

Capstone Project - 3

Credit Card Default Prediction

Team Members

Arbaaz Malik

Huzaifa khan

Table of contents:-

- Overview and Objective
- Data Description
- Exploratory Data Analysis
- Feature Engineering
- Handling Imbalance data
- Model Building
- Model Evaluation
- Conclusion

Overview and Objective

Overview

Credit card is a commonly used transaction method in modern society and one of the main business of banks. For banks, it helps the bank to generate interest revenue but at the same time, it raise the liquidity risk and credit risk to the bank. In order to control the cash flow and risk, detecting the customers with default payment next month could play an important roles of estimating the potential cash flow and risk management.

Objective

The main objective of this project is to 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

Understanding attributes of dataset better:-

1. **ID:** ID of each client
2. **LIMIT_BAL:** Amount of given credit in NT dollars (includes individual and family/supplementary credit)
3. **SEX:** Gender (1=male, 2=female)
4. **EDUCATION:** (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
5. **MARRIAGE:** Marital status (1=married, 2=single, 3=others)
6. **AGE:** Age in years

Continued....

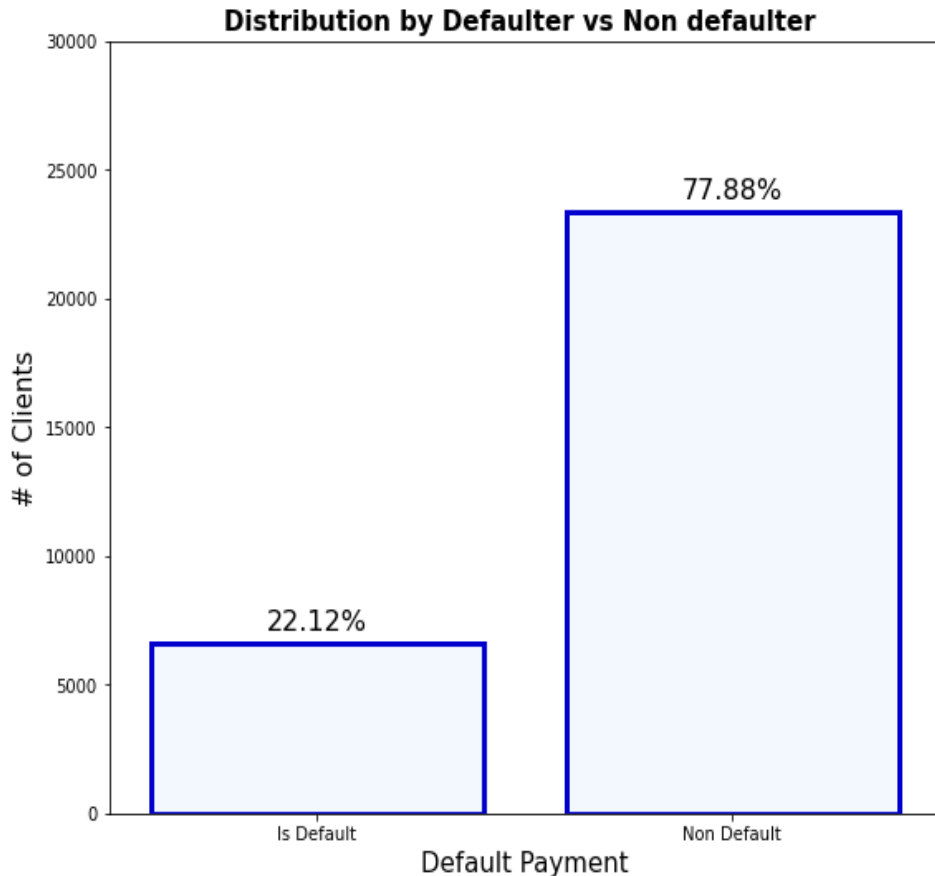
7. **PAY_0-6**: Repayment status in September- April, 2005 (-2 = Unused, -1 = pay duly, 0=Revolving Credit, 1=payment delay for one month, 2=payment delay for two months, 8=payment delay for eight months, 9=payment delay for nine months and above)
8. **BILL_AMT1-6**: Amount of bill statement in September- April, 2005 (NT dollar)
9. **PAY_AMT1-6**: Amount of previous payment in September- April, 2005 (NT dollar)
10. **default.payment.next.month**: Default payment (1=yes, 0=no)

Exploratory Data Analysis (EDA)

Dependent variable

We can see that the dataset consists of more than 20000 clients who are not expected to default payment whereas around 5300 clients are expected to default the payment. Here, there is huge difference between non-defaulter(0) and defaulter(1).

Approx 78% are Non Defaulters and 22% are Defaulters respectively.



