Capstone Project 3

Credit Card Default Prediction

Kiran Ahire

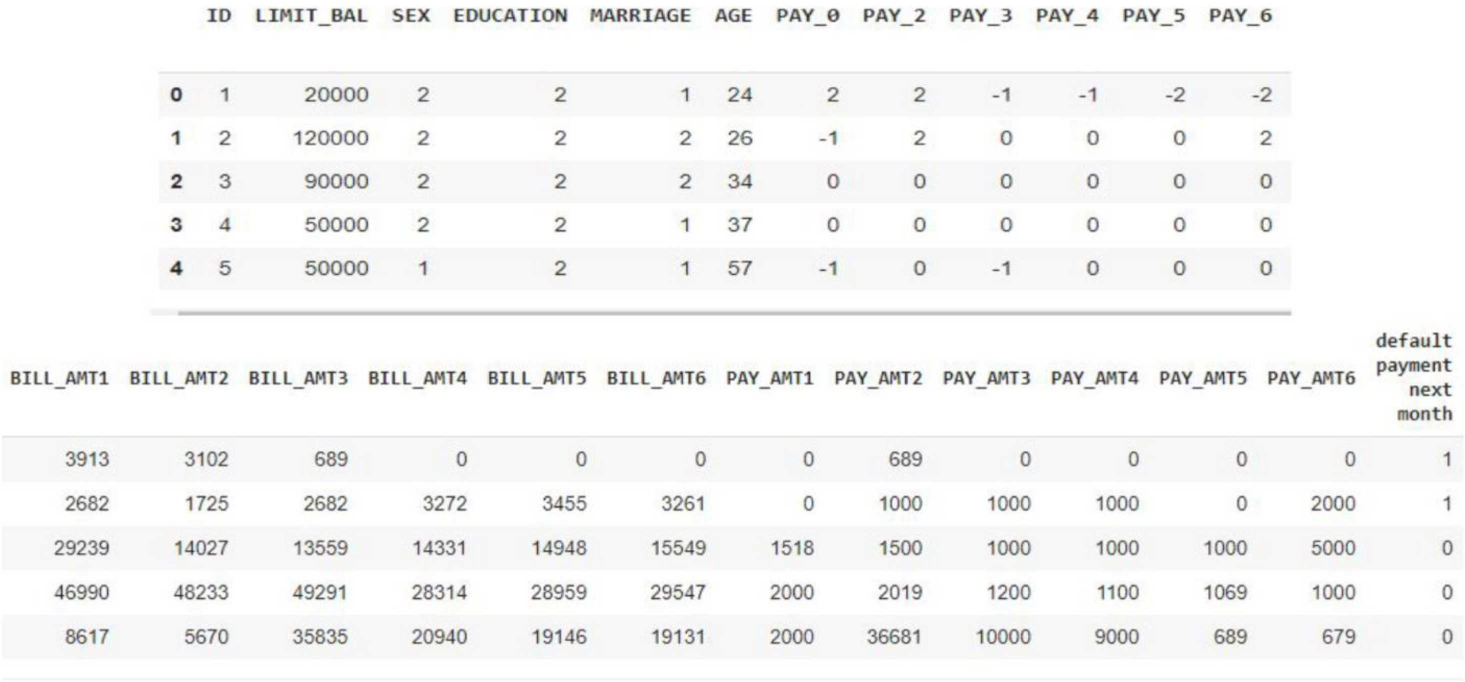
Points for Discussion

- Business Objective
- Data Summary
- Feature Summary
- Data Preprocessing
- Exploratory Data Analysis
- Corelation between data
- Algorithms implementation
- Conclusion

Business Objective

- We are all aware what is Credit card. It is type of payment card in which charges are made against a line of credit instead of the account holder's cash deposits. When someone uses a credit card to make a purchase, that person's account accrues a balance that must be paid off each month.
- 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 provided dataset contains information about Gender, Amount of given credit, Education, Age, Repayment Status, Amount of Bill Statement, Amount of previous payment and the target is Default payment next month. This project is aimed at predicting the case of customers default payments in Taiwan.

