Capstone Project Classification

(Abhishek Mishra,Kurva Mallesh, Arunesh Mishra)
Data science trainees,
Alma Better, Bangalore

Abstract:

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Our EDA can make us understand data which variable is very important and check how every variable connected with dependent variable.

We make some models to predict the label column based on feathers.

Keywords: EDA, Predicting whether a customer will default on his/her credit card

1.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.

Data Description: -

Attribute Information:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

- **ID**: ID of each client
- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- **SEX**: Gender (1=male, 2=female)
- **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others)
- **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
- **AGE**: Age in years
- **PAY_0**: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
- **PAY_2**: Repayment status in August, 2005 (scale same as above)
- **PAY_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)

- **PAY_5**: Repayment status in May, 2005 (scale same as above)
- **PAY_6**: Repayment status in April, 2005 (scale same as above)
- **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
- **BILL_AMT6:** Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
- **Default payment next month**: Default payment in June, 2005 (1=yes, 0=no)

1. Data wrangling step:-

1. We check that in the row data 2st row is the actual column name so we need to replace it.
2. In the dataset we found there are not null values.
3. After that we found that all the data types are in the form of object so we need to convert it into int64 form.

2. EDA

Exploratory Data Analysis (EDA):

After the data wrangling step, we performed EDA by comparing different parameters which are involved in the dataset. EDA helps us to find the different relations among the parameters. It involves the visualization of the data by comparing the different parameters to find out the best among all.

In the process of understanding the data we found that **Default payment next month** is the label column and rest all are feathers we did some analysis. We explain each one by one.

Check counts for all pay 0 to 6 columns:-

1. Notice code 0 and -2 are in the PAY columns but are not included in the data description.

```
 0    95919
-1    34640
-2    24415
 2    18964
 1     3722
 3     1430
 4      453
 7      218
 5      137
 6       74
 8       28
Name: value, dtype: int64
```

2. We found there are 95919 values as listed as 0. Using some Google search we found that 0 meaning the payment wasn't due, which makes sense that most customers were using the revolving credit.
3. Also we found 24415 values as -2 which mean No consumption.

Check label data column: -

```
0    0.7788
1    0.2212
Name: default payment next month, dtype: float64
```

1 So, we have 77.88% data as credit not default and 22% data as credit default.

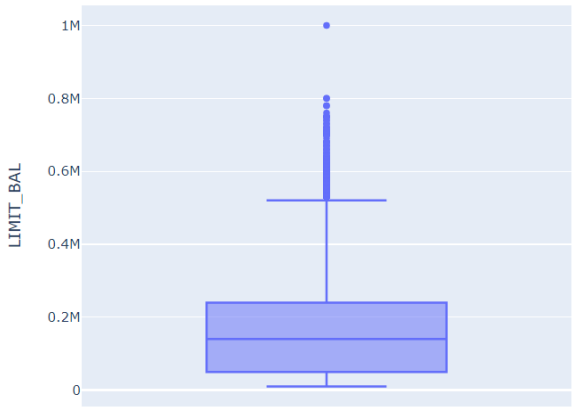
2 So, our data is highly imbalanced.

Rename the Column Name:-

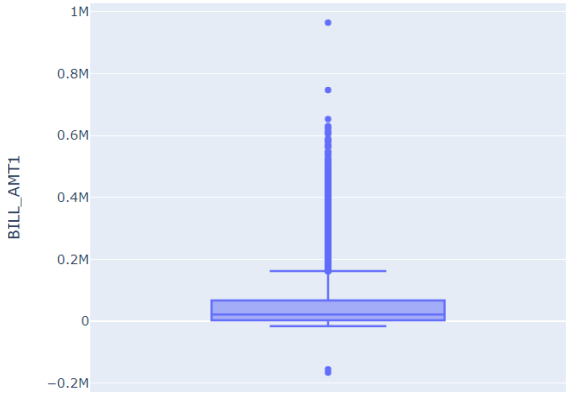
1. We found that there is a column name pay_0 in the data description only pay_1 is described so we replace it.
2. Also we found there is no duplicate row or column.

Identify outliers: -

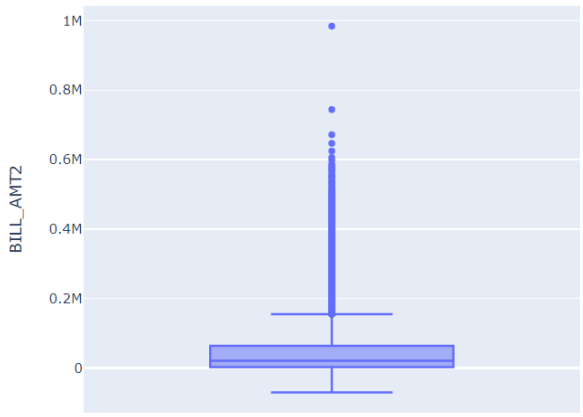
1. LIMIT_BAL1 outliers.



