Capstone Project

Credit Card Default Prediction

Sumaya Bai A R

CONTENT

Summary

One of the major concerns of Banks are defaulters and fradulers.credit card companies are those companies who have a significant interest in predicting which customers will default on their payments because such defaults cost them money, and thus, they would rather not extend money to individuals with a high probability of default. A good prediction model will enable them to lend to good customers.

It is always beneficial for these kind of financial institution to know their defaulters beforehand.

Here, I've used Credit Card Default Prediction Dataset.

Dataset name : Credit Card Default Prediction Dataset.

Shape : Rows : 30000 Columns : 25

Problem Statement

The main motive behind this project is to identify the defaulters of credit card with their historical data such as demographic condition, repayment statuses, bill statements, history of payments with the help of Machine learning techniques.

The issuer of credit will get to know the likelihood of defaulters and they can decide to lend them credit or not to lend.

It would also help the issuer have a better understanding of their customers, so as to help them build their future strategy.

All of these could be achieved using Machine Learning techniques.

Dataset Variable Description

1. X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
2. X2: Gender (1 = male; 2 = female).
3. X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
4. X4: Marital status (1 = married; 2 = single; 3 = others).
5. X5: Age (year).
6. X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -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.
7. X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
8. X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Data Wrangling

I've renamed the column names for better and easy understanding.

Hence forth below columns are renamed as :

```
Unnamed: 0
X1
X2
X3
X4
X5
X6
X7
X8
X9
X10
X11
X12
X13
X14
X15
X16
X17
X18
X19
X20
X21
X22
X23
Y
```

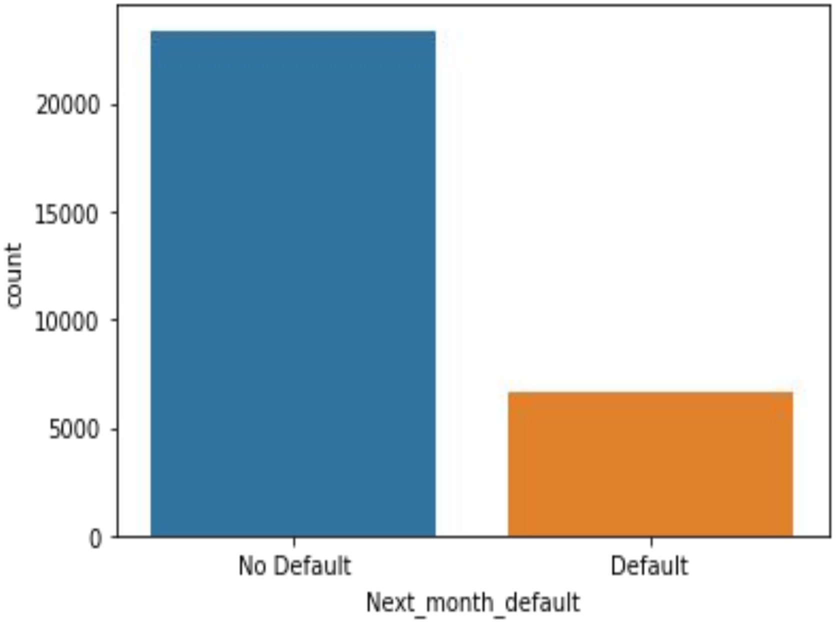
Column Name

Renamed Column names

```
ID
Credit_given
Gender
Education
Marriage
Age
Paymment_his_april
Paymment_his_may
Paymment_his_june
Paymment_his_july
Paymment_his_august
Paymment_his_sept
BILL_AMT_april
BILL_AMT_may
BILL_AMT_june
BILL_AMT_july
BILL_AMT_august
BILL_AMT_sept
PAY_AMT_april
PAY_AMT_may
PAY_AMT_june
PAY_AMT_july
PAY_AMT_august
PAY_AMT_sept
Next_month_default
```