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

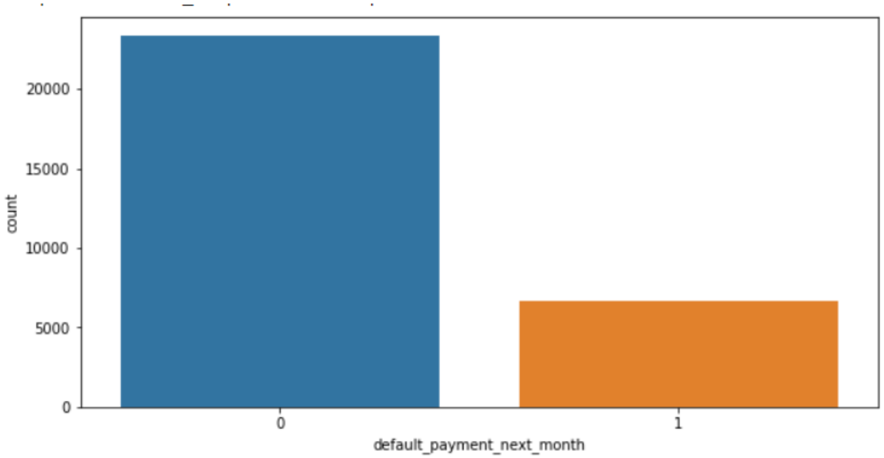
Feature Summary

- X1 - Amount of the given credit
- X2 - Gender
- X3 - Education
- X4 - Marital status
- X5 - Age
- X6 - X11: History of past payment from April to September
- X12-X17: Amount of bill statement April to September
- X18-X23: Amount of previous payment April to September
- Y – Default payment

Data Preprocessing

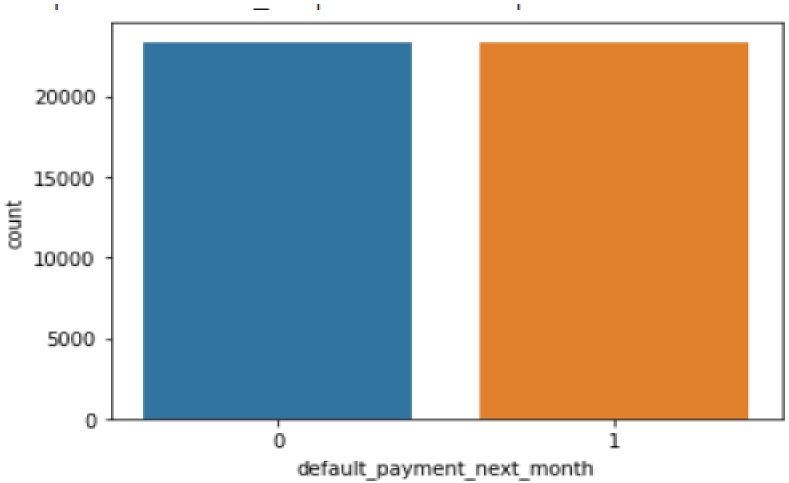
- The dataset contains 30000 rows and 25 columns
- There are no missing values present
- There are no duplicate values present
- There are no null values present
- ‘Default payment next month’ is the dependent variable which we need to predict for our new observations
- There 9 categorical variables in our dataset
- I rename some of the features for convenience.

Analysis of Dependent variable



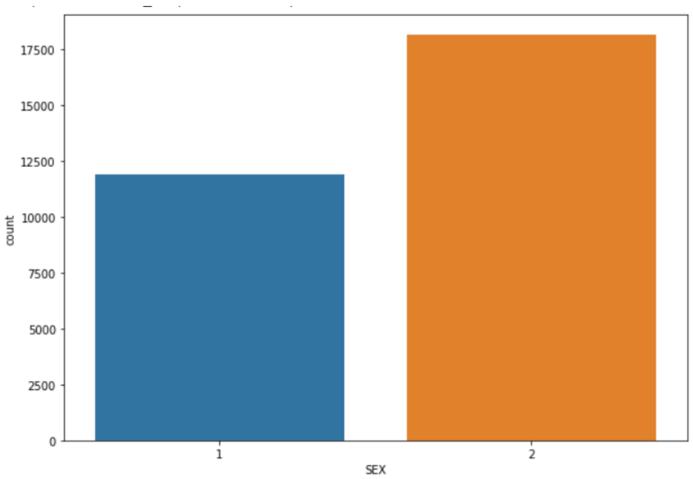
- In our dataset ‘default payment next month’ is dependent variable.
- From the above data analysis we can say that
- 0 - Not Default
- 1 - Default
- Numbers of Defaulters are less than the Non Defaulters in the given dataset
- From the graph we can see that both classes are not in proportion i.e imbalanced dataset so we normalize our dataset.

