**Credit Card Default Prediction**

**Harsh Durugkar**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

Credit Card Default Prediction is a Supervised Machine Learning Classification Project.The main purpose is to build model allows us to effectively combine static and dynamic features to provide superior predictive performance for financial data.

For this project i use credit card user dataset & This Dataset is from Taiwan.

In this dataset contain the all information of credit card users like User ID, Limit Bal, Include the information of users is Male or Female, Education, Marriage, Age, History of past payments from April to September, Amount of bill statement from April to September, Amount of Previous Payment from April to September, Default payment information.While working with this project we mainly use python programming languages. As well as for data wrangling we use numpy and pandas library and then for data visualization we use Matplotlib & Seaborn.

***Keywords: exploratory data analysis, supervised ml classification, random forest classification, Logistic Regression***

**1.Problem Statement**

1)As we know in today’s times, credit cards have huge risks behind the high returns of banks. The increasing number of credit card users is all about an increase in the number of credit card defaults and that’s why the result is amounts of bills & repayment information data have chances to create a risk.

2) The Credit card default prediction is based on the data of all credit card customers. The method which we use to predict and analyze credit card customer default behavior is a typical classification problem.

3) According to the Federal Reserve economic data, the default rate on credit loans across all commercial banks is at an all-time high for the past 66 months and it is likely to continue to climb throughout 2020.

4) That’s why, banks must have a risk prediction model and be able to classify the most relative characteristics that are indicative of people who have a higher probability of default on credit.

5) The main purpose is to build a model that allows us to effectively combine static and dynamic features to provide superior predictive performance for financial data.

**2. Introduction**

We Have a DataSet for our Credit Card Default Prediction Project this is Supervised Machine Learning Classification Project & Dataset we used is a Taiwan Dataset, for performing exploratory data analysis and we discuss about this dataset to cover some querry with analysing as well as for making model that allows us to effectively combine static and dynamic features to provide superior predictive performance for financial data.While Analysing the data we work with Credit Card Default Prediction Dataset, in this dataset include the information such as like User ID, Limit Bal, Include the information of users is Male or Female, Education, Marriage & other information with respected all 25 columns.

1. **Work of Flow**

In work of flow we first collect and understanding the data. While understanding the data we found there are 30000 number of rows and 25 number of columns. Then we find is any data miss or not in the dataset. But , there are no null value in our dataset. So, data is perfect for start the project .Then we start to performing exploratory data analysis(EDA).Next step is to start the preparation of data for model building and after prepation we start model selection and evaluation and last we move on our conclusion.

1. **Data Review**

There are total 30000 rows and 25 columns in our Dataset , here’s the list of Cloumns which is present in our Dataset :-

1. ID :- Contain Id Number of Credit Card Users.
2. Limit Bal :- Include the information of Limit Balance.
3. Sex :- Include the information of users is Male or Female.
4. Education :- Include the information of Education of Users.
5. Marriage :- Is user single or married.
6. Age :- Age information of users.
7. Pay-0 to Pay-6 :-History of past payments from April to September.
8. Bill-Amt1 to Bill-Amt6 :- Amount of bill statement from April to September.
9. Pay-Amt1 to Pay-Amt6 :- Amount of Previous Payment from April to September. 10) Default Payment Next Month : - Default payment information.
10. **Points We Cover With this Project**

1.Visualize the data of Defaulters vs Non-Defaulters.