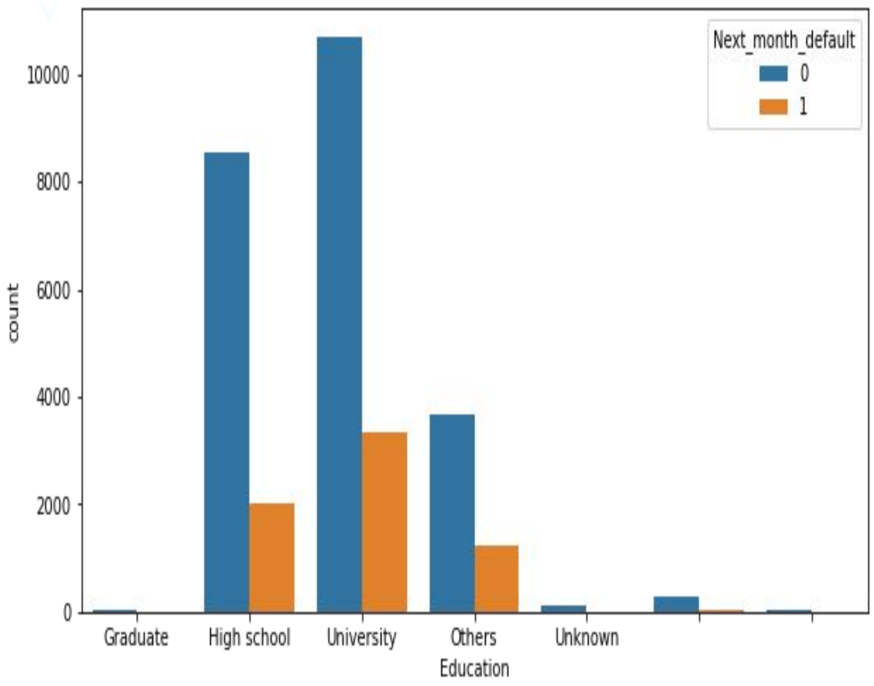
Exploratory Data Analysis

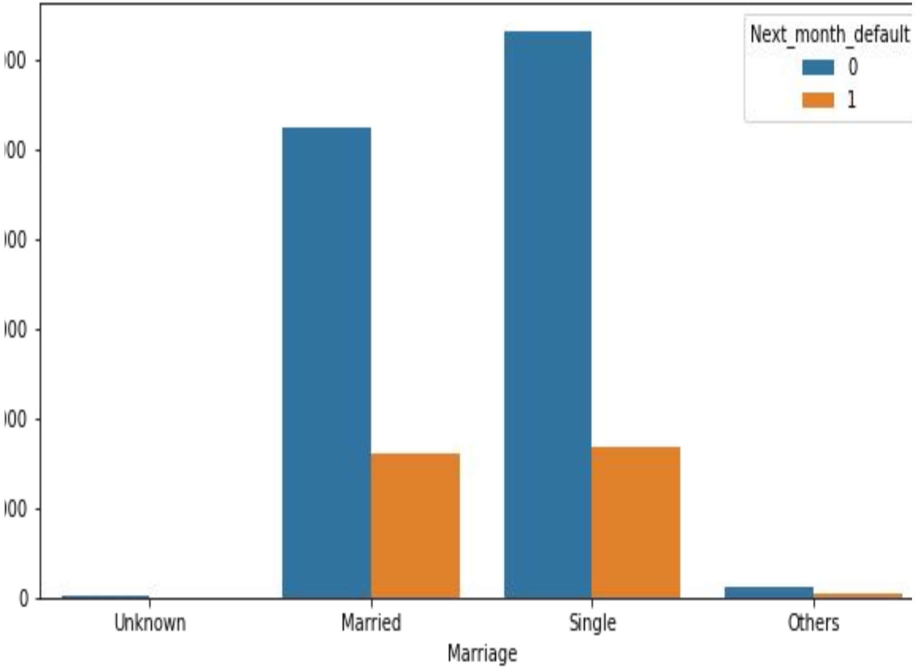
Count of Defaulters



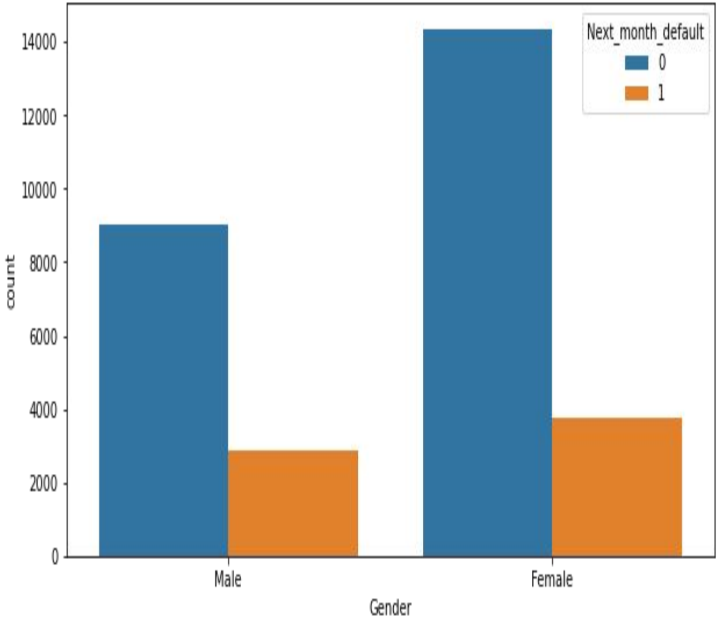
- This graph shows the number of defaulters.
- We can see that there are more of no defaulters and less of defaulters.
- This is because our dataset is imbalanced