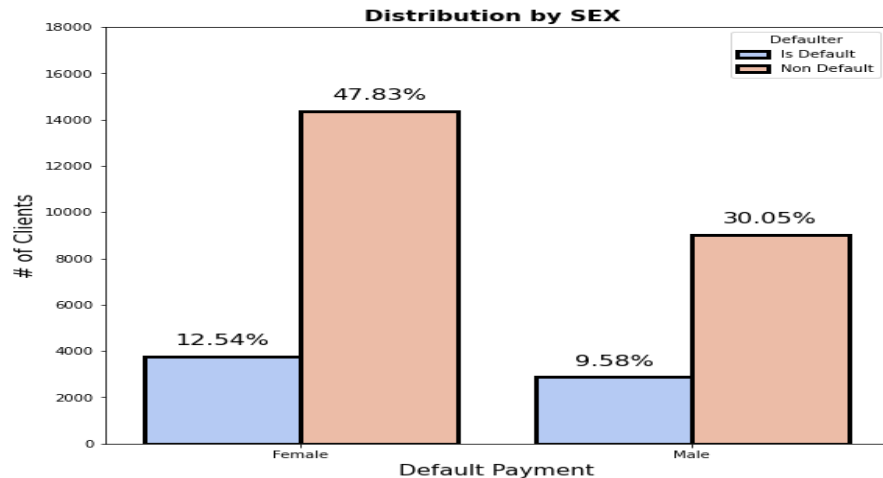
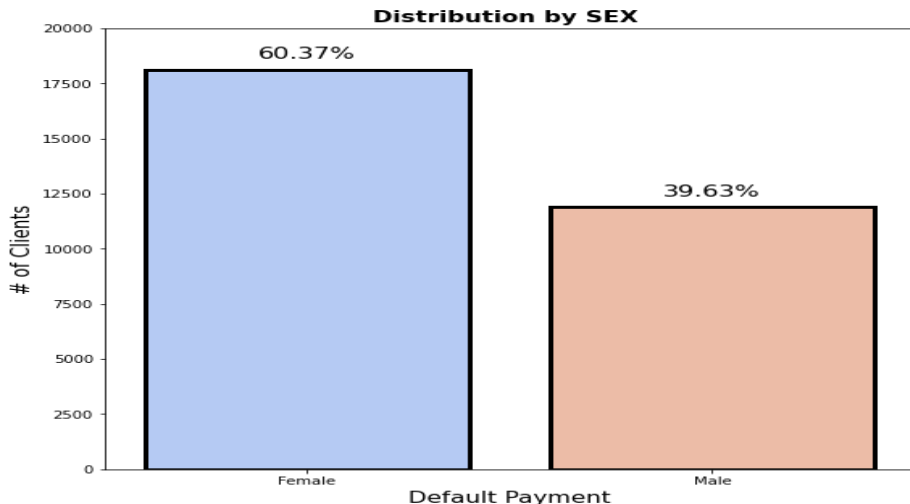
EDA Continued...

SEX

Number of Male credit holder (represented as 1) is less than Female (represented as 2).

Approximately 40% are male and 60% are Female.

It is evident from the second graph that the number of defaulter have high proportion of females.



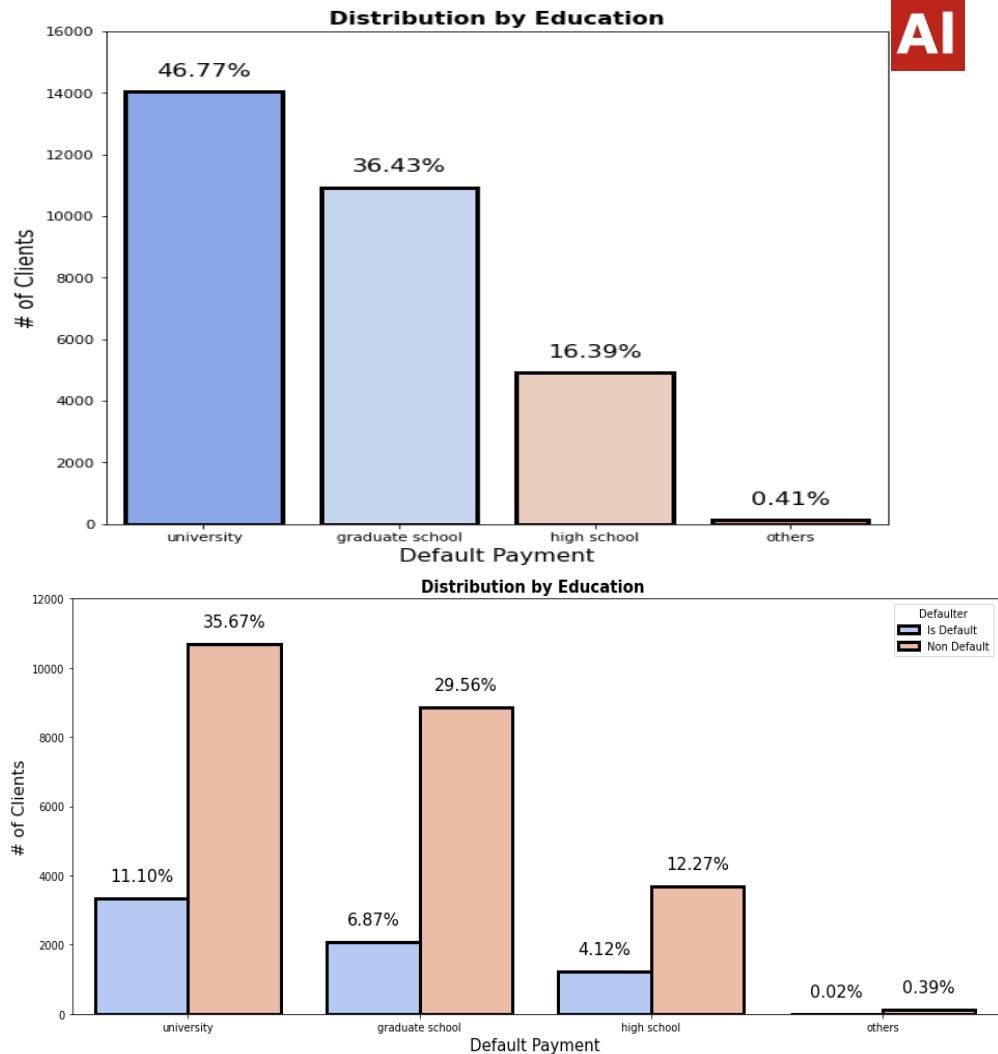
EDA Continued...