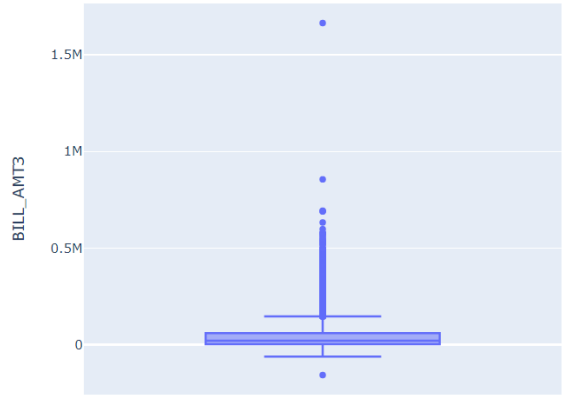
2. BILL_AMT1 outliers.



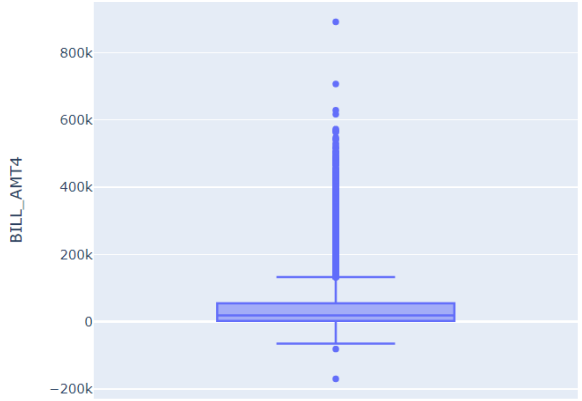
3. BILL_AMT 2outliers.



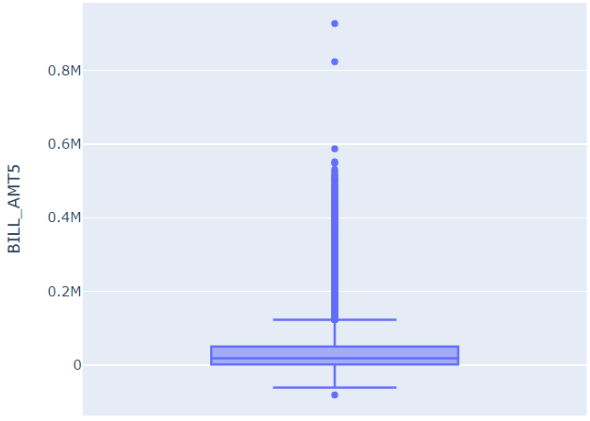
4. BILL_AMT 3 outliers.



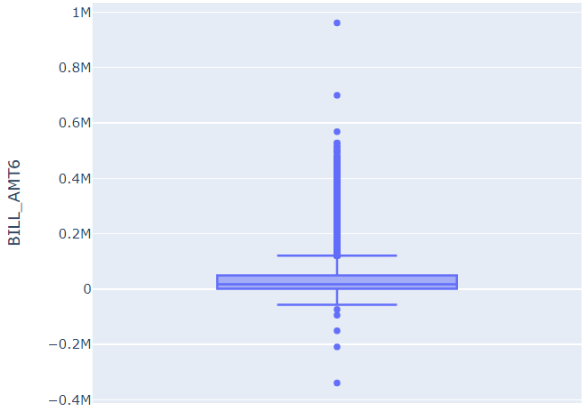
5. BILL_AMT 4 outliers.



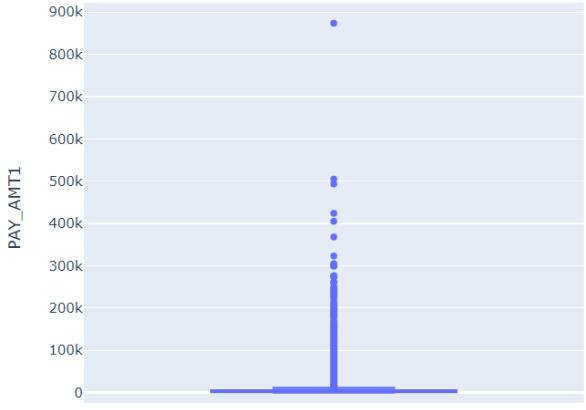
6. BILL_AMT5 outliers.