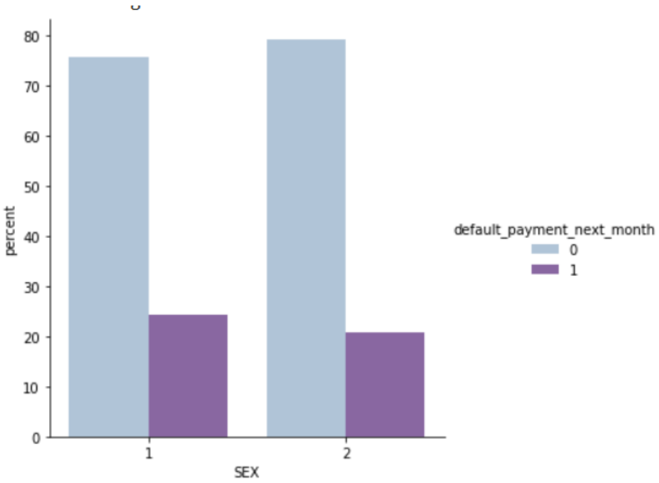
SMOTE



- **SMOTE** - Synthetic Minority Oversampling Technique is 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 we get this balanced dataset.

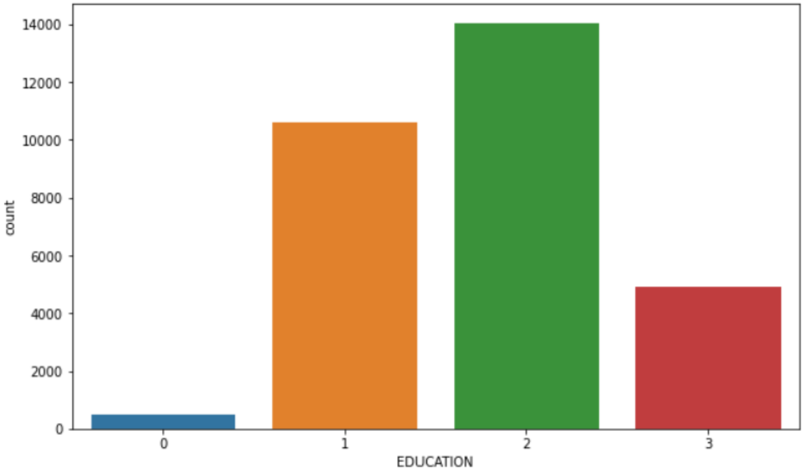
Analysis of SEX

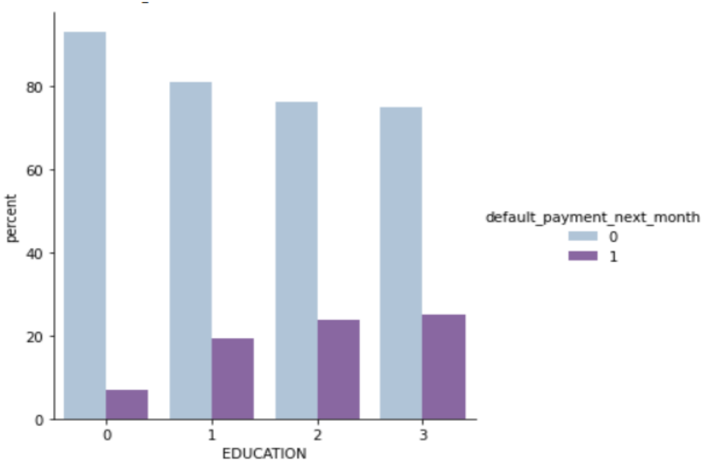




- From the above data analysis we can say that
- 1 - Male
- 2 - Female
- Number of Male credit holder is less than Female.
- From the above graph we can say that number of defaulter have high proportion of Males.

