Capstone Project - 3

Credit Card Default Detection

Team Members

Amir Khan
Saurabh Daund
Het Kothari
Kamya Malhotra
Mouleena Jaiswal

Table of contents:

Introduction

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.

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 Overview

- Understanding attributes of dataset better

ID	ID of each client
LIMT_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, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0-6	Repayment status in September-April, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months,…)
BILL_AMT1-6	Amount of bill statement in September-April, 2005 (NT dollar)
PAY_AMT1-6	Amount of previous payment in September-April, 2005 (NT dollar)
default.payment.next.month	Default payment (1=yes, 0=no)

Continued...

- Summarize the data by identifying key characteristics, such as data volume and total number of variables in the data.

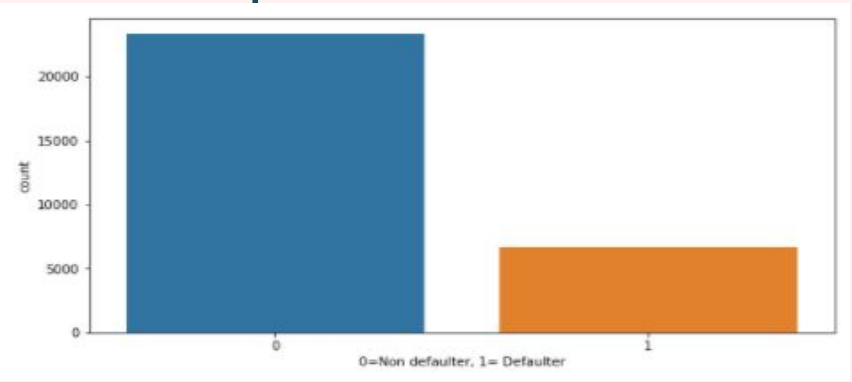
  Number of rows in our dataset are 30000

  Number of columns in our dataset are 25.

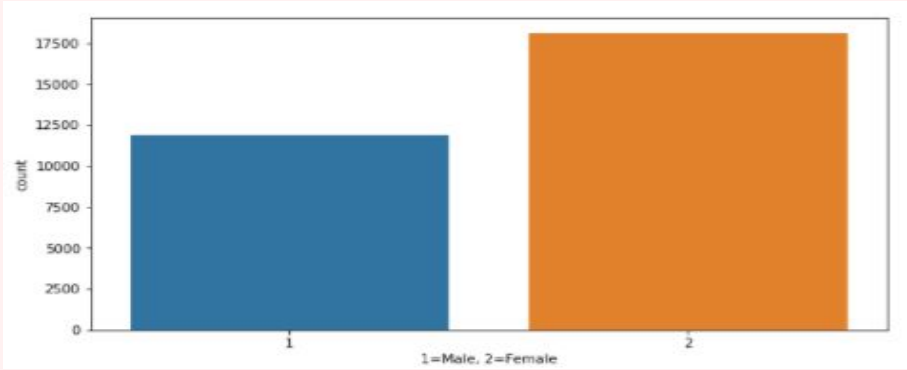
- Understand the problems with the data, such as missing values, inaccuracies, and outliers.
- There are no NAN/NULL values in our dataset.