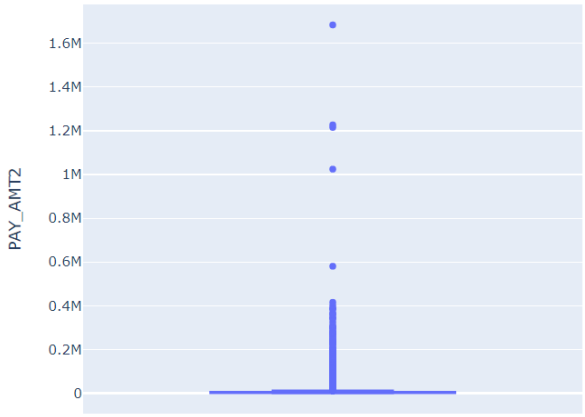
7. BILL_AMT6 outliers.



8. PAY_AMT1 outliers.

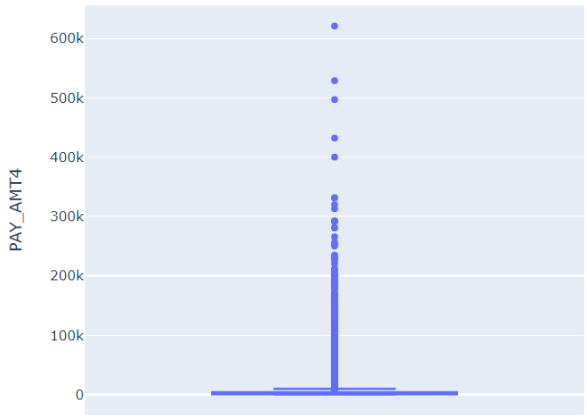


9. PAY_AMT2 outliers.

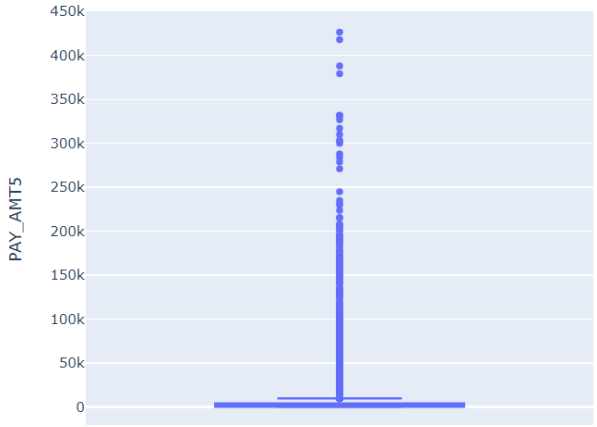


10. PAY_AMT3 outliers.

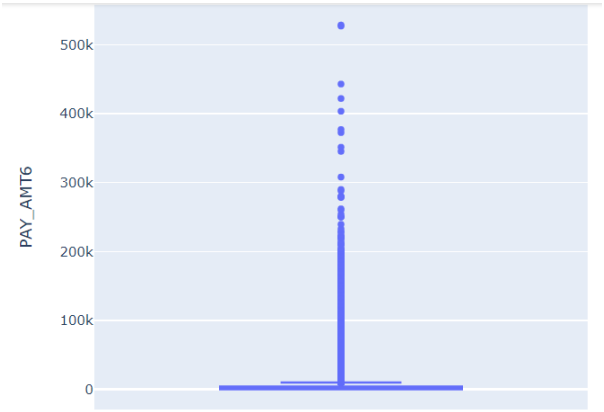
11. PAY_AMT4 outliers.