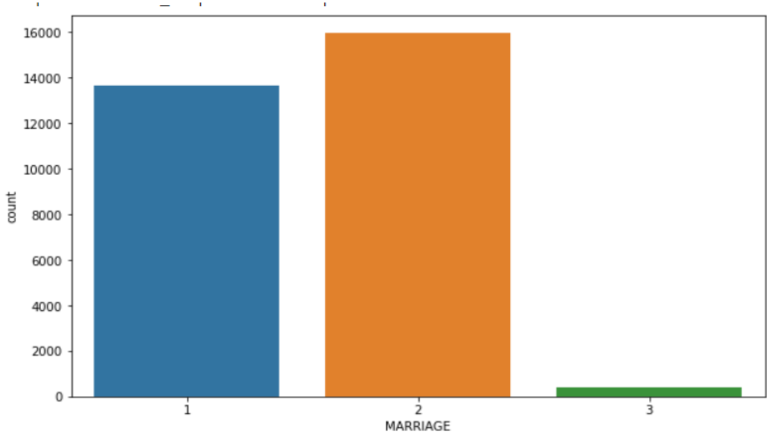
Analysis of Education

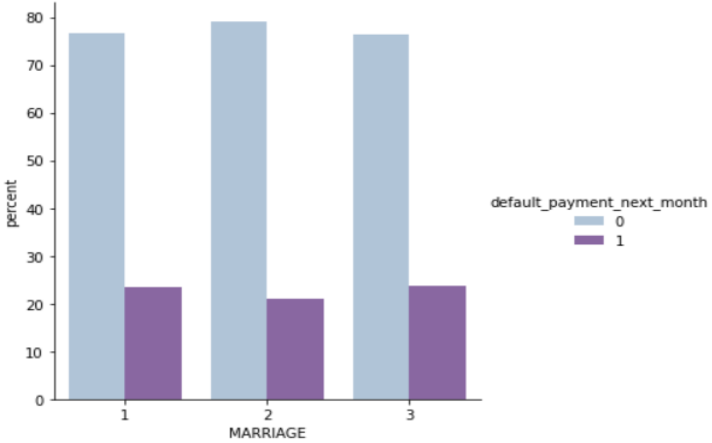




- 1 = graduate school; 2 = university; 3 = high school; 0 = others
- From the above data analysis we can say that
- More number of credit holders are university students followed by Graduates and then High school students.
- From the above plot it is clear that people in Other category have higher default payment as compared to graduates and university people

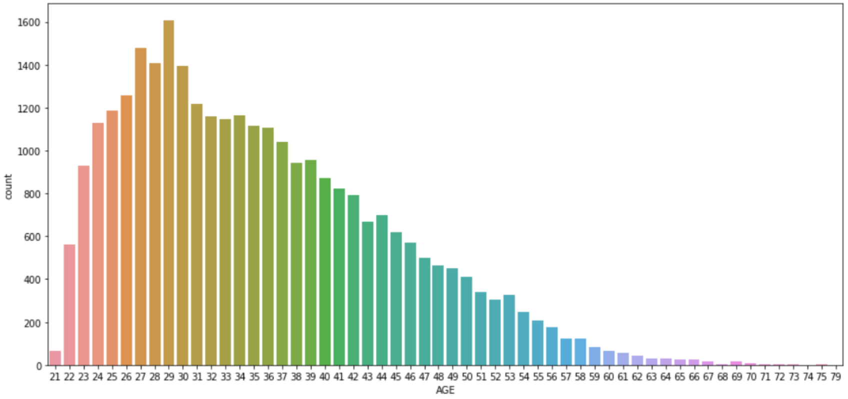
Analysis of Marriage

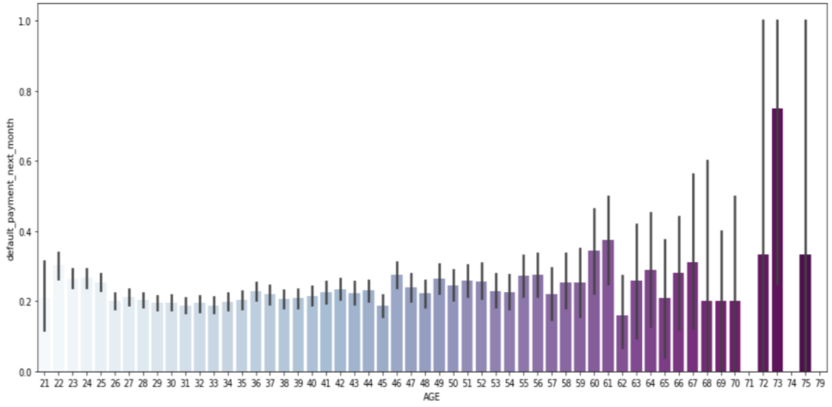




- 1 = married; 2 = single; 3 = others
- From the above data analysis we can say that
- 1 - married
- 2 - single
- 3 - others
- More number of credit cards holder are Single.
- from the above graph we can say that defaulter rate of Married and others category more than those of single.