EDUCATION

More number of credit holders are university students followed by Graduates and then High school students.

From the 2nd plot it is clear that those people who are university students have higher default payment w.r.to graduates and high school people.

From university 11% are default, from graduate 7% are default, and from high school 4% are default.



EDA Continued...

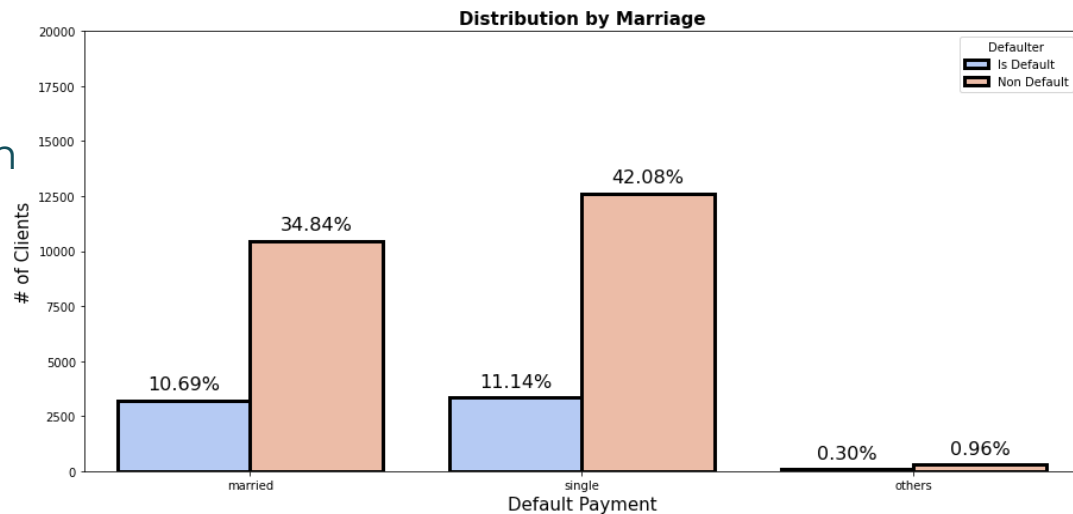
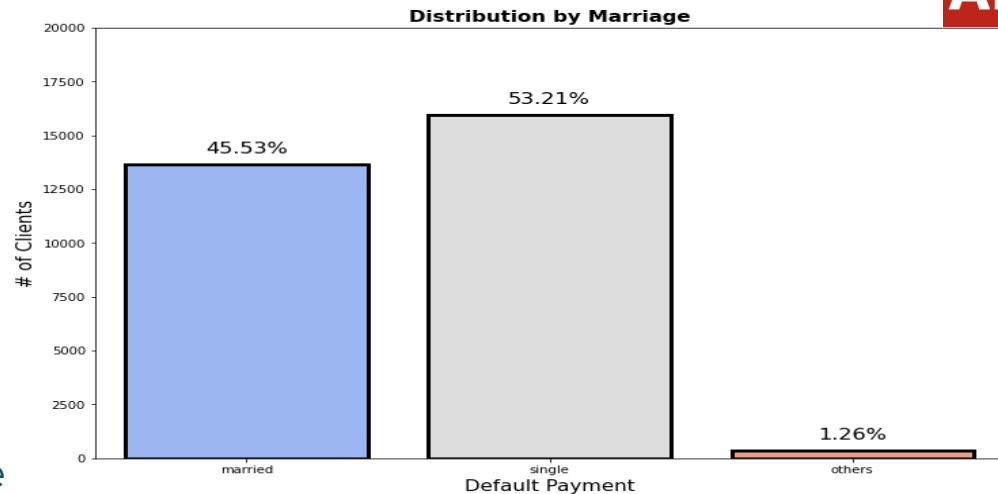
AI

MARRIAGE

From graph 1 we can say that more number of credit cards holder are Single as compared to married and others

Here it seems that married ,single are most likely to default.

From single 11% are default and from married approx 11% are defaulter.

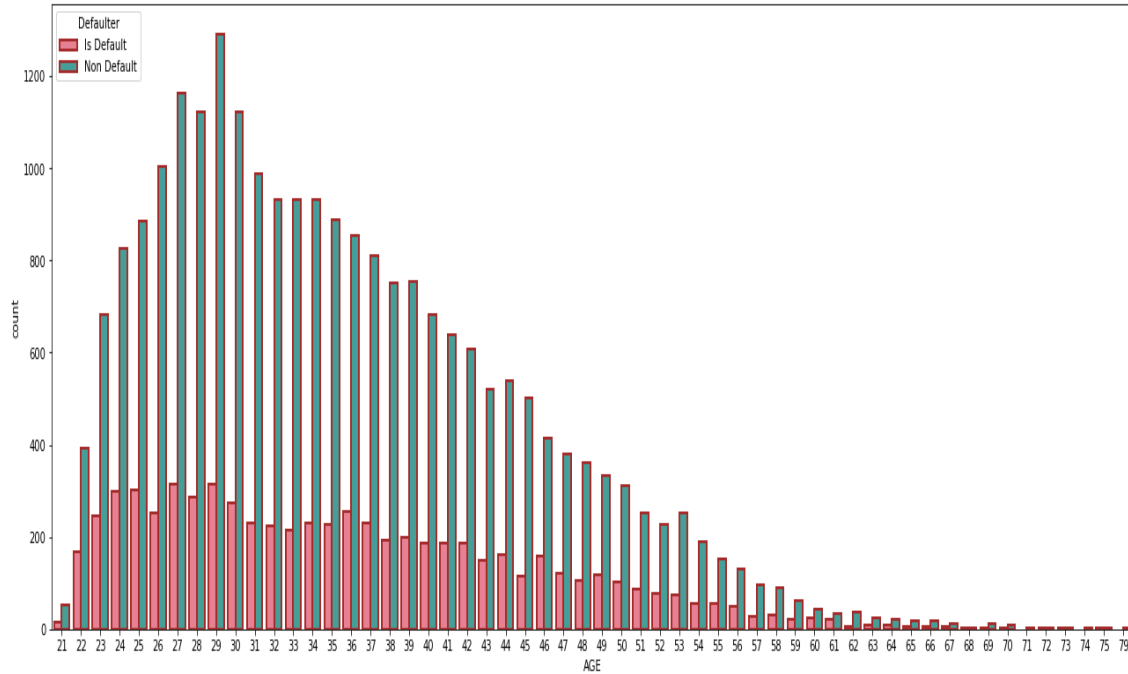


EDA Continued...

AGE

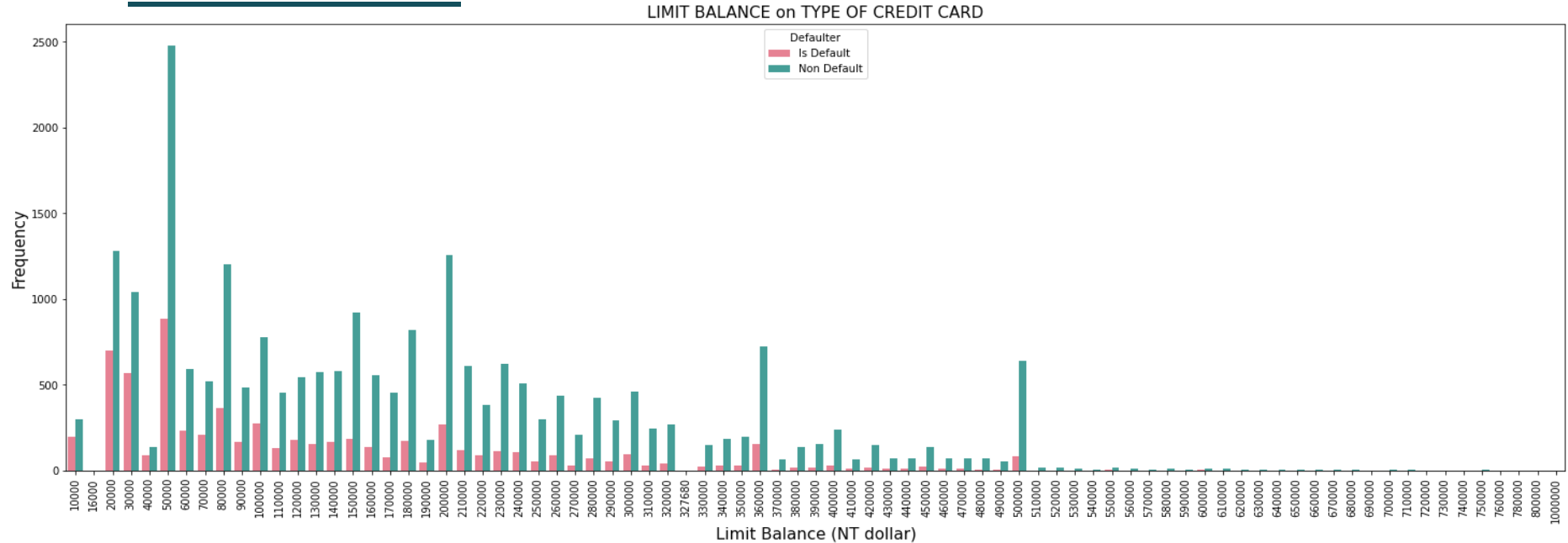
More number of credit card holders age between 26-32 years and 29 years age is the highest uses of credit card. Age above 60 years old rarely uses the credit card.

Also more number of Defaulters are between 27-29 years.



EDA Continued...

LIMIT BALANCE

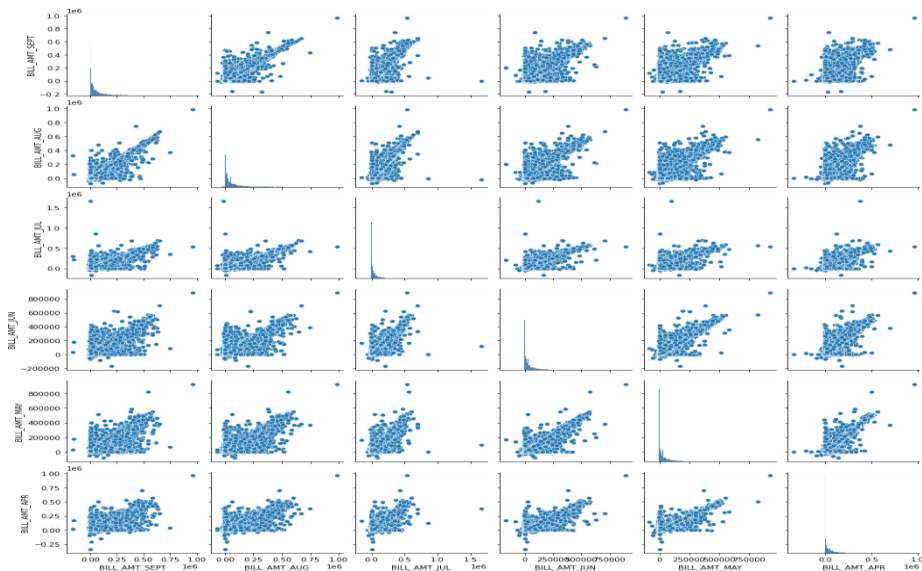


Maximum amount of given credit in NT dollars is 50,000 followed by 20,000 and 30,000. And Defaulters are between this Limit Balance only.

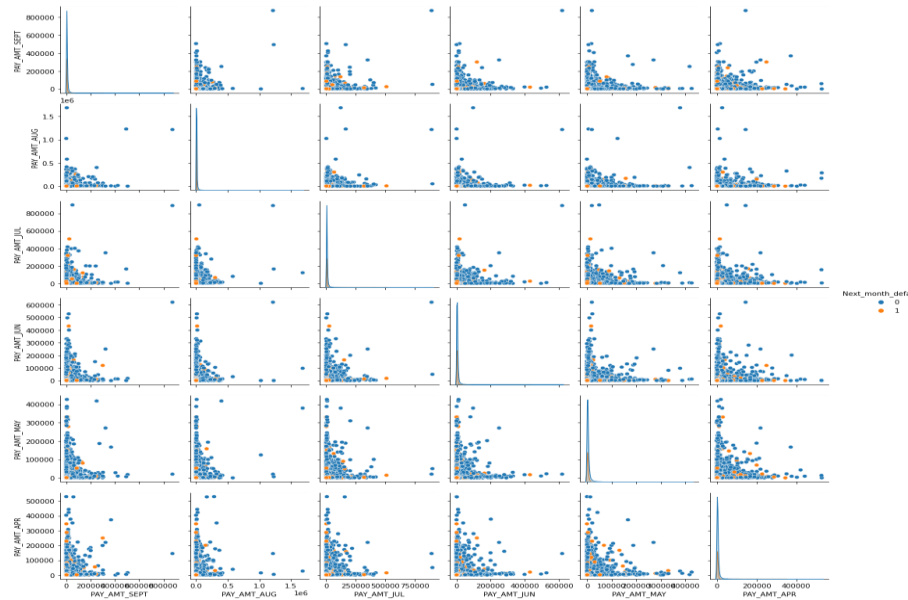
EDA Continued...



Pairplot of bill



Pairplot of pay



The 1st pairplot shows the distribution of bill amount statements for each month explicitly for defaulters and non-defaulters.

The 2nd pairplot shows the distribution of payment statements for each month explicitly for defaulters and non-defaulters.

FEATURE ENGINEERING



ONE HOT ENCODING:

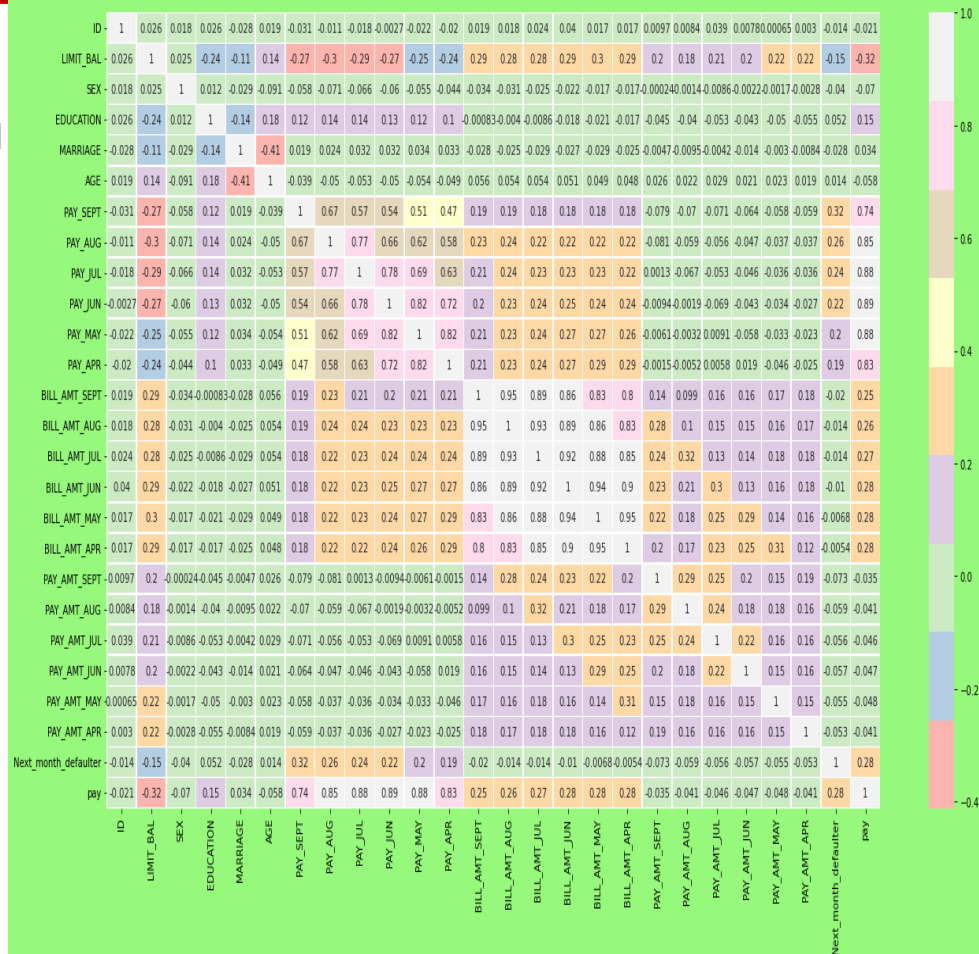
One hot encoding is a process by which categorical variables are converted into a numerical variables.

Here we perform one hot encoding on 'EDUCATION', 'MARRIAGE', and 'SEX'.

Correlation Analysis:

We draw heatmap to find correlation between different independent features and dependent feature.

We remove columns which are not important for further analysis such as ID, AGE,DEFAULTER and PAY.



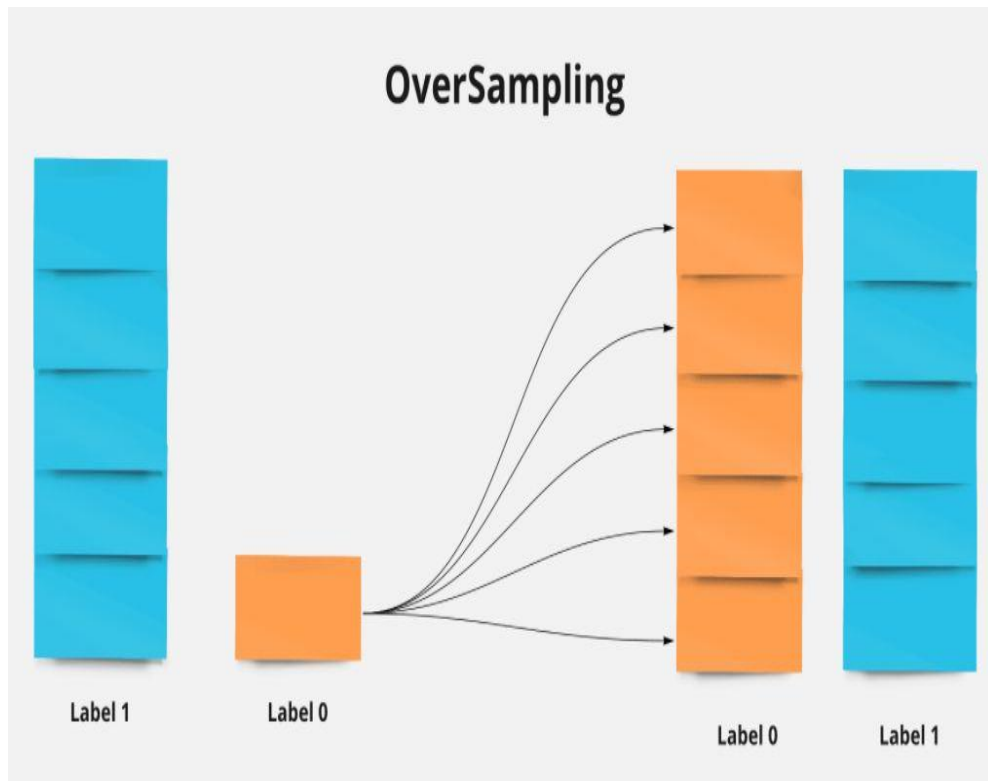
RESAMPLING:

Why resampling?

- Our dataset is imbalanced

Solution:

- Create new training dataset:
(on X train and y train)
- ❑ Oversampling training data consists in over-sizing the minority class by adding observations.



MODEL BUILDING:

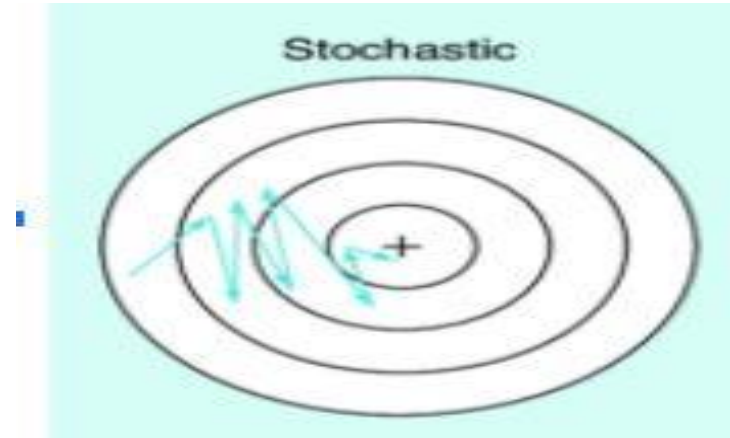
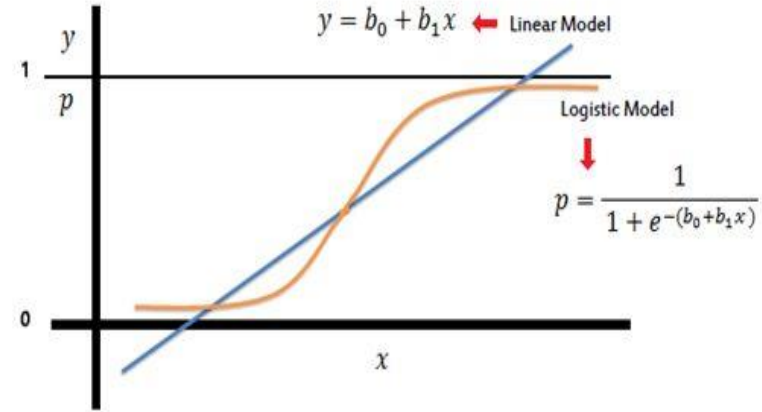
LOGISTIC REGRESSION

Logistic Regression is a Machine Learning algorithm and it is basically used for binary classifications like yes-no, true-false, male-Female, etc.

It takes the linear combination and applies a sigmoid function (logit). The Sigmoid curve gives a value between 0 & 1.

Stochastic Gradient Descent

In SGD, it uses only a single sample, i.e., a batch size of one, to perform each iteration. The sample is randomly shuffled and selected for performing the iteration.

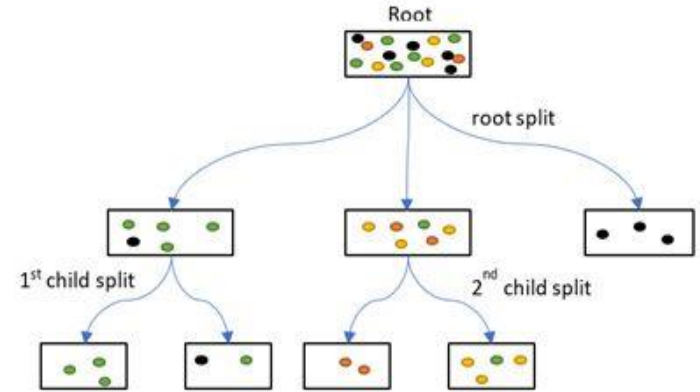


MODEL BUILDING (continued):

DECISION TREE CLASSIFIER

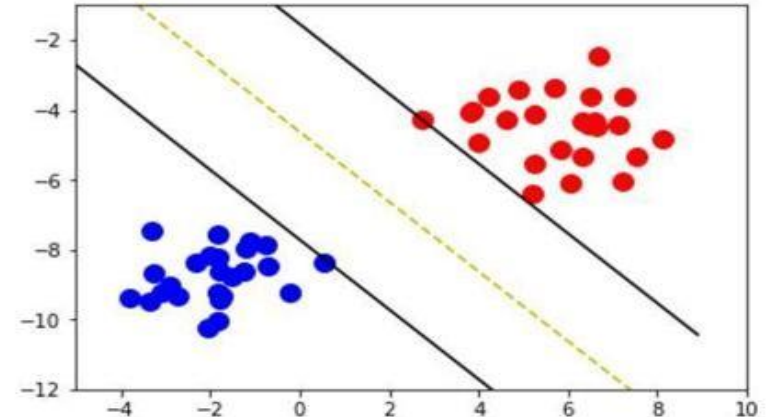
It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

The objective of Decision tree algorithm is to find the relationship between the target column and the independent variables and Express it as a tree structure.



SUPPORT VECTOR MACHINE

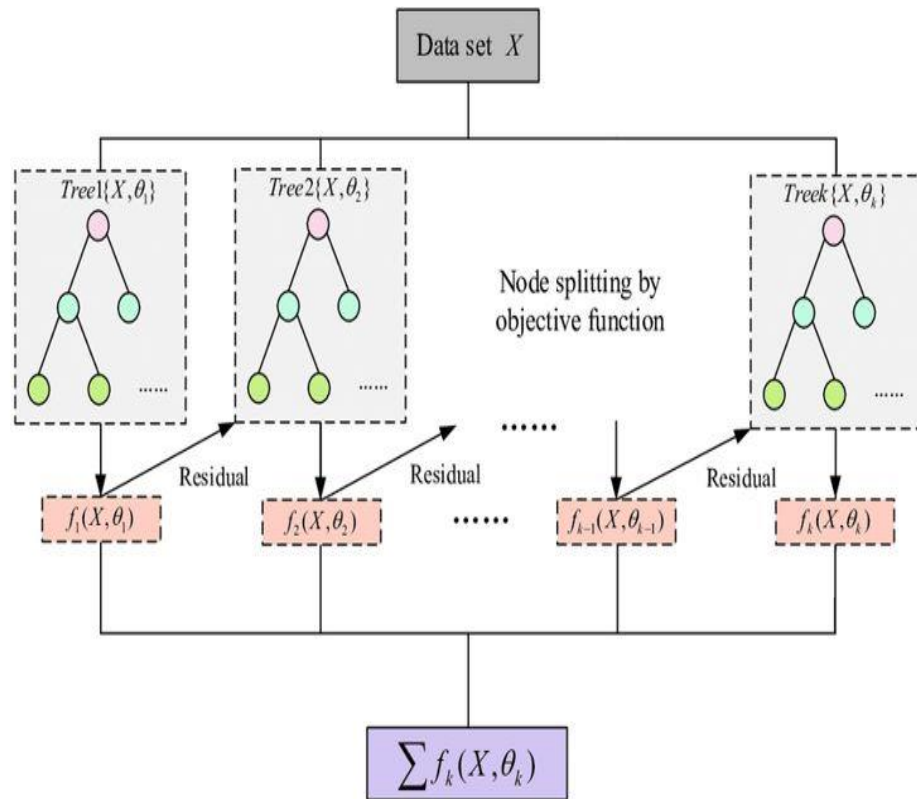
- Maximize the distance from the yellow line (decision boundary) that separates the data
- Black lines are support vectors that used to determine the decision boundary
- Can be used to classify non-linear relationship



MODEL BUILDING (continued):

XG BOOST CLASSIFIER

- Stands for:– e**X**treme **G**radient **B**oosting.
- XGBoost is a powerful iterative learning algorithm based on gradient boosting.
- Regularisation to avoid overfitting
- Tree pruning using depth-first approach
- It is generally used for very large dataset

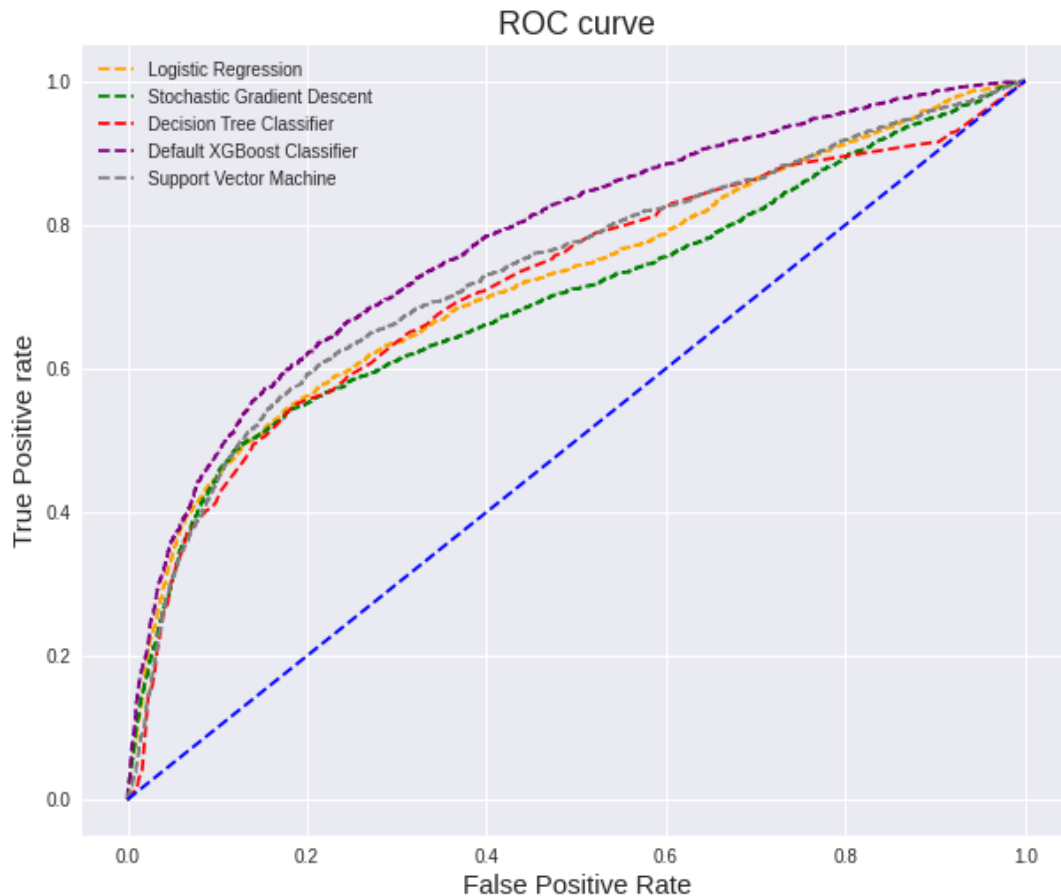


MODEL EVALUATION:

	Models Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score	ROC_AUC Score
0	Logistic Regression	0.675171	0.693889	0.385014	0.6320	0.478516	0.671786
1	Stochastic Gradient Descent	0.501131	0.226222	0.223053	0.9995	0.364714	0.502393
2	Decision Tree Classifier	0.771388	0.718889	0.407666	0.5850	0.480493	0.671071
3	Default XGBoost Classifier	0.732309	0.763333	0.474787	0.6120	0.534731	0.709286
4	Support Vector Machine	0.731087	0.776333	0.497025	0.5430	0.518996	0.693000

Let us look at ROC Curve!!

Receiver Operating Characteristic (ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate (1-specificity). For plotting ROC, it is advisable to assume $p > 0.5$ since we are more concerned about success rate.



Conclusion:

- 1.** We rename the column for each features for better understanding.
- 2.** From above graph we check the distribution of defaulter vs non defaulter and we see around 78% are non defaulter and 22% are defaulter. Also we check for Marriage, Education, Sex with respect to defaulter and we found in marriage more number of defaulter is Female, in Education more no. of defaulter is university and in Marriage more no. of defaulter is single.
- 3.** We see that with the help of correlation matrix age and marriage are highly negatively correlated to each other and we drop some columns we are not in used.
- 4.** After that we build the Five models Logistic Regression, Stochastic gradient descent, Decision Tree, Default XGBoost Classifier & Support vector machine and inspite of all the models, the best accuracy is obtained from the Default XGBoost Classifier.

Continued...

- a.)** Using a Logistic Regression classifier, we can predict with 69.38% accuracy, whether a customer is likely to default next month.
 - b.)** With Stochastic Gradient Descent classifier, 58.84% customer is likely to default next month.
 - c.)** With Decision Tree classifier, 71.88% customer is likely to default next month.
 - d.)** With Default XGBoost Classifier, 76.33% customer is likely to default next month.
 - e.)** And with Support Vector Machine classifier, 77.63% customer is likely to default next month.
- 5.** From above table Default XGBoost Classifier and Support Vector Machine are giving us the best Precision, 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.

THANK YOU!!