12. PAY_AMT5 outliers.



13. PAY_AMT6 outliers.

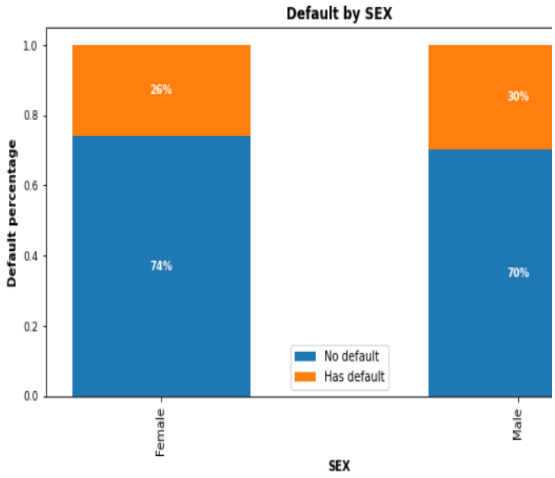


<u>we can create a column to check if the customer has default or not:-</u>

1. So we have another column has_def in this column we can see whether the customer have any deff or not.

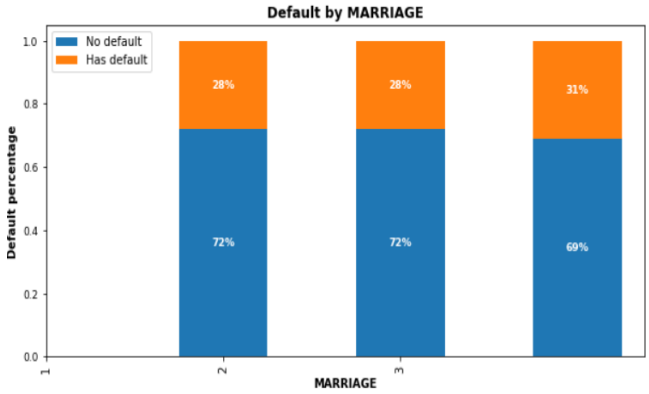
default percentages than customers with grad school education did.

Learn about SEX column:-



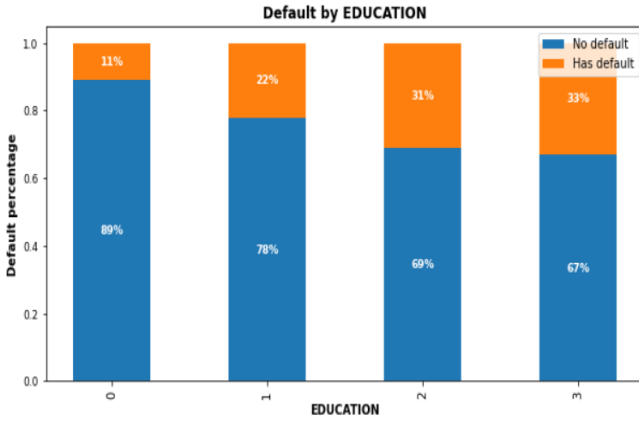
- 30% male have default payment while 26% female have default payment, the difference between them is not significant.
- Also we can observe that female have more count than male.

Learn about MARRIGE Column:-



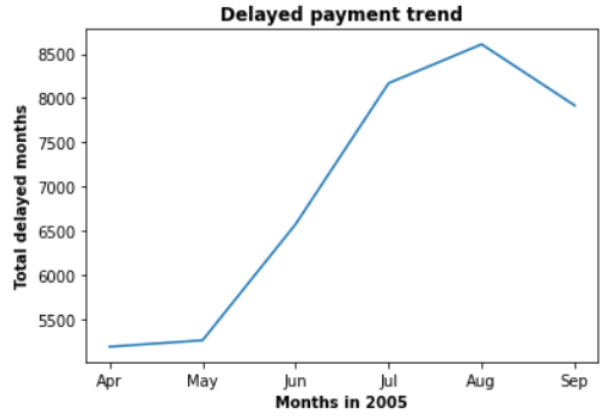
- 'MARRIAGE' column: what does 0 mean in 'MARRIAGE'? Since there are only 54 observations of 0,
- We will combine 0 and 3 in one value as 'others'.

Learn about Education Column:-



- The data indicates customers with lower education levels default more.
- Customers with high school and university educational level had higher

Let's check the change status of the payment in months:-

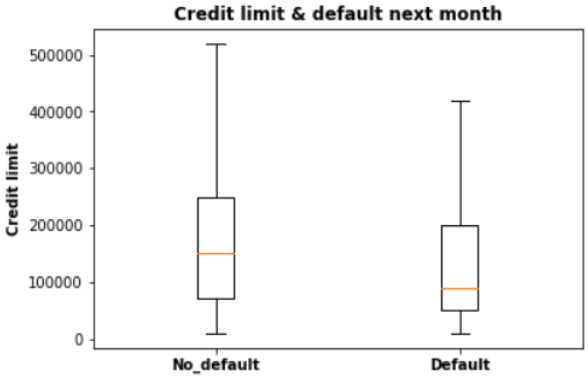


- There was a huge jump from May 2005 to July 2005.

- When delayed payment increased significantly, then it peaked at August 2005.
- Things started to getting better in September 2005

Check the relation between credit limit and the default payment next month: -

- Unsurprisingly, customers who had higher credit limits had lower delayed payment rates.

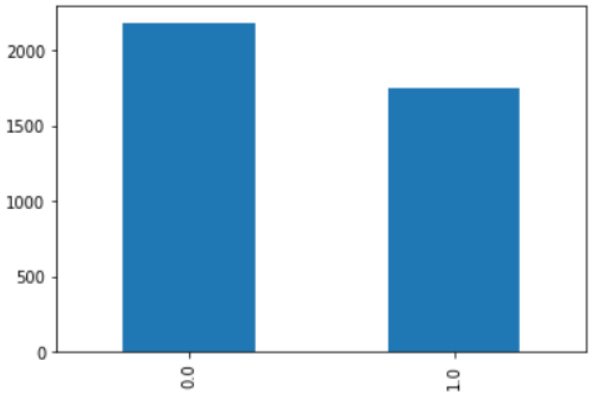


Check there are some negative bills:-

bill_cycle	amount
BILL_AMT1	590
BILL_AMT2	669
BILL_AMT3	655
BILL_AMT4	675
BILL_AMT5	655
BILL_AMT6	688

- The minimal of those 6 bill columns are negative numbers.
- In general, there are 590-688 bills with negative amounts each month, which is less than 2% of total 30,000 records monthly.