2. Visualize the data of Male vs Female for Credit.

3. Visualize the data of Education of Credit Card Holders.

4. Visualize the data From Marriage Column.

5. Visualize the data of Number of People By Age.

6. Visualize the Data Distribution from LIMIT\_BAL Column.

7. Visualize the data of default payment next month with limit Balance.

8. Visualize the data of default payment next month with Sex Column.

9. Visualize the data of default payment next month with Education Column.

10. Visualize the data of default payment next month with Marriage Column.

11. Visualize the data of default payment next month with Age Column.

12. Corelation between dependent and independent variable.

1. **Performing EDA**

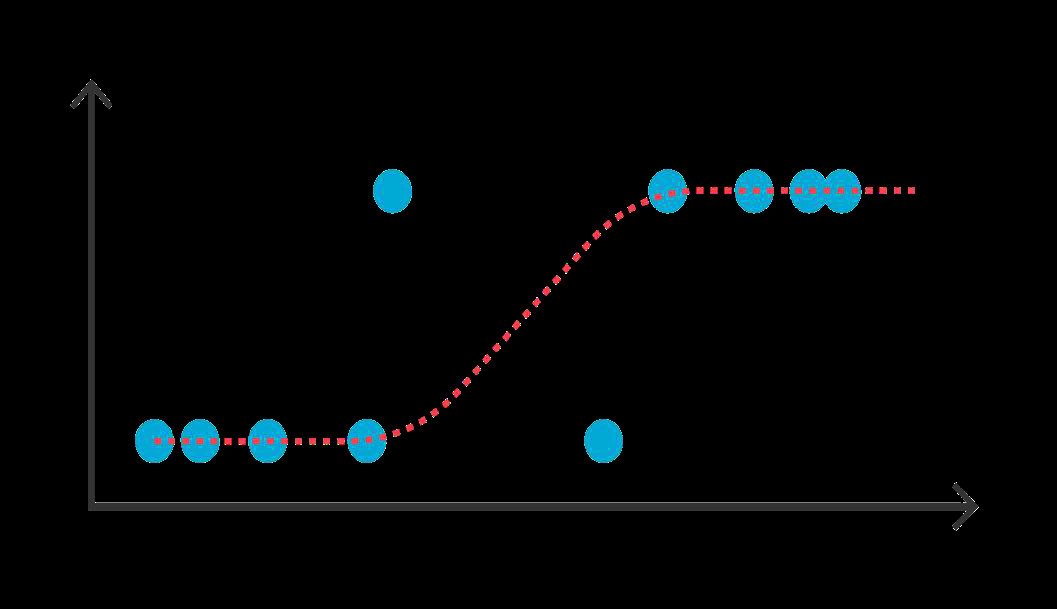
So we move on our Next part of this project which is Exploratory data analysis (EDA).

Exploratory Data Analysis is an process to analyze the data using some visual techniques. It is used to discover trends, patterns, as well as to check assumptions with the help of statistical knowledge and graphical representations. It means trying to understand the given data much better, so that we can make it more sense out of it. It is also used to produce a value distribution and identify missing values, and outliers.

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing tasked in Python uses data visualization to draw meaningful patterns and insights.

1. **Performing Model Building :**
2. Logistic Regression
3. Random Forest Classifiers
4. Support Vector Classifier
5. XGBoost Classifiers
6. **Logistic Regression :-** Logistic Regression is similar to Linear Regression. It is also used to find the relationship between the Dependent variable and one/more Independent Variable, also it’s used to make predictions for a categorical variable as well as used to handle the classification problems.

We perform Logistic Regression with respect to parameter :- {'C': 0.01, 'penalty': 'l2'}

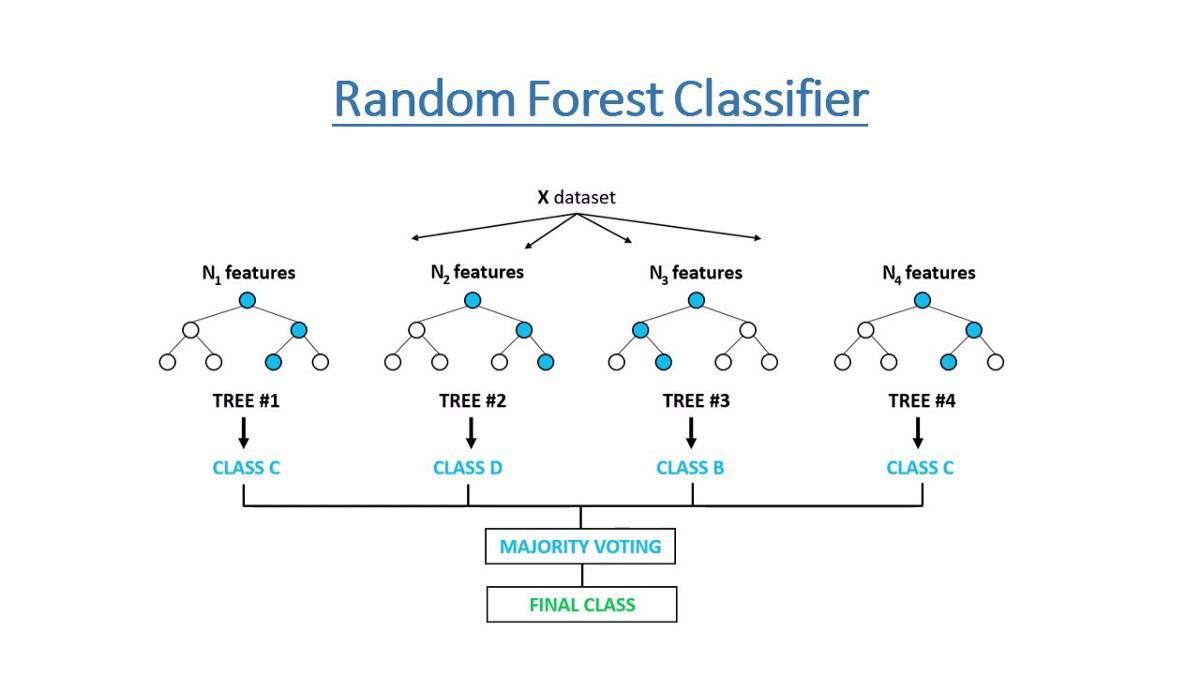


Library used for Logistic Regression :-

from sklearn.linear\_model import

LogisticRegression

1. **Random Forest Classifiers :-** Random Forest Classifier is a technique that makes an aggregated prediction using a group of decision trees trained Using the bootstrap method with extra randomness, while growing trees with the help of searching for the best features among a randomly selected feature subset.

