

Predicting whether a customer will default on his/her credit card

Meenakshi Singh

Adityasingh Thakur

Tushar Wagh

Abstract: Understanding the history of clients will act as a valuable screening method for banks by providing information that can categorize clients as defaulters on a loan. Customer credit rating is a grade process where the consumer is categorized by the grade. Credit scoring model used to ascertain credit risk from new and existing customer. Credit rating is an assessment used to measure the creditworthiness of the customer. For the huge customers related dataset, we can use various classification techniques used in the field of data mining. The main idea is by analysing the customer data and by combining machine-learning algorithm to identify the default credit card user. Default is a keyword, used for predicting the customer who can't repay the amount on time. Predicting future credit default accounts in advance is highly tedious task. Modern statistical techniques are usually unable to manage huge data. The proposed work focuses mainly on supervised learning classification technique.

Keywords: Credit card, Logistic Regression, Decision Tree, Xgboost, Support Vector Classifier

I. 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. We can use the K-S chart to evaluate which customers will default on their credit card payments

II. Objective

- Objective of our project is to predict which customer might default in upcoming months. Before going any further let's have a quick look on definition of what actually meant by Credit Card Default.
- We are all aware what is credit card. It is type of payment card in which charges are made against a line of credit instead of the account holder's cash deposits. When someone uses a credit card to make a purchase, that person's account accrues a balance that must be paid off each month.
- Credit card default happens when you have become severely delinquent (usually a young person who regularly performs illegal or immoral acts) on your credit card payments. Missing credit card payments once or twice does not count as a default. A payment default occurs when you fail to pay the Minimum Amount Due on the credit card for a few consecutive months.

III. Introduction

When a customer applies for and receive a credit card, it becomes a huge responsibility for customer as well as the credit card issuing company. The credit card company evaluates the customer's credit worthiness and gives him/her a line of credit that they feel a customer can be responsible for. While most people will use their card to make purchases and then diligently make payments on what they charge, there are some people who, for one reason or another do not keep up on their payments and eventually go into credit card default.

Credit card default is the term used to describe what happens when a credit card user makes purchase by charging them to their credit card and they do not pay their bill. It can occur when one payment is more than 30 days past due, which may rise your interest rate. Most of the time, the term default is used informally when the credit card payment is more than 60 days past due. A default has a negative impact on the credit report and most likely lead to higher interest rates on future borrowings.

As more and more consumers rely on the credit card to pay their everyday purchases in online and physical retail store, the amount of issued credit cards and the overwhelming amount of credit card debt by the cardholders have rapidly increased. Therefore, most financial institutions have to deal with the issues of credit card default in addition to the credit card fraud such as credit card dump. Both the credit card verification applied to the cardholders and the default risk management after card issued are crucial to the future success of most financial institutions. Under those circumstances, to reduce the delinquent payment of credit card holders, and perform the effective management of credit card risk will help financial and banking institutions to achieve competence advantages.

IV. Data Description

The suggested system uses the original UCI repository report. There are 25 factors and 30,000 documents for customers. This dataset contains information on the credit card clients, regular charges, demographic factors, credit records, payment. 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, 5=unknown, 6=unknown)
- ☐ **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 (1=yes, 0=no)

V. Exploratory Data Analysis

This dataset contains information on default payments, demographic factors, credit limit, history of payments, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It includes 30,000 rows and 25 columns, and there is no credit score or credit history information.

Overall, the dataset is very clean, but there are several undocumented column values. As a result, most of the data wrangling effort was spent on searching information and interpreting the columns.

More details about the data cleaning can be found in this Colab Notebook. The purpose of exploratory data analysis is to identify the variables that impact payment default likelihood and the correlations between them. We use graphical and statistical data exploratory analysis tools to check every categorical variable. Each starts with a visualization and is followed by a statistical test to verify the findings.

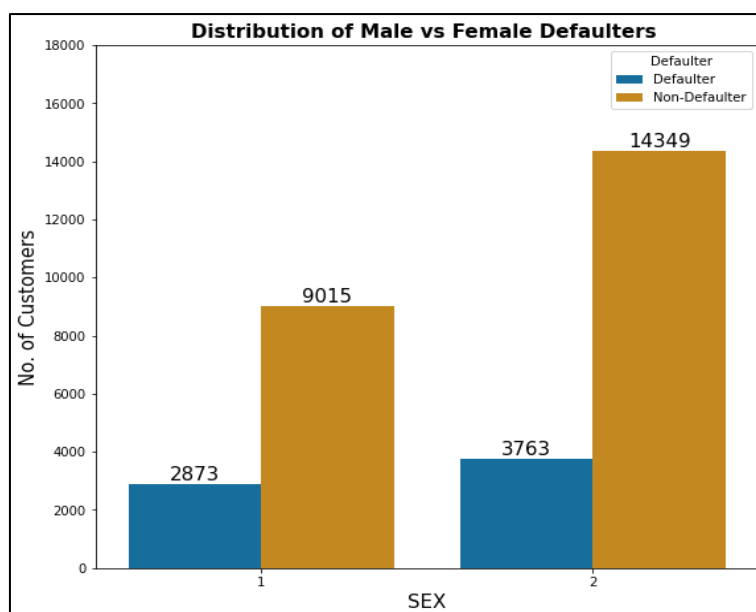
The main findings from exploratory analysis are as following:

- Females have more delayed payment than males in this dataset.
- Customers with higher education have less default payments and higher credit limits.
- Customers aged between 30-50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates. However, the delayed rate drops slightly again in customers older than 70.
- There appears to be no correlation between default payment and marital status.
- Customers being inactive doesn't mean they have no default risk. We found 317 out of 870 inactive customers who had no consumption in 6 months then defaulted next month.

1. Gender Variable

Females have more delayed payments than males in this dataset.

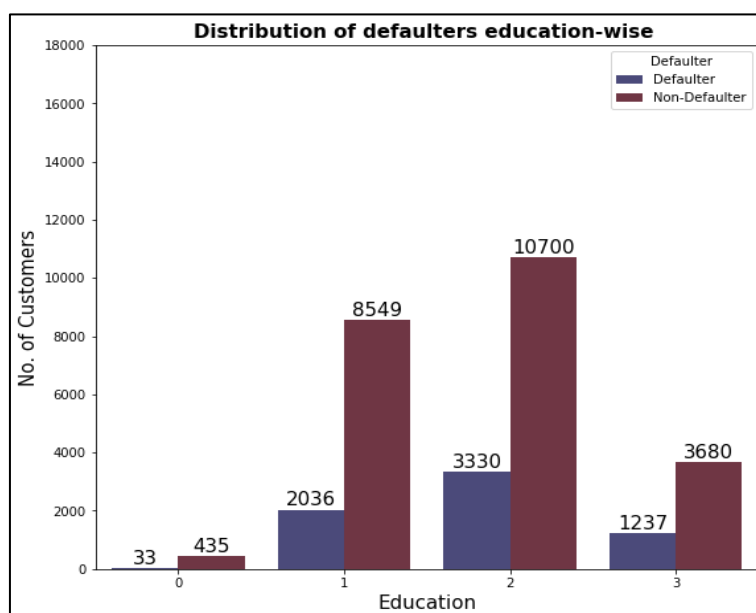
Which gender group tends to have more delayed payment? Since there are more females than males in the dataset, we use percentage of default within each sex group. Figure shows 30% females have default payment while only 26% males have default payment. The difference is not significant.



2. Education Variable