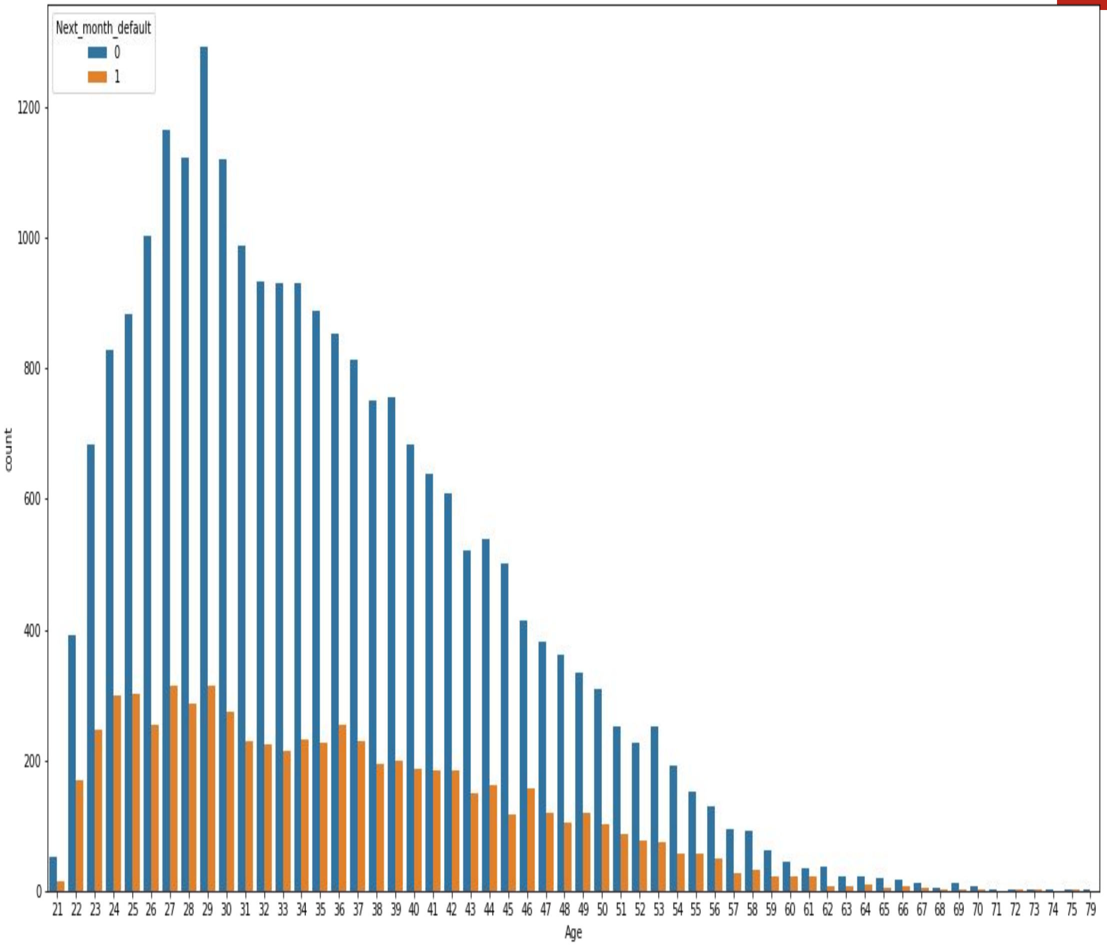
EDA on demographic factors.