We perform Random Forest Classifiers with respect to parameter :- {‘max\_depth’: 30, ‘n\_estimators’:200}

Library I’m used for Random Forest Classifiers :- from sklearn.ensemble import RandomForestClassifier

1. **Support Vector Classifiers :-** Support vector classifiers are a set of supervised learning methods used for classification, regression and outlier detection. The big advantage of support vector machines is that Effective in high dimensional spaces as well as it’s still effective in cases where the number of dimensions is greater than the number of samples.

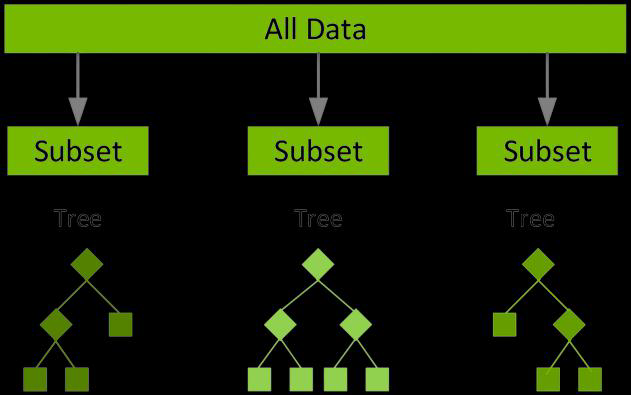
We perform Support Vector Classifiers with respect to parameter :- {‘c’: 10, ‘kernal’:

‘rbf’}

Library I’m used for Support Vector Classifiers :- from sklearn.model\_selection import GridSearchCV

1. **XGBoost Classifiers :-** XGBoost, which also stands for Extreme Gradient Boosting, is a scalable & distributed gradient-boosted decision tree (GBDT) machine learning library. It is the leading machine learning library for regression, classification, and ranking problems as well as it is provides parallel tree boosting.

We perform XGBoost Classifiers with respect to parameter :- {‘max\_depth’: 15, ‘min\_child\_weight’: 8}



Library I’m used for XGBoost Classifiers :- import xgboost as xgb from xgboost import XGBClassifier

1. **Conclusion :**

1) We observe 78% of people are Non-defaulters and the remaining 22% are Defaulters.

2) Male credit holder is less Than Female Credit Card Holder and if we compare male/female with defaulters list we observe that, in defaulters list male credit holder is Higher than Female Credit Holder.

3) Highest Number of credit holders are university students then 2nd Highest are Graduate Students then 3rd Highest from High school Students & Remaining from Others.

4) Highest Number of credit holders are Single, then 2nd Highest are Married & remaining are from Others category. As well as we observe married people have less number of defaulters with the comparison of other’s marriage person category list.

1. The Maximum amount of given credit in NT dollars is 50,000 followed by 30,000 and 20,000.

6) We observe Most of credit card holders' ages start from 24-32 Years and people's age above 61 year , they use credit cards very rarely.

7) We find the relationship between age and defaulter’s & we can say that people who are 60 years or older , that may be they don’t use their credit card frequently.

8) In both cases they have a negative impact on the bank, since false positives leads to unsatisfied customers and false negative leads to financial loss.

9)XGBoost Classifier having Recall, F1\_score, and ROC Score values equals 82%, 77%, and 86% and Random forest Classifier having Recall,F1score and ROC Score values equals 81%, 75%, and 84%.

1. XGBoost Classifier and Decision Tree Classifier are giving us the best Recall, F1\_score, and ROC Score among other algorithms.

11) We observe XGBoost classifier and decision tree classifier are the best to predict whether the credit card user is defaulter or non-defaulter.

12) Random Forest is Higher Precision than Logistic Regression. That's why Random forest is better than logistic regression and it’s suitable for our machine learning model.

**References-**

1. W3Schools
2. GeeksforGeeks
3. Analytics Vidhya
4. H. Kim, H. Cho and D. Ryu, "An empirical study on credit card loan delinquency", Econ. Syst., vol. 42, no. 3, pp. 437-449, Sep. 2018.