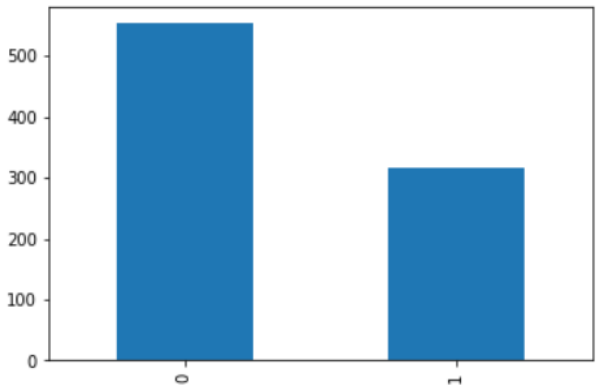
Check Why some bill statement amounts are greater than credit limit:-



- The common sense is that the bill statement amount shouldn't exceed credit limit.

- However, there are 3931 customers whose bill amounts are greater than credit limit.
- Could the difference be late payment interest assuming these customers had delayed payment?

Check customers who had no consumption in 6 months then default in the next month:-



- There are 870 customers whose bill amount was 0 in 6 months
- 317 customers had default payment next month which is against common sense.
- We will investigate this in the data analysis process.

1.EDA Conclusion : -

- Notice code 0 and -2 are in the PAY columns but are not included in the data description.
- We found there are 95919 values are listed as 0. Using some Google search we found that 0 meaning the payment wasn't due, which makes sense that most customers were using the revolving credit.
- Also we found 24415 values as -2 which mean No consumption.
- There are no duplicate IDs or rows.
- 30% male have default payment while 26% female have default payment, the difference is not significant.
- Also we can see Female have more count than male
- 'EDUCATION' column:
  notice 5 and 6 are both recorded as 'unknown' and there is 0 which isn't explained in the dataset description.
- Since the amounts are so small, let's combine 0,4,5,6 to 0 which means “other'.
- 'MARRIAGE' column: what does 0 mean in 'MARRIAGE'? Since there are only 54 observations of 0,
- We will combine 0 and 3 in one value as 'others'.
- There was a huge jump from May 2005 to July 2005
- When delayed payment increased significantly, then it peaked at August 2005.
- Things started to get better in September, 2005.
- Unsurprisingly, customers who had higher credit limits had lower delayed payment rates.
- The minimal of those 6 bill columns are negative numbers.
- In general, there are 590-688 bills with negative amounts each month, which is less than 2% *of total *30,000 records monthly.
- The common sense is that the bill statement amount shouldn't exceed credit limit.
- However, there are 3931 customers whose bill amounts are greater than credit limit.

- Could the difference be late payment interest assuming these customers had delayed payment?
- There are 870 customers whose bill amount was 0 in 6 months
- 317 customers had default payment next month which is against common sense.

---

2. Prepare for Modelling:-

- We use pair plots for understanding the data.
- Also created bins for AGE columns.
- This dataset is also imbalanced, with 78% non-default vs. 22% default.
- We use SMOTE because the class is highly Imbalance