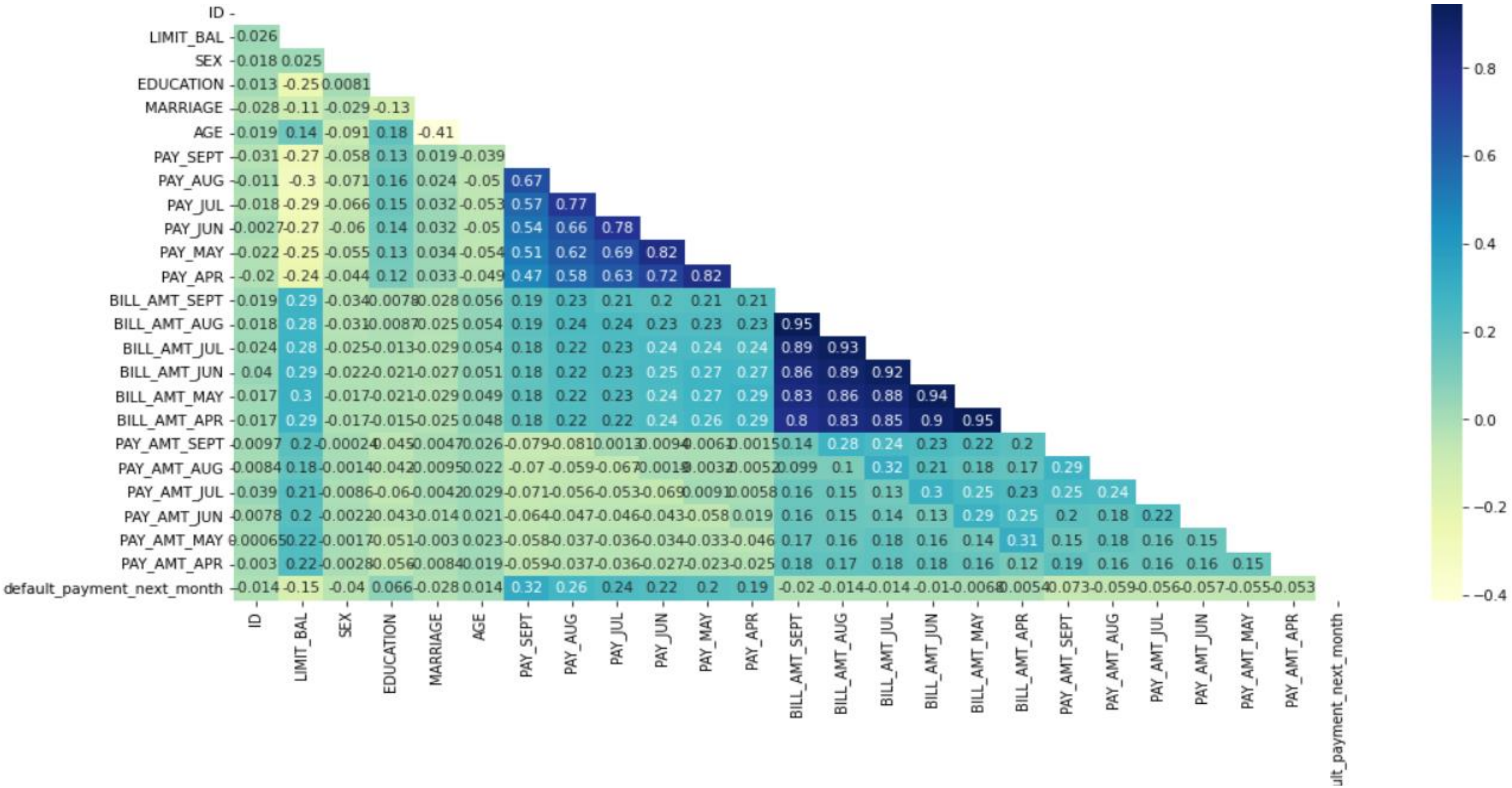
Analysis of Age





- From the above data 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.
- default rate is higher in people whose age is more than 60.