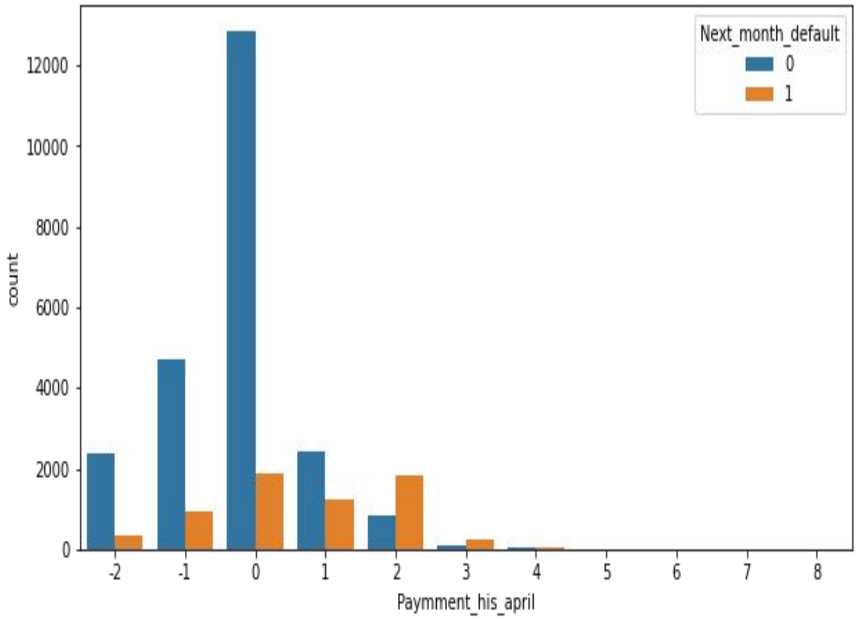


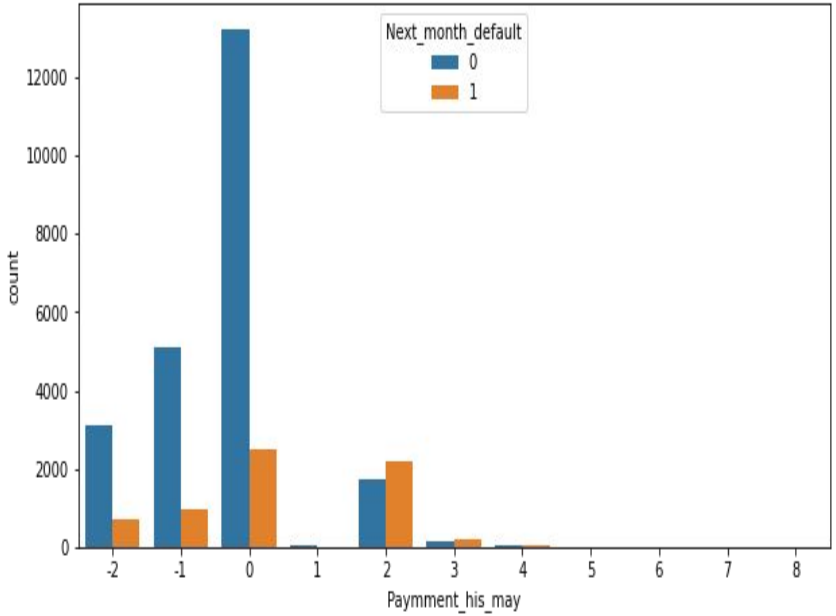
Gender and Age



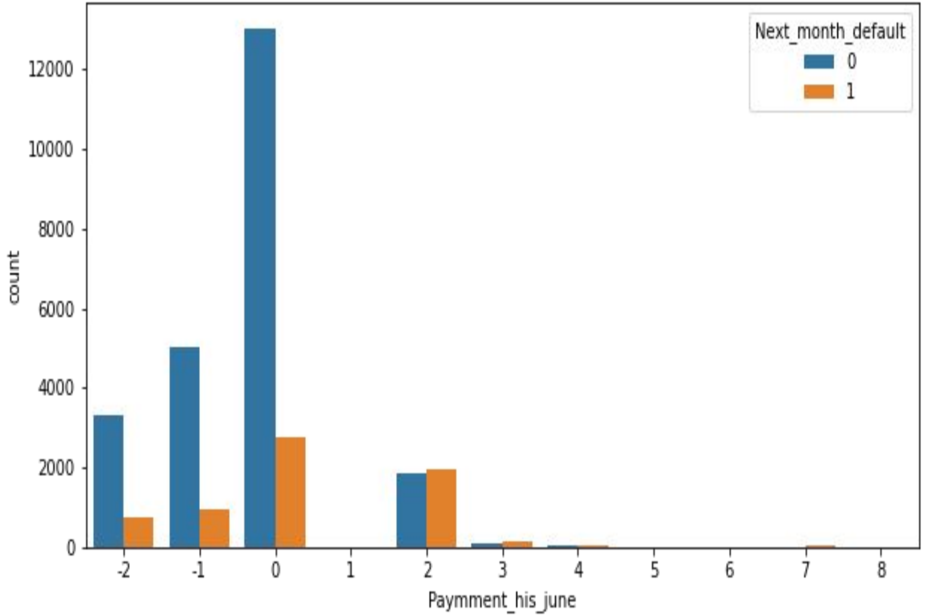


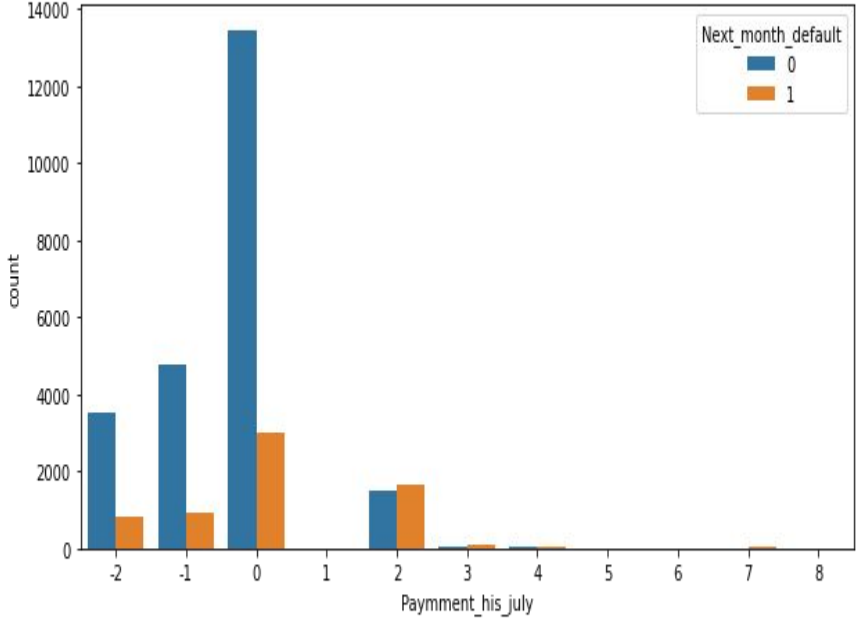
EDA on the Payment History April and May



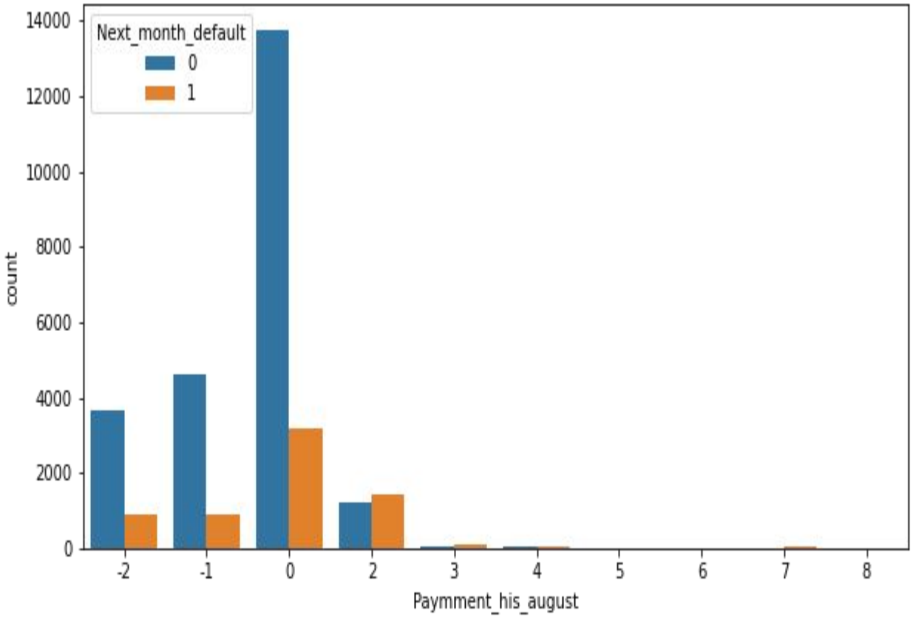


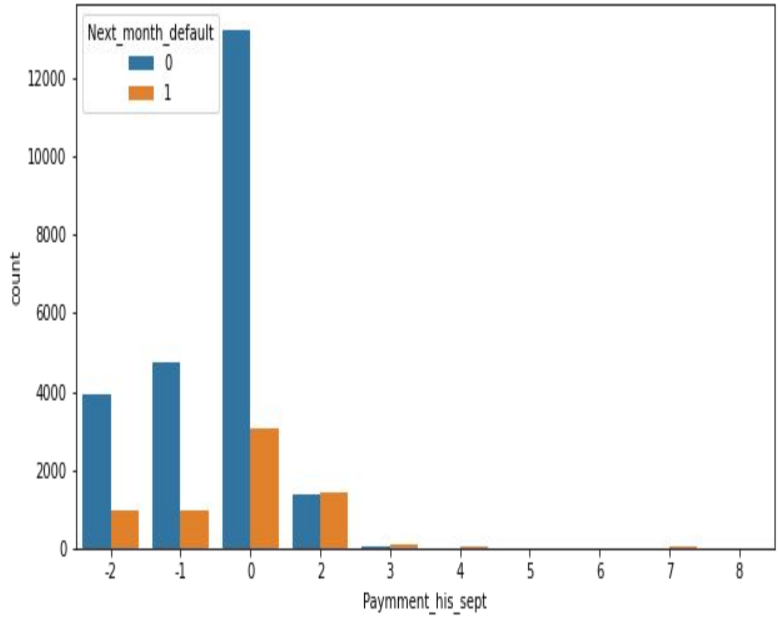
June and July





August and September





Dealing with the Imbalanced Data

From this exercise, we are trying to predict if the customer will make a default next month or not. But from the previous EDA we can understand that, Next_month_default feature seems to be imbalanced.