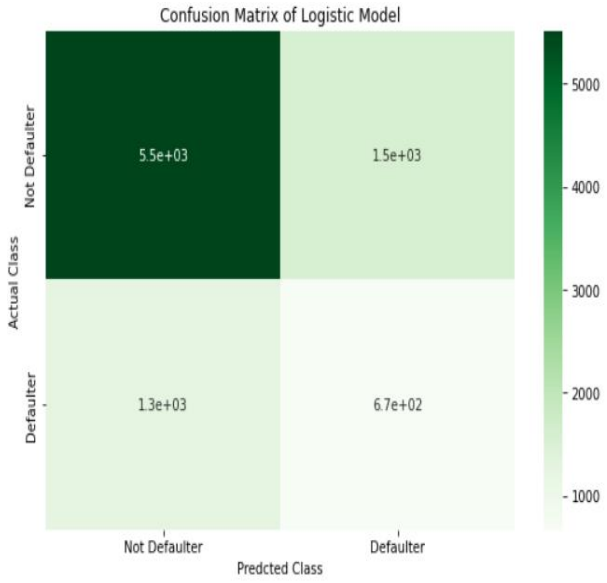
---

3. Classification Predictive Modelling:-

Logistic Regression: -

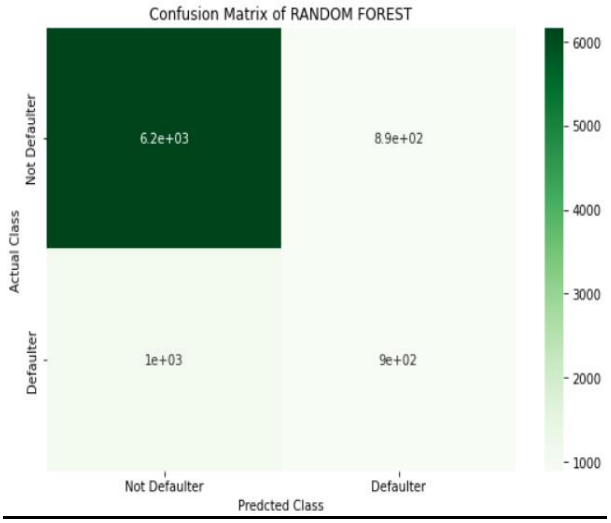
Using LOGISTIC regression, we get our scores like-

- Training accuracy: 0.5787536800785084
- Testing accuracy: 0.6876666666666666
- Precision score of logistic model: 0.3032083145051966
-
- Recall score of logistic model: 0.3458762886597938
-
- F1 score of logistic model: 0.32313989886828803
-
- ROC AUC score of logistic model: 0.5637313454630414
-
- Confusion matrix of logistic model : [[5518 1542]
  [1269 671]]



RANDOM FOREST Model:-

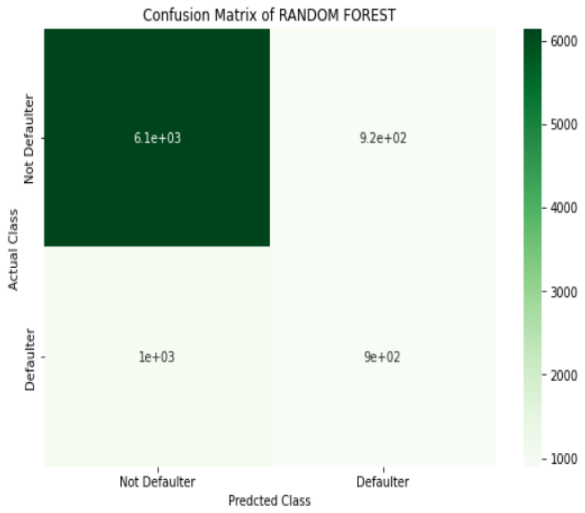
- Training Accuracy of Random Forest: 0.9981292934249264
- Testing Accuracy of Random Forest: 0.7853333333333333
- Precision score of logistic model: 0.5022371364653244
- Recall score of logistic model: 0.46288659793814435
- F1 score of logistic model: 0.48175965665236054
- ROC AUC score of logistic model: 0.6684121374959844
- Confusion matrix of RANDOM FOREST
  [[6170 890]
  [1042 898]]



RANDOM FOREST (hyper parameter tuning): -

Using RANDOM FOREST, we get our scores like-

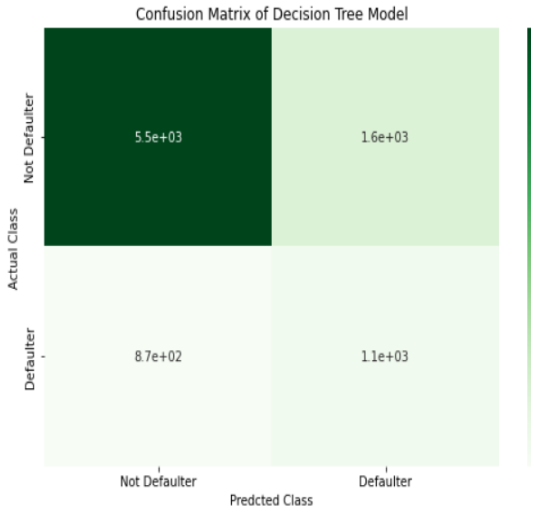
- The accuracy on train data is 0.9980986261040236
- The accuracy on test data is 0.7824444444444445
- Precision score of RANDOM FOREST: 0.4950440528634361
- Recall score of RANDOM FOREST: 0.4634020618556701
- F1 score of RANDOM FOREST: 0.4787007454739084
- ROC AUC score of RANDOM FOREST: 0.6667576881516312
- Confusion matrix of RANDOM FOREST model : [[6143 917] [1041 899]]