EDA Univariate analysis

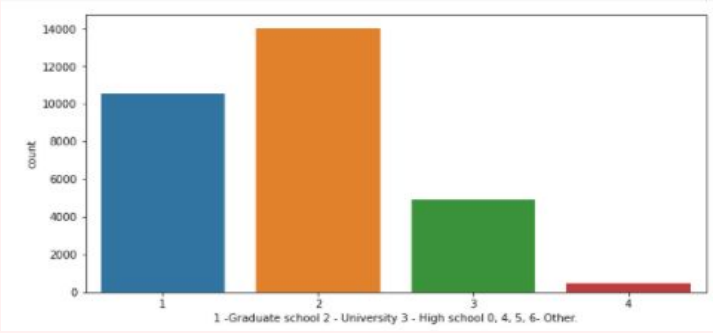
Dependent variable



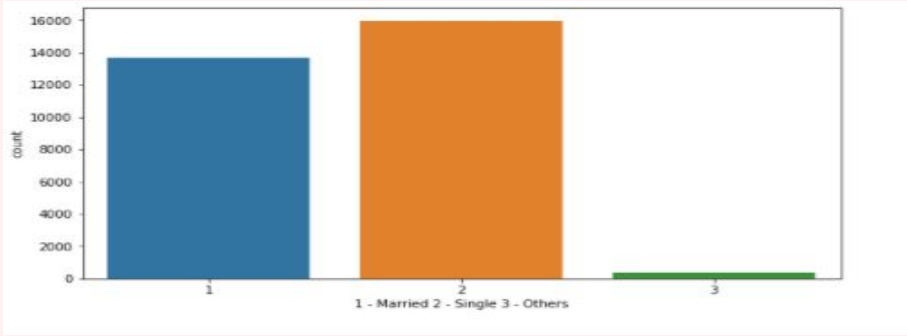
Sex



Education

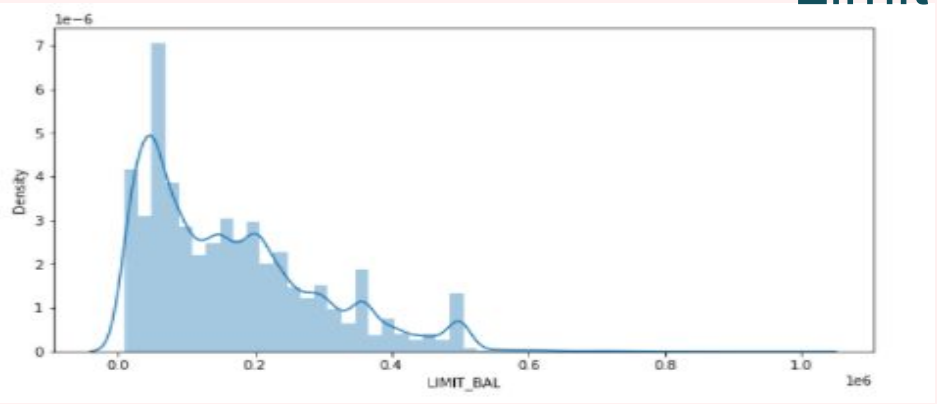


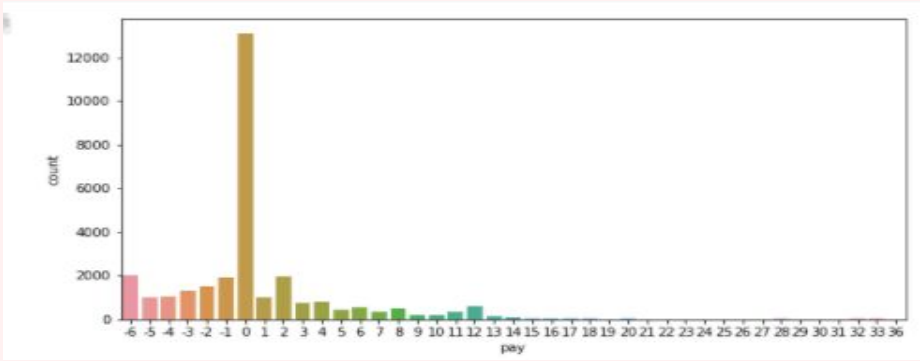
Marriage



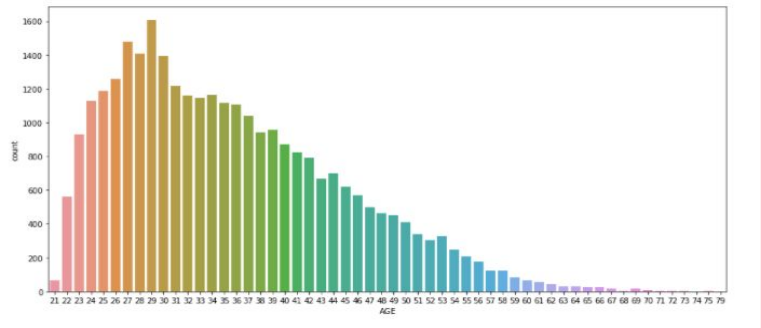
EDA Univariate analysis

Limit Balance



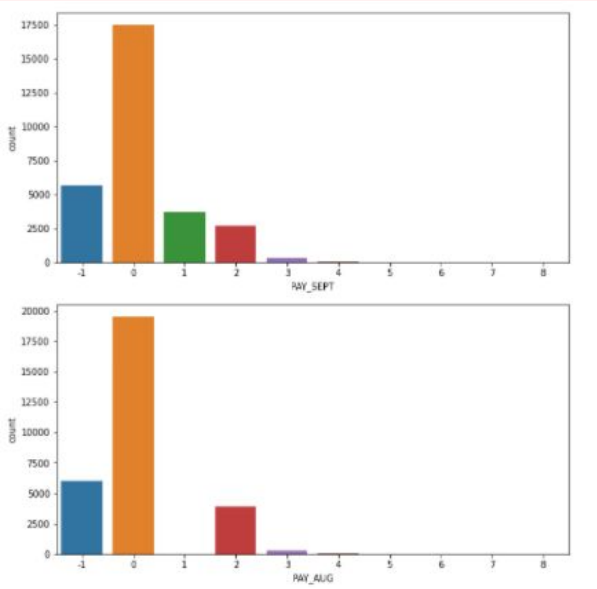


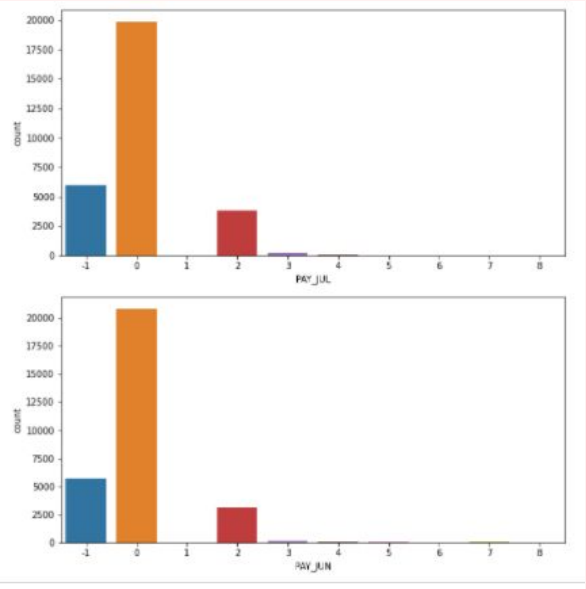
Age

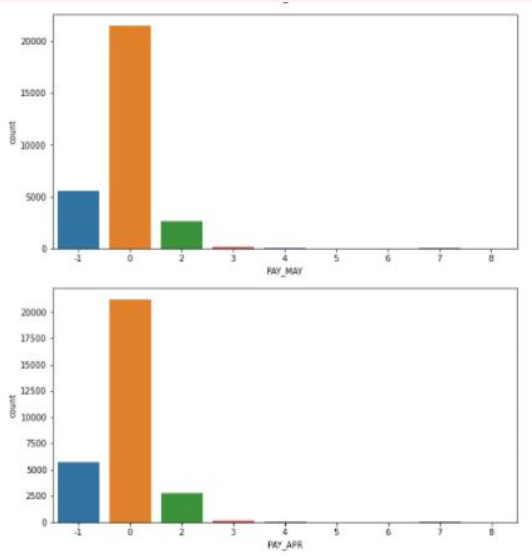


EDA Univariate analysis

Pay_Apr-Sept





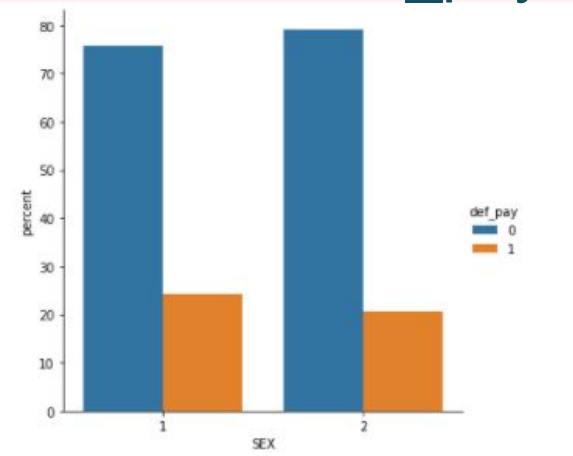


Insights from Univariate Analysis:

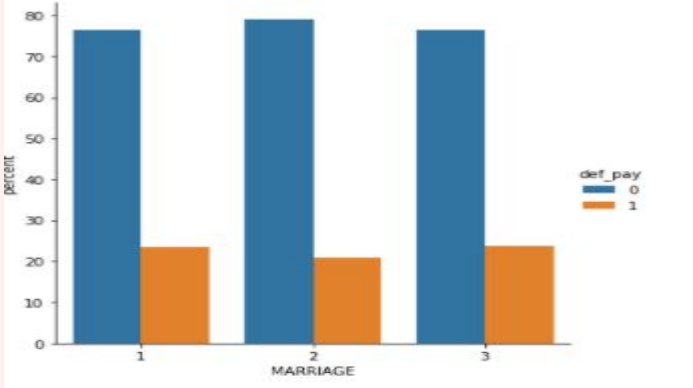
- We can see that the dataset consists of more than 20000 clients who are not expected to default payment whereas around 5300 clients are expected to default the payment.Here ,there is huge difference between non-defaulter(0) and defaulter(1)
- Number of Male credit holder(represented as 1) is less than Female(represented as 2).
- More number of credit holders are university students(represented as 2) followed by Graduates(represented as 1) and then High school students(represented as 3).Here 0, 4, 5, and 6 can be treated as other.
- More number of credit cards holder are Single.
- Mostly ,Payment are not due(0) for april to september.
- 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.Maximum amount of given credit in NT dollars is 50,000 followed by 20,000 and 30,000.