Lets understand what is Data Imbalance.

It simply means that the number of observations is not the same for all the classes in a classification dataset.This problem is very common to happen. It is happens when one class dominate another set of class during the time taking survey.

There are many methods to overcome this problem.

I've particularly used the below listed techniques to overcome the data imbalancing problem so to yield better accuracy from the predictive model.

Data Imbalancing Techniques.

1. Random Oversampling:

   Random oversampling randomly duplicate examples in the minority class. It works by selecting examples from minority classes with replacement and adds into the training set.

2. Random Undersampling :

   Random undersampling randomly delete examples in the majority class.It works by selecting examples from majority classes and deletes them from the training dataset.

Both these sampling techniques are known as naive techniques as it assumes nothing about the data.

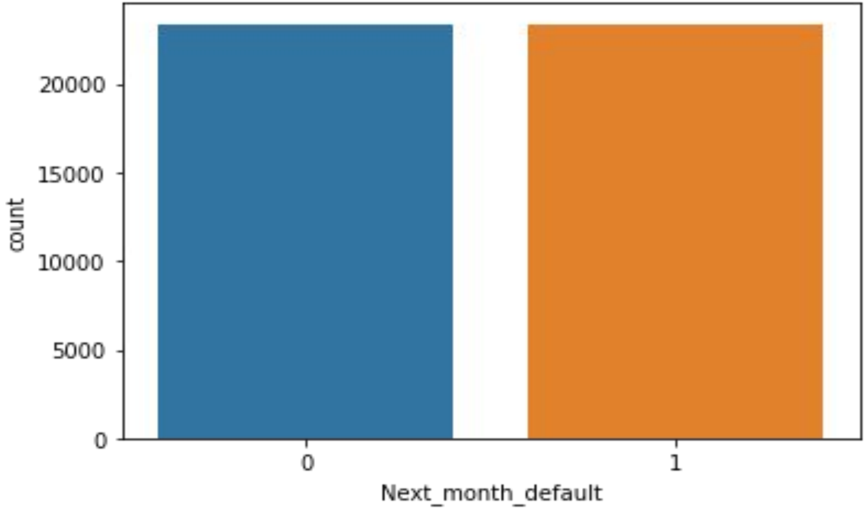
These methods are prone to overfit which is a huge disadvantage of random sampling techniques.

Data Imbalancing Techniques

Due to the above disadvantage, Synthetic Minority Oversampling Technique (SMOTE) works differently than over typical random sampling methods.

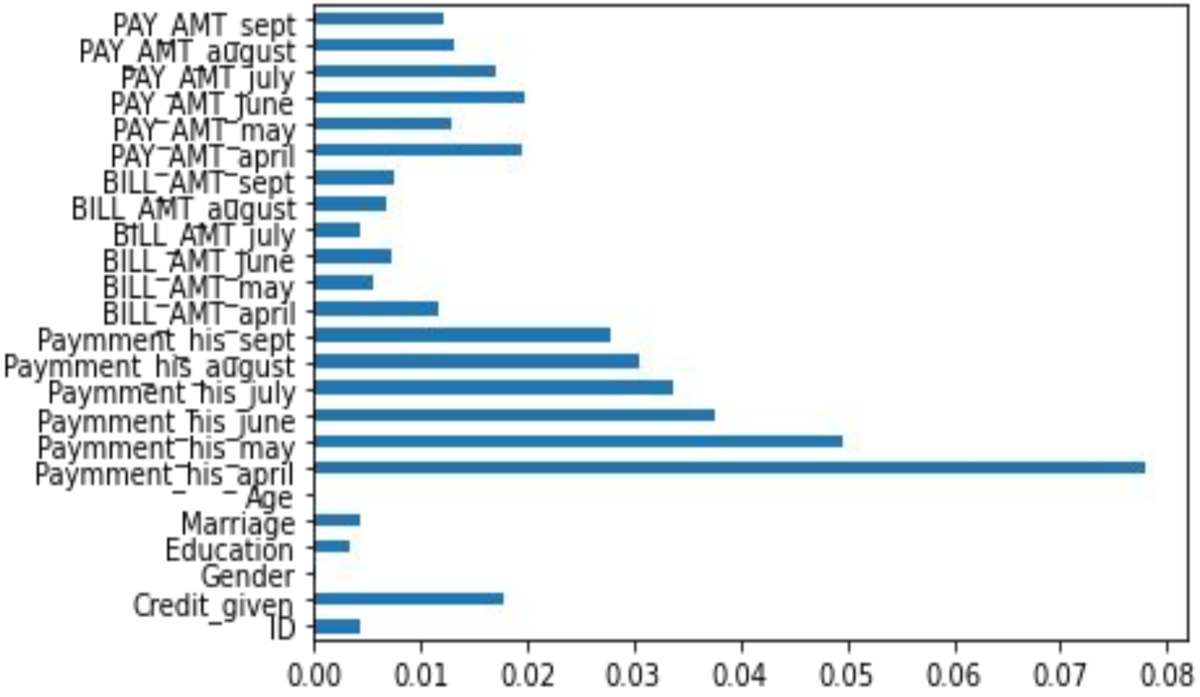
3. SMOTE :

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.



Feature Selection

Mutual into classif technique to select the best features



Feature Scores.

Feature	Score
Paymment_his_april	0.073723
Paymment_his_may	0.050523
Paymment_his_june	0.037313
Paymment_his_august	0.034425
Paymment_his_july	0.033068
Paymment_his_sept	0.029598
PAY_AMT_april	0.022693
PAY_AMT_june	0.017575
PAY_AMT_july	0.015232
PAY_AMT_may	0.015139
PAY_AMT_august	0.014097
Credit_given	0.012509
BILL_AMT_april	0.010464
PAY_AMT_sept	0.010322
BILL_AMT_may	0.006582

Independent and dependent Variable

Dependent Variable : Next_month_default is our dependent or the target variable.

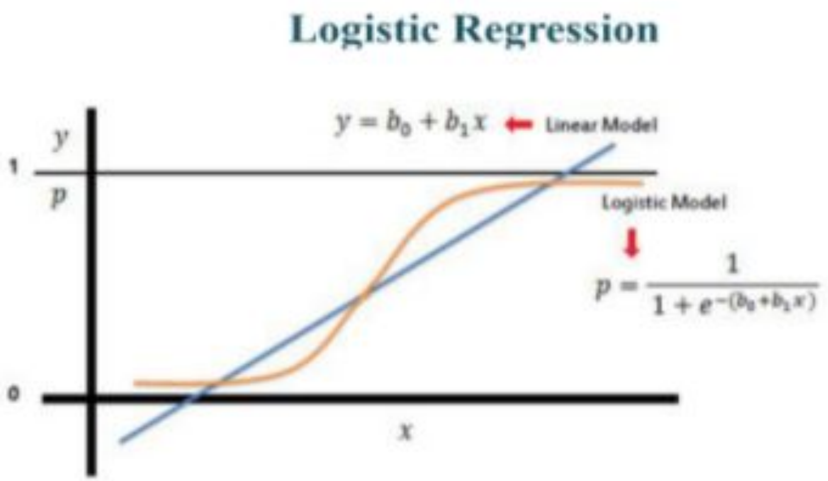
Independent Variable : From the above Feature selection we can see that ID, gender, Marriage,Education and age doesn't seem to have much importance.
So our independent variable will be all other variables in the dataset except the above mentioned.

Modelling

Algorithms used :

1. Logistic Regression.
2. Random Forest Classifier
3. K- Nearest Neighbour
4. Decision Tree classifier

Logistic Regression



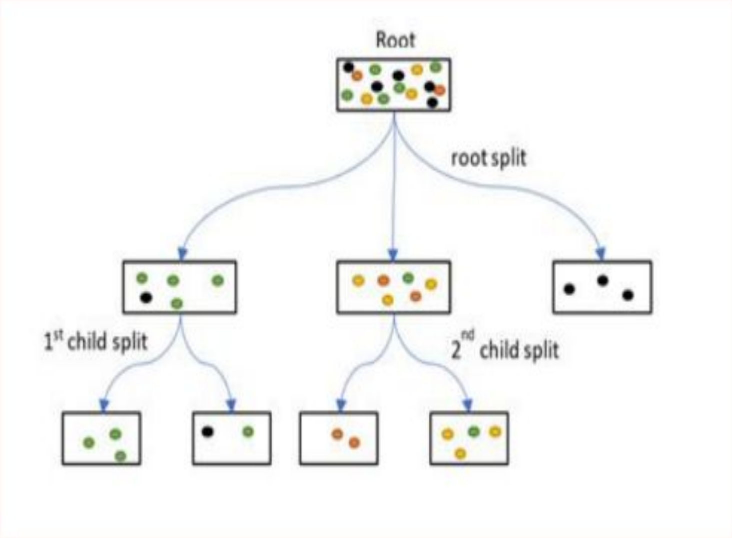
- Takes the linear combination and apply a sigmoid function
- It is one of simplest parametric classification model.

Logistic Regression and Logistic regression using Grid Search

	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	Logistic Regression	0.813333	0.735084	0.233865	0.354839	0.605081
1	Logistic Regression Tuned	0.813333	0.735084	0.233865	0.354839	0.605081

Decision Tree

Decision Tree classifier is a tree like structure algorithm, which basically is a graphical representation of getting all the possible solutions to a problem based on given condition.

