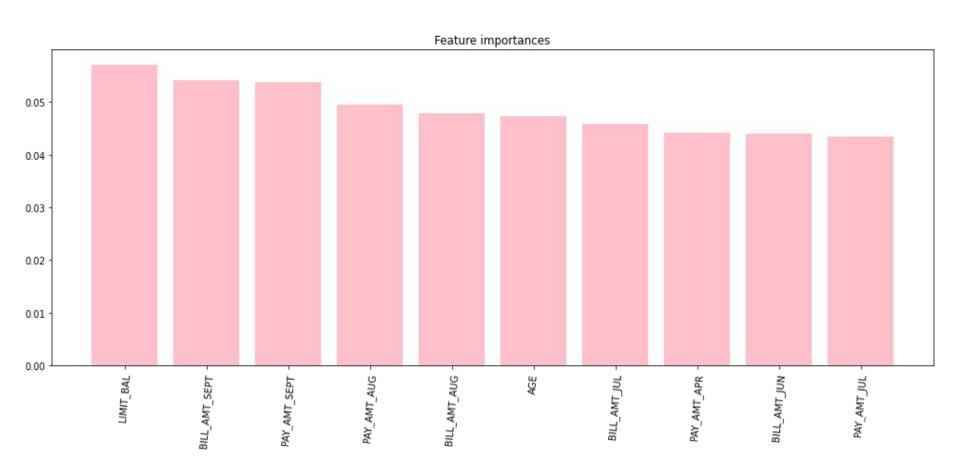






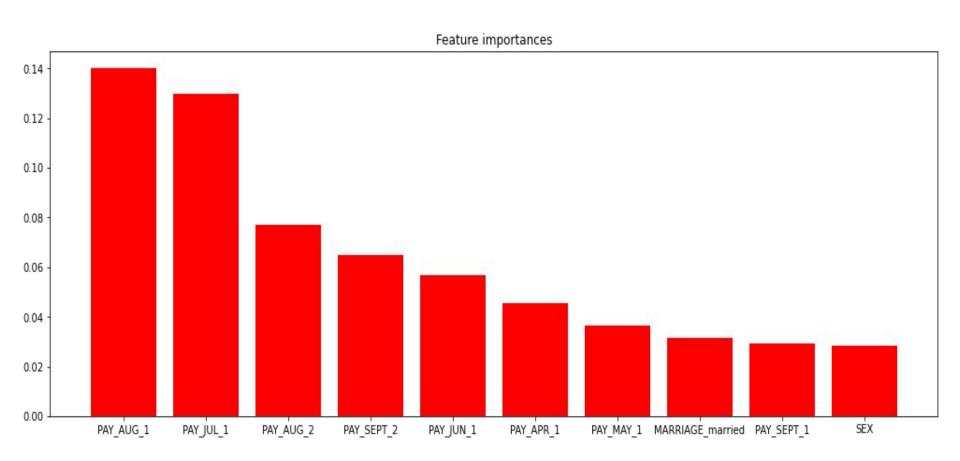
RandomForest







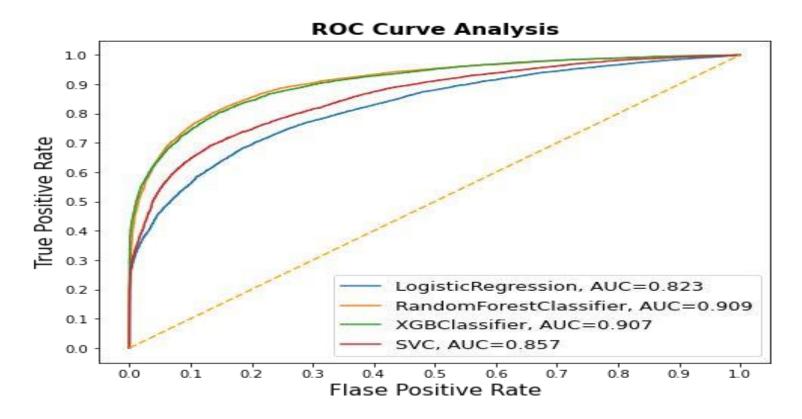
XGBoost feature Importance





ROC AUC for all the Models

ROC - AUC curve is a performance measurement for the classification problems at various threshold settings.





Model Validation & Selection (continued)

- Observation 1: As seen in the ROC AUC plot, Logistic Regression is not giving great results.
- Observation 2: Support Vector Classifier and Decision tree performed equally good.
- Observation 3: We are getting the best results from Random forest and XGBoost Classifier.





Challenges

- A huge amount of data needed to be dealt while doing the project which is quite an important task and also even small inferences need to be kept in mind.
- As dataset was quite big enough which led more computation time.







- After performing the various model we the get the best accuracy from the Random forest and XGBoost classifier.
- Logistic Regression is the least accurate as compared to other models performed.
- XGBoost has the best precision and the recall balance.
- Higher recall can be achieved if low precision is acceptable.
- We can deploy the model and can be served as an aid to human decision.
- Model can be improved with more data and computational resources



THANK YOU