Decision Tree Classifier (hyper parameter tuning): -

Using Decision Tree Classifier, we get our scores like-

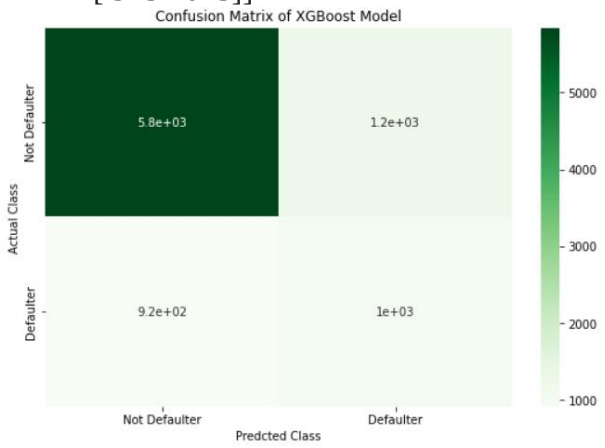
- Training accuracy of decision tree classifier: 0.7702097644749755
- Testing accuracy of decision tree classifier: 0.7242222222222222
- Precision score of Decision Tree model: 0.398729446935725
- Recall score of Decision Tree model: 0.55
- F1 score of Decision Tree model: 0.46230502599653384
- ROC AUC score of Decision Tree model: 0.6610481586402266
- Confusion matrix of Decision Tree model : [[5451 1609] [873 1067]]



XGBoost model:-

Using XGBoost, we get our scores like-

- Training Accuracy of XGBClassifier: 0.7917382237487733
- Testing Accuracy of XGBClassifier: 0.763
- Precision score of XGBoost model: 0.4569772625947392
- Recall score of XGBoost model: 0.5283505154639175
- F1 score of XGBoost model: 0.4900788907482668
- ROC AUC score of XGBoost model: 0.6779146345024969
- Confusion matrix of XGBoost model : [[5842 1218]
  [915 1025]]



2. Creating Data Frame of all Evaluation Matrix with respect of models: -

	Classification Models	Training Accuracy	Testing Accuracy	Precision Score	Recall Score	F1 Score	ROC-AUC Score
0	Logistic Regression	0.678754	0.687667	0.303208	0.345876	0.323140	0.563731
1	Random Forest	0.998129	0.786333	0.502237	0.462887	0.481760	0.668412
2	Random Forest tuning	0.998089	0.782444	0.495044	0.463402	0.478701	0.666758
3	Decision Tree Classifier	0.770210	0.724222	0.398729	0.550000	0.462305	0.661048
4	XGBoost Classifier	0.791738	0.763000	0.456977	0.528351	0.490079	0.677915

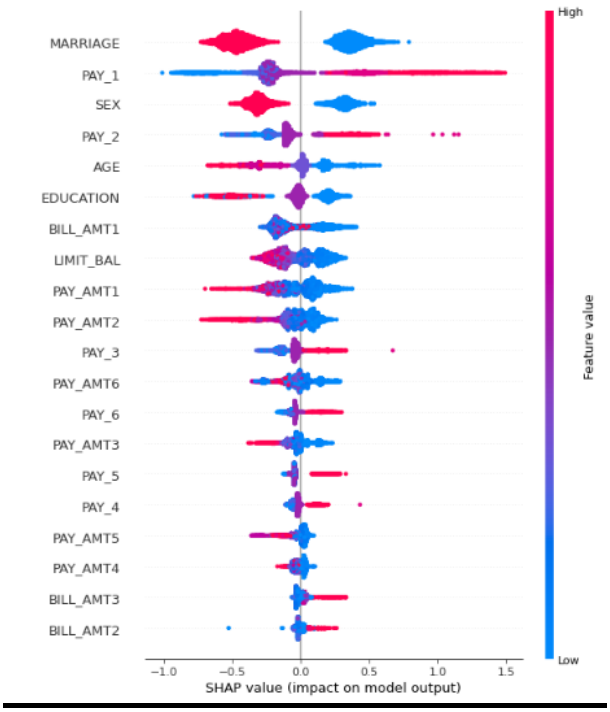
3. Creating Data Frame of all Evaluation Matrix with respect of models 2 : -

- Rename default default payment next month to is defaulter
- Use smote because the data is imbalanced
- Create another feather named Payement_Value after adding all pay columns
- Create another feater Dues.
- Replace 5,6,0 to 4 in education column
- Replace 0 with 3 in marriage column
- Using one hot encoding on Education, marriage column
- Using level encoding on Sex column.

	Classification Models	Training Accuracy	Testing Accuracy	Precision Score	Recall Score	F1 Sc
0	Logistic Regression	0.725205	0.721030	0.718182	0.727497	0.722
1	Random Forest	0.999393	0.835354	0.849723	0.814786	0.467
2	Decision Tree Classifier	0.775290	0.749368	0.771962	0.707782	0.738
3	XGBoost Classifier	0.790302	0.781013	0.807000	0.738651	0.771

About models -2:-

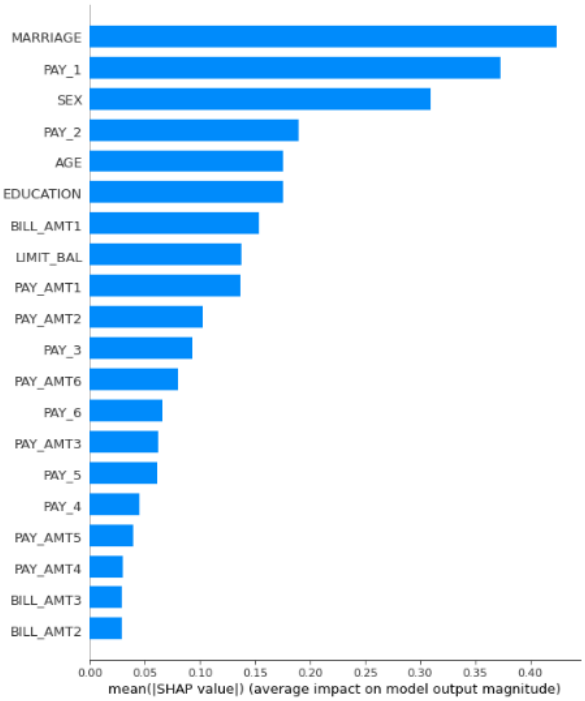
- Using a Logistic Regression classifier, we can predict with 72% accuracy, whether a customer is likely to default next month.
- Using Decision Tree classifier, we can predict with 83% accuracy whether a customer is likely to default next month or not.
- Using Random Forest, we can predict with 74% accuracy whether a customer will be defaulter in next month or not.
- By applying XGBoost Classifier with recall 78%, we can predict with 81.60% accuracy whether a customer is likely to default next month.



- Red column shows the high features values
- Blue column shows low feature values.
- On X axis there are shap values, positive will tell you about defaulter and negative values will tell customers will not default in next month.
- On y-axis, features are ordered in decreasing order in sense of importance for the XGBoost model to predict the default.

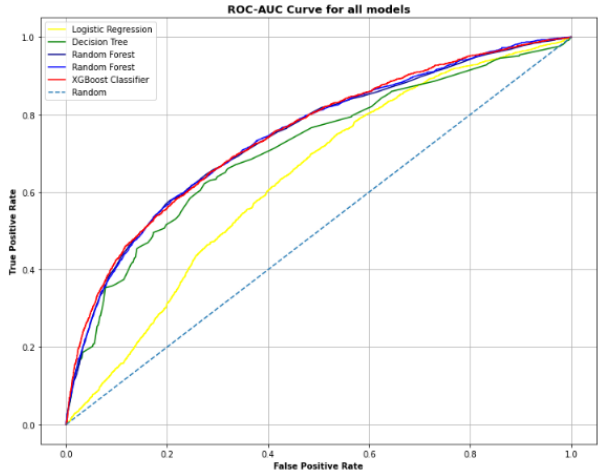
4. model expansibility: -

Model explainability using shap



- 'MARRIGE' is the most important features and pay_1 and SEX are also important.
- pay_amt4 is the least important featers.

5. Plot ROC-AUC Curve:-



6. Conclusion from Model Training: -

Prepare for Modelling

- We use **pair plots** for understanding the data.
- Also created **bins** for **AGE** columns.
- This dataset is also **imbalanced**, with **78%** non-default vs **22%** default.
- We use **SMOTE** because the class is highly **Imbalance**

MODELS

- Using a **Logistic Regression** classifier, we can predict with **68.37%** accuracy, whether a customer is likely to default next month.
- Using **Decision Tree** classifier, we can predict with **73.83%** accuracy whether a customer is likely to default next month or not.
- Using **Random Forest**, we can predict with **78.38%** accuracy whether a

customer will be defaulter in next month or not.

- By applying **XGBoost Classifier** with recall **75%**, we can predict with 81.60% accuracy whether a customer is likely to default next month.

Model Explanation

- 'MARRIGE' is the most important features and pay_1 and SEX are also important.
- pay_amt4 is the least important feature.

References-

1. scikit-learn
2. Matplotlib
3. Seaborn
4. MachineLearningMastery
5. GeeksforGeeks
6. Analytics Vidhya
7. Wikipedia