EDA Bivariate analysis

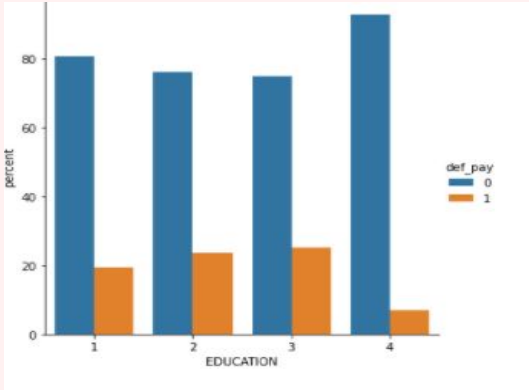
Sex and def_pay



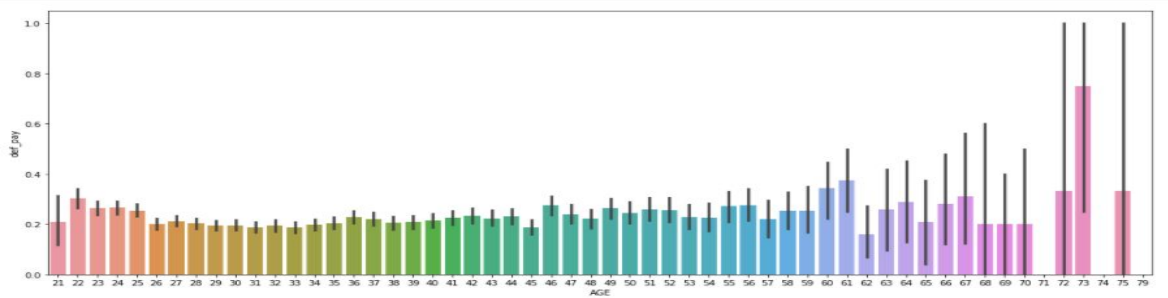
Marriage and def_pay



Education & def_pay

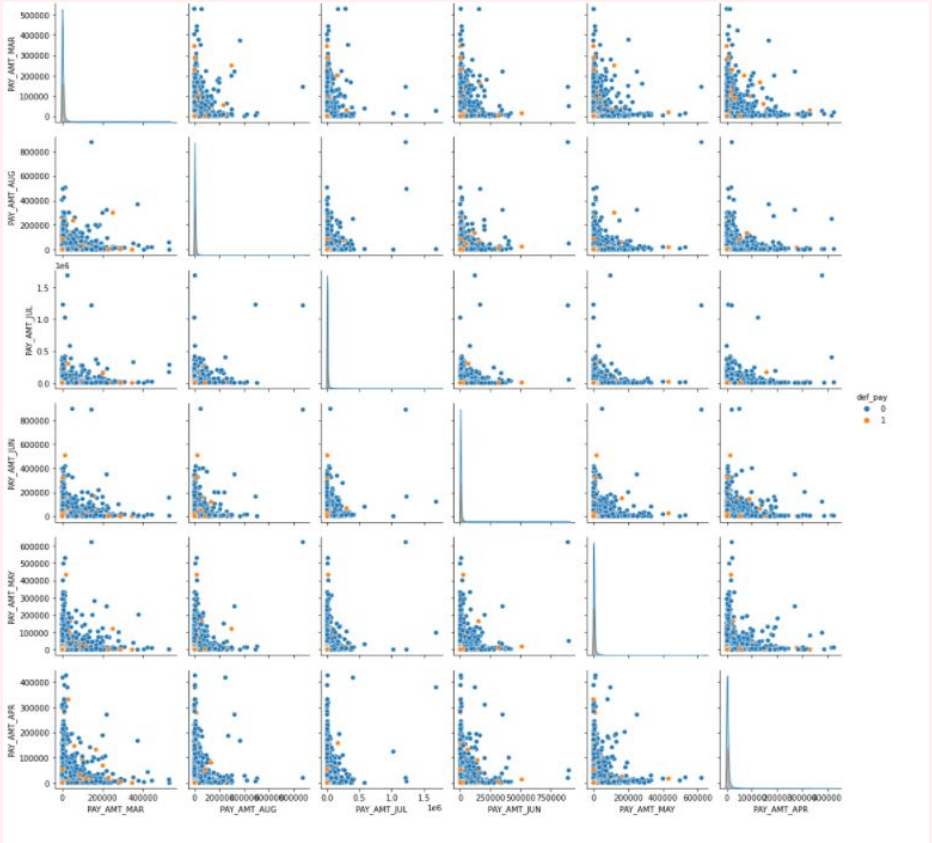


Age and def_pay

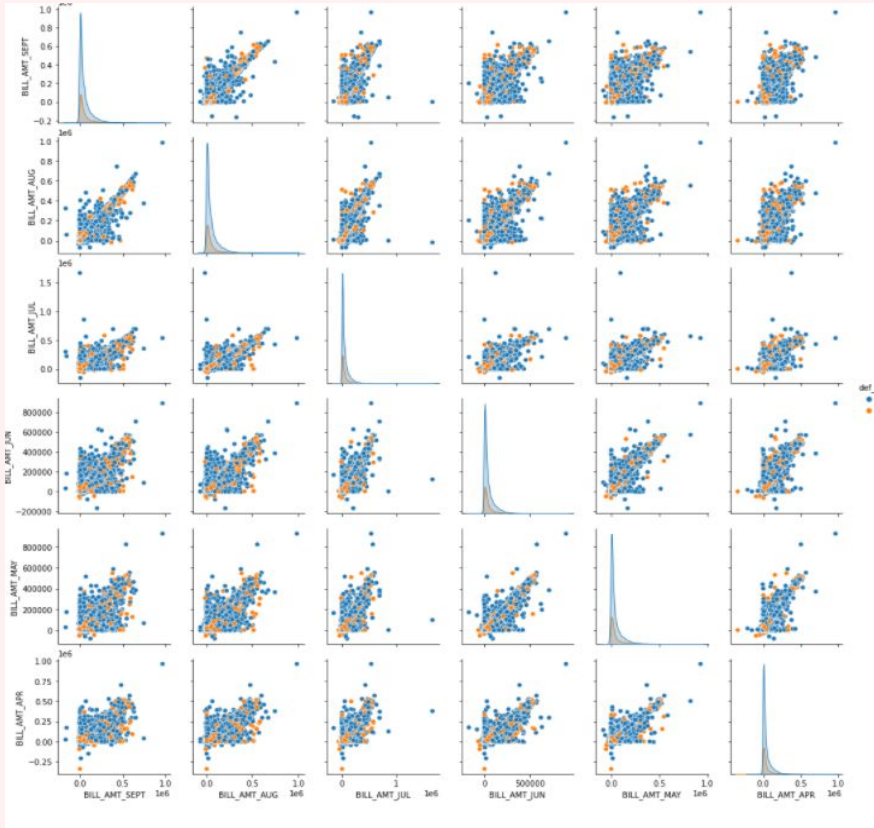


EDA Bivariate analysis

Pairplot of pay

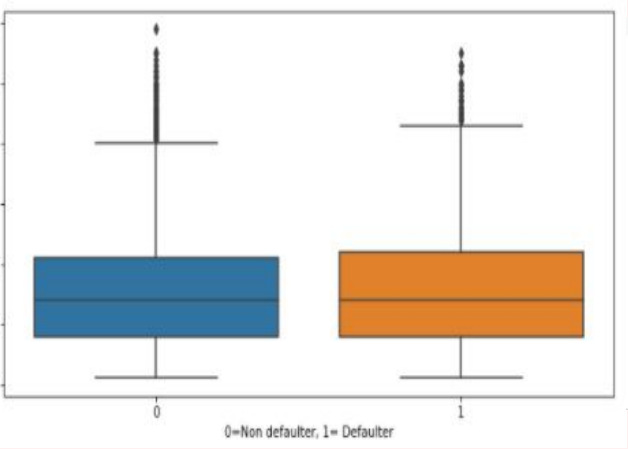


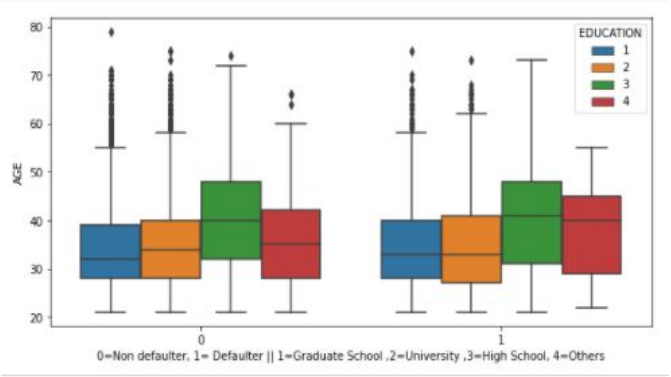
Pairplot of bill

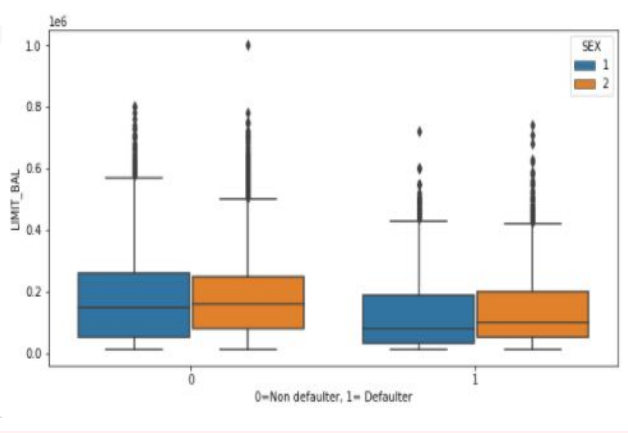


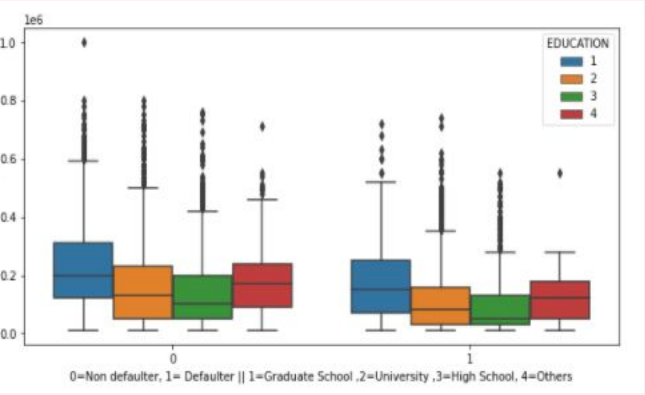
EDA Bivariate analysis

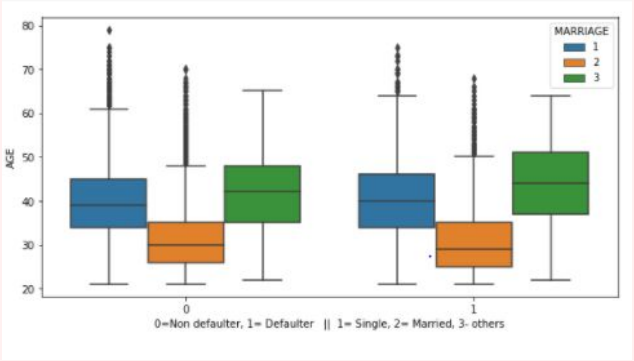
Boxplot

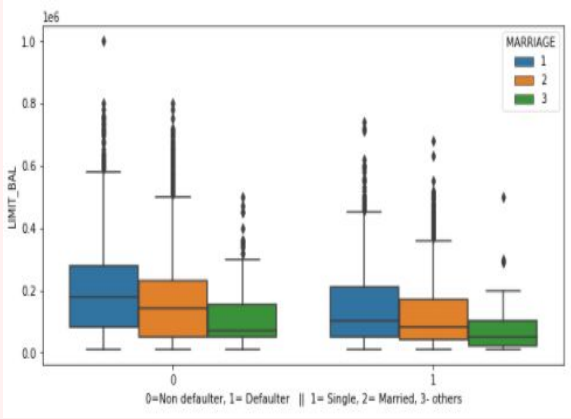












Insights from Bivariate Analysis:

- It is evident from the above graph that the number of defaulter have high proportion of females.
- From the above plot it is clear that those people who are high school students have higher default payment wrt graduates and university people
- Here it seems that married ,single and other are most likely to default.
- Most number of defaulters are between the age of 25-30 years old.
- The pairplot shows the distribution of bill amount statements for each month explicitly for defaulters and non-defaulters.
- The pairplot shows the distribution of payment statements for each month explicitly for defaulters and non-defaulters.

Feature Engineering

One Hot Encoding :