Corelation between variables



	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_SEPT	PAY_AUG	PAY_JUL	PAY_JUN	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	ult_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_SEPT	-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_JUL	-0.018	-0.29	-0.066	0.15	0.032	-0.053	0.57	0.77																	
PAY_JUN	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_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													
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	0.95												
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	0.93											
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	0.86	0.89	0.92										
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	0.83	0.86	0.88	0.94									
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	0.8	0.83	0.85	0.9	0.95								
PAY_AMT_SEPT	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_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					
PAY_AMT_JUN	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	

One Hot Encoding

- Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers. This is required for both input and output variables that are categorical.
- One hot encoding is **one method of converting data to prepare it for an algorithm and get a better prediction**.
- With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector.
- Here we performed One Hot Encoding on ‘EDUCATION’ , ‘MARRIAGE’ , ‘PAY_SEPT’, ‘PAY_AUG’, ‘PAY_JULY’, ‘PAY_JUNE’, ‘PAY_MAY’
- We have applied Label encoding on feature ‘SEX’

Models Implemented

1. Logistic Regression
2. Random Forest

Logistic Regression

- **Logistic Regression** - Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.
- The function used in Logistic Regression is sigmoid function or the logistic function given by:

$$f(x)= 1/1+e^{(-x)}$$

- Our best parameters from listed hyperparameters are :

{'C': 0.1, 'penalty': 'l2’}

- From the regression model we get the following results

The accuracy on test data is **0.751377991051164**

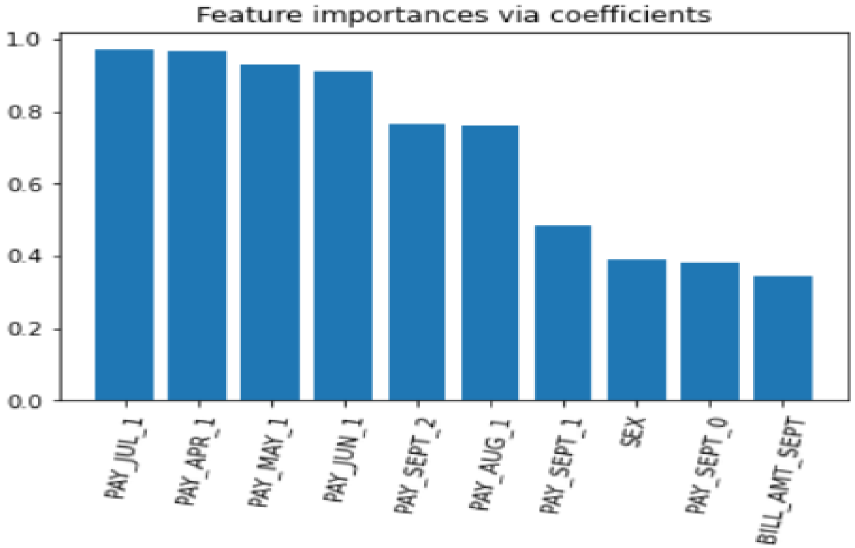
The precision on test data is **0.6883268482490272**

The recall on test data is **0.7876224398931434**

The f1 on test data is **0.7346345514950168**

The roc_score on test data is **0.755437386018206**

Feature Importance



- From the above feature importance graph we can say that the most important feature that make an impact on dependent variable are PAY_JUL_1,PAY_MAY_1,PAY_APR_1