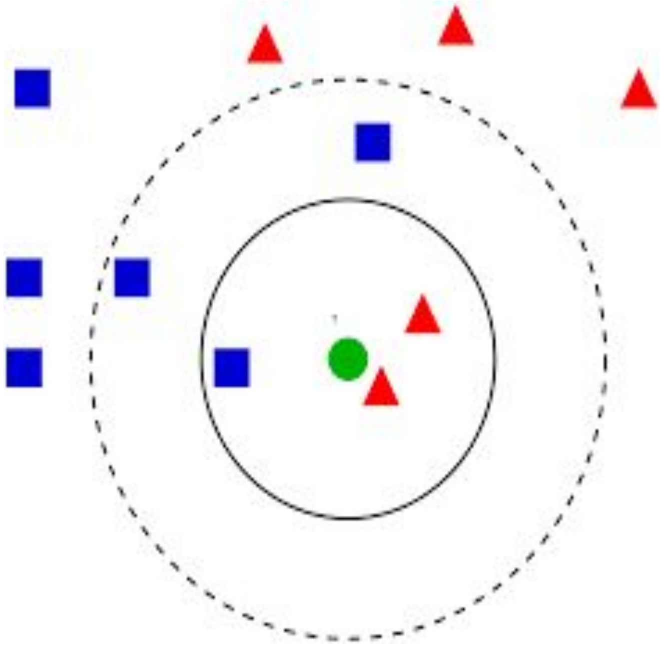


Decision Tree and Decision tree with grid search

	Model	Accuracy	Precision	Recall	F1 Score	ROC
2	Decison Tree	0.737333	0.405267	0.420653	0.412817	0.623523
3	Decison Tree Tuned	0.824667	0.690100	0.365224	0.477656	0.659550

K- Nearest Neighbour

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

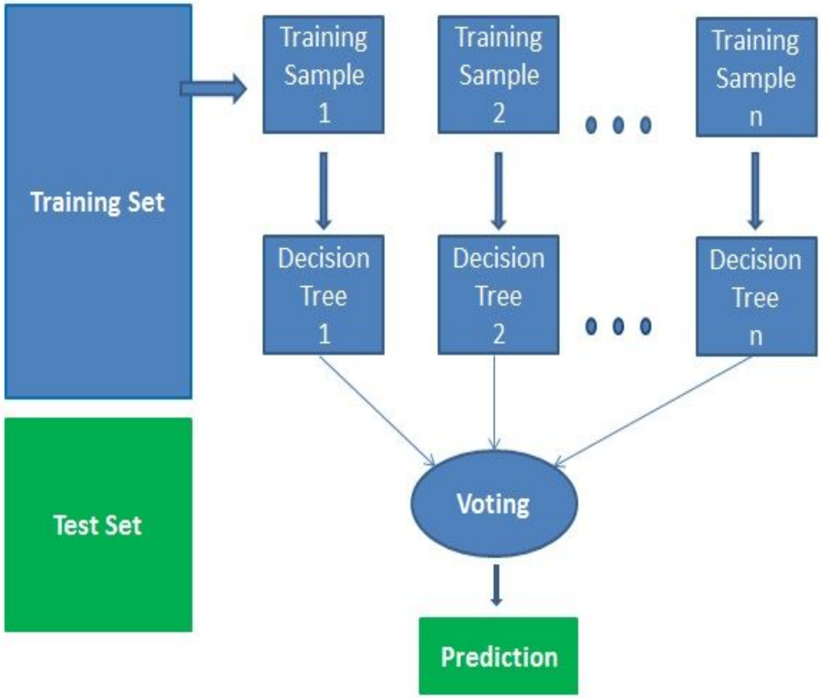


K- Nearest Neighbour

	Model	Accuracy	Precision	Recall	F1 Score	ROC
5	KNN	0.808000	0.601227	0.372058	0.459662	0.651329

Random Forest Classifier

Random Forest is a supervised learning algorithm, it creates a forest and makes it somehow random. The "forest“ it builds, is an ensemble of Decision Trees.



Random Forest Classifier

	Model	Accuracy	Precision	Recall	F1 Score	ROC
4	Random Forest Classifier	0.814333	0.632680	0.367502	0.464938	0.653749

Model Evaluation

	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	Logistic Regression	0.813333	0.735084	0.233865	0.354839	0.605081
1	Logistic Regression Tuned	0.813333	0.735084	0.233865	0.354839	0.605081
2	Decison Tree	0.737333	0.405267	0.420653	0.412817	0.623523
3	Decison Tree Tuned	0.824667	0.690100	0.365224	0.477656	0.659550
4	Random Forest Classifier	0.814333	0.632680	0.367502	0.464938	0.653749
5	KNN	0.808000	0.601227	0.372058	0.459662	0.651329

Conclusion

- The best **accuracy** is obtained for the **Decision Tree with grid search.**
- 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.

THANK YOU!!