Dummify features like EDUCATION and SEX

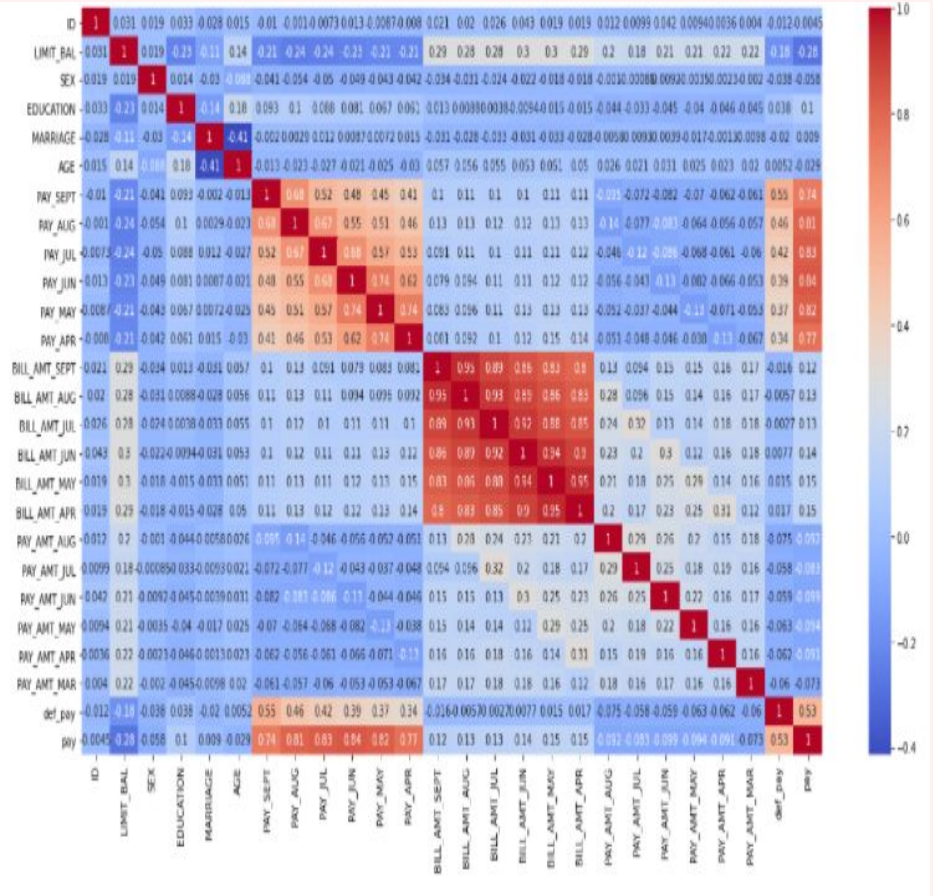
Correlation Analysis:

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

If correlationbetween independent features are high and has very less relation with dependent feature, remove them.

We remove columns which are not important for further analysis such as

`ID,AGE,MARRIAGE,BILL_AMT_MAY,BILL_AMT_APR,BILL_AMT_JUN,PAY_AMT_MAY,PAY_AMT_APR,PAY_AMT_MAR`



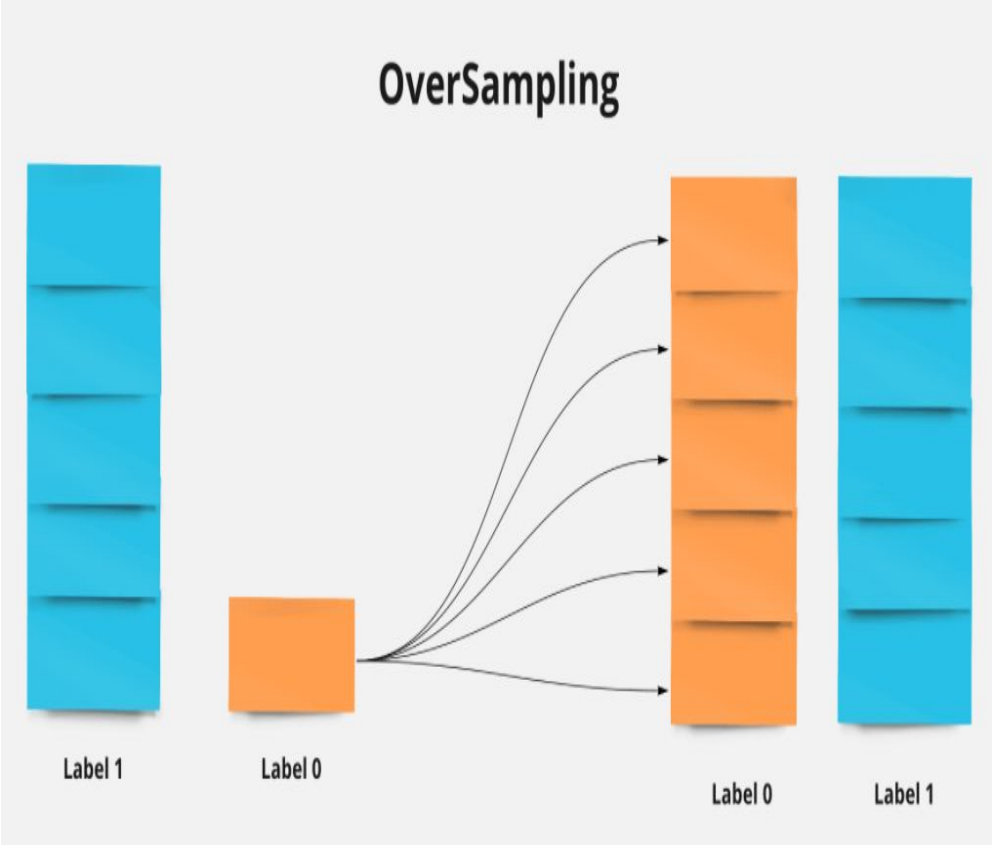
Resampling

Why resampling?

- Our dataset is imbalanced

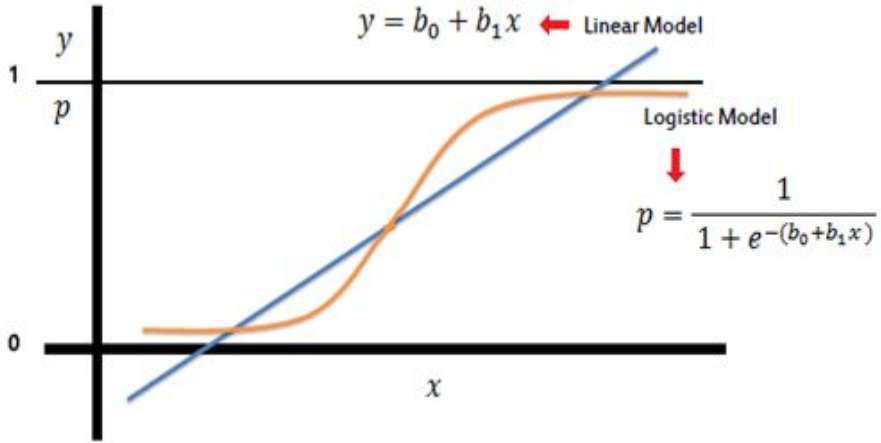
Solution

- Create new training dataset:
  - Oversampling training data
    - consists in over-sizing the minority class by adding observations.



Model Creation

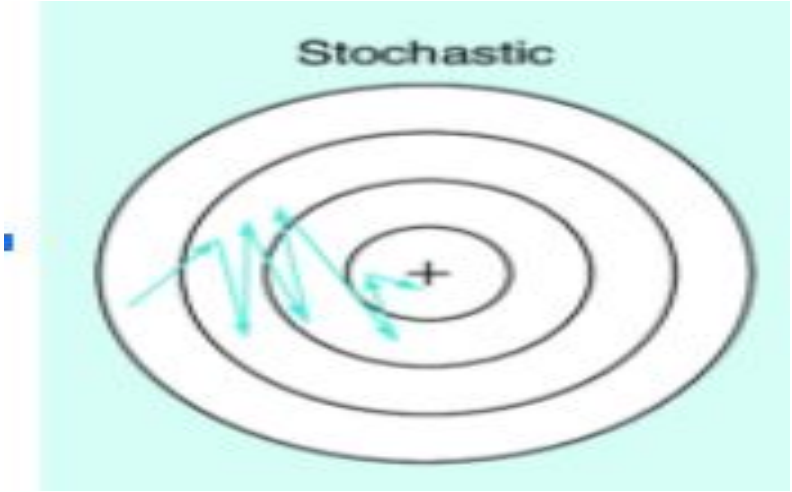
Logistic Regression



- Simplest parametric model in classification
- Take the linear combination and apply a sigmoid function (logit)

Model Creation(continued)

Stochastic Gradient Descent

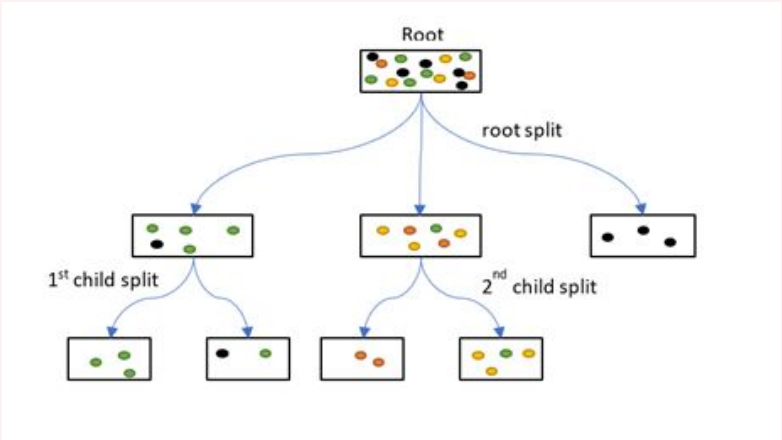


In SGD, it uses only a single sample, i.e., a batch size of one, to perform each iteration. The sample is randomly shuffled and selected for performing the iteration.

Model Creation(continued)

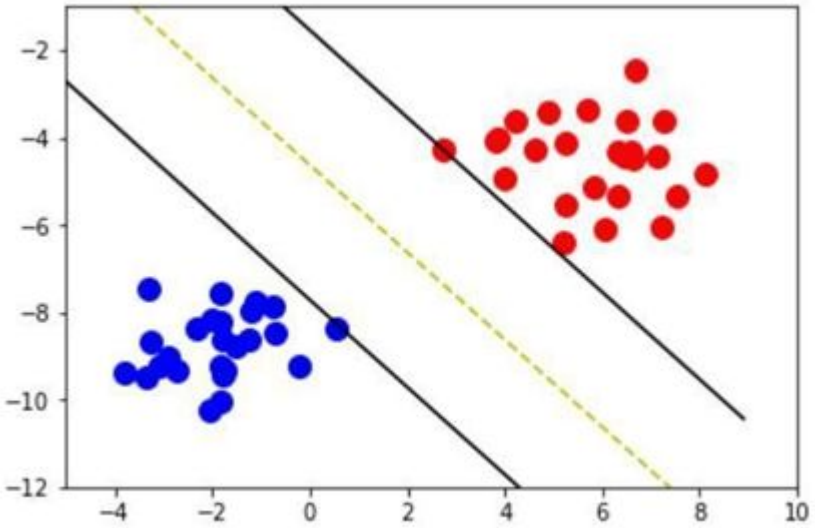
Decision Tree Classifier

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



Model Creation(continued)

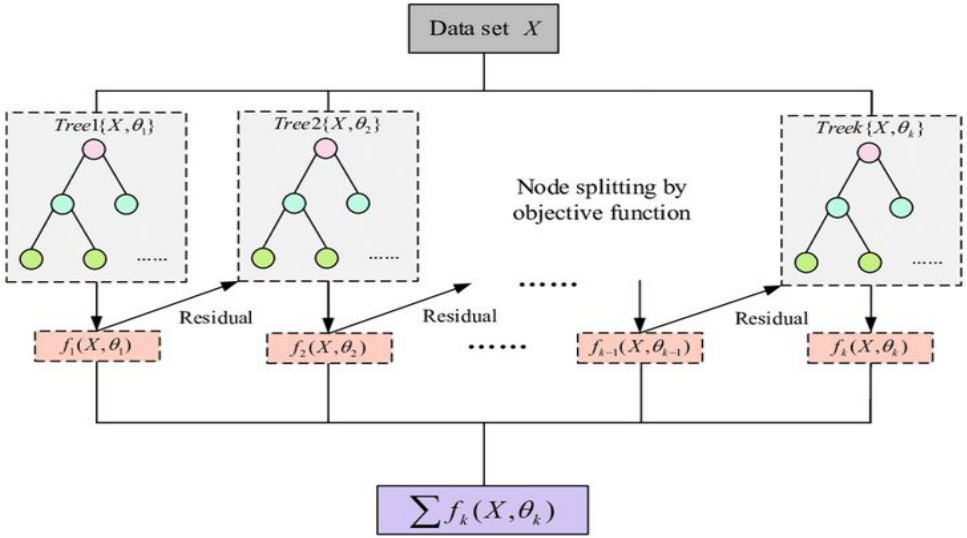
SVM



- 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 Creation(continued)

XGBOOST



- Stands for:

–eXtreme Gradient Boosting.

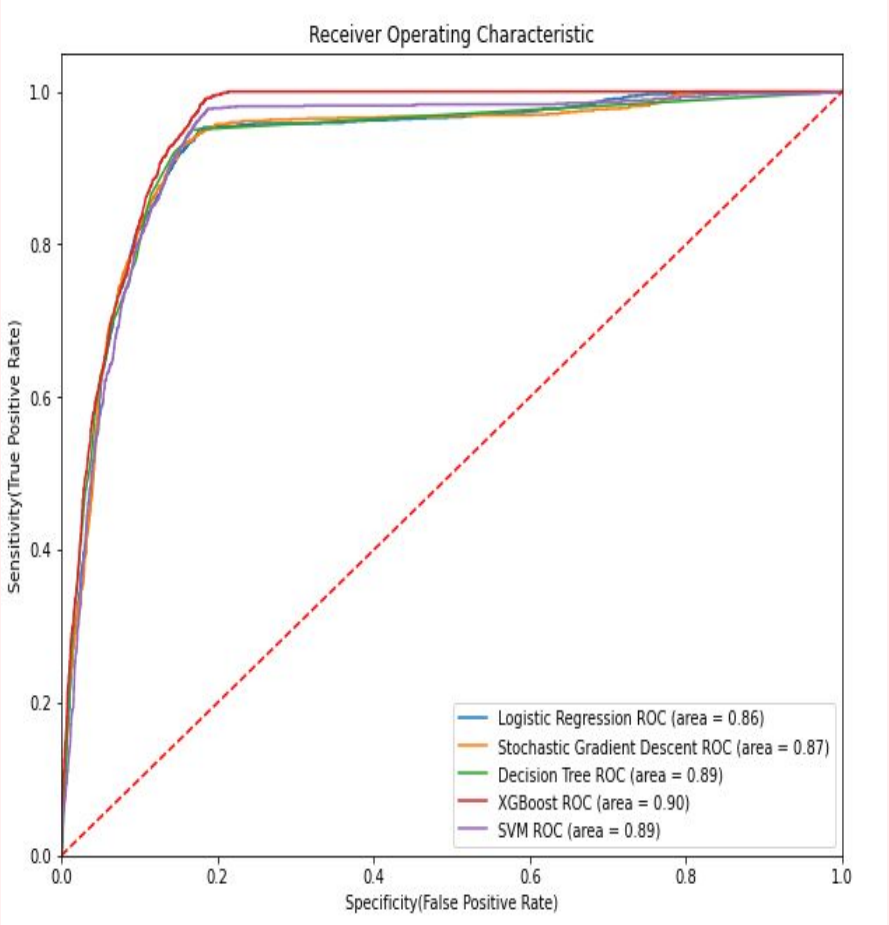
- XGBoost is a powerful iterative learning algorithm based on gradient boosting.
- Regularisation to avoid overfitting
- Tree pruning using depth-first approach
- It is generally used for very large dataset

Model Evaluation

	Model	Accuracy	Precision	Recall	F1 Score	ROC Score
0	Logistic Regression	0.880648	0.559933	0.843645	0.673115	0.865301
1	Stochastic Gradient Descent	0.879674	0.556955	0.850334	0.673064	0.867505
2	Decision Tree Classifier	0.849105	0.490901	0.969900	0.651869	0.899205
3	Default XGBoost Classifier	0.852028	0.495907	0.962375	0.654535	0.897795
4	GridSearch XGBoost Classifier	0.872610	0.540409	0.838629	0.657274	0.858516

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

- The best **accuracy** is obtained for the **Logistic Regression** and **Stochastic Gradient Descent.**
- 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.
- Furthermore, **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 **Default XGBoost Classifier** having **Recall**, **F1-score**, and **ROC Score** values equals 96%, 65%, and 89% and **Decision Tree Classifier** having **Recall**, **F1-score**, and **ROC Score** values equals 97%, 65%, and 90%.
- **Default XGBoost Classifier** and **Decision Tree 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.
-

Challenges

- Large dataset to handle.
- Need to Remove outliers
- Carefully handled feature imbalanced data .
- Carefully tuned Hyperparameters .

THANK YOU!!