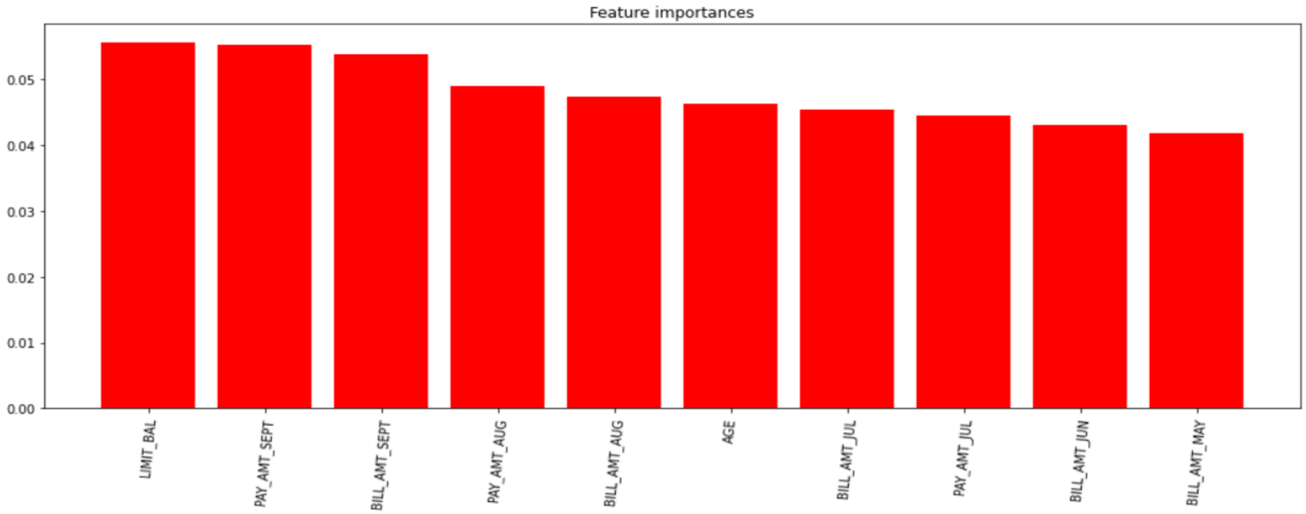
Random Forest

- **Random Forest Classifier** - Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.
- Our best parameters from listed hyperparameters are :

  {'max_depth': 30, 'n_estimators': 200}
- From the model we get the following results

  The accuracy on test data is 0.8333441411062836
  The precision on test data is 0.8007782101167316
  The recall on test data is 0.8565482796892342
  The f1 on test data is 0.8277248960986728
  The roc_score on test data is 0.8347638512777109

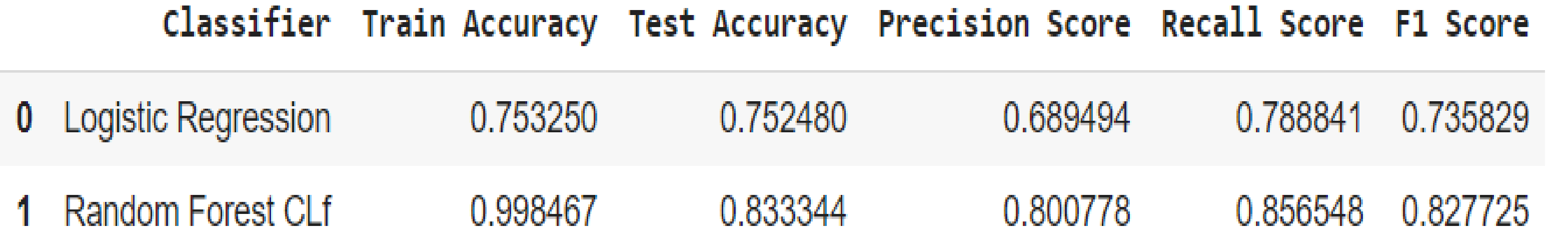
Feature Importance



- from the above feature importance graph we can say that the most important feature are LIMIT_BAL,PAY_AMT_SEPT

Evaluating the model

- We implemented 2 algorithms Logistic Regression, Random Forest Classifier. The results of our evaluation are:



	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.753250	0.752480	0.689494	0.788841	0.735829
1	Random Forest CLf	0.998467	0.833344	0.800778	0.856548	0.827725

Conclusion

- Recent two month payment status and credit limit are strongest default predictor.
- From above table we can see that **Logistic Regression** having **Recall**, **F1-score** equals 78%, 73% resp and **Random forest Classifier** having **Recall**, **F1-score** values equals 85%, 83%resp.
- The best **accuracy** is obtained for the **Random forest** than **Logistic Regression.**
- If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. Since Random Forest has slightly lower recall but much higher precision than Logistic Regression, I would recommend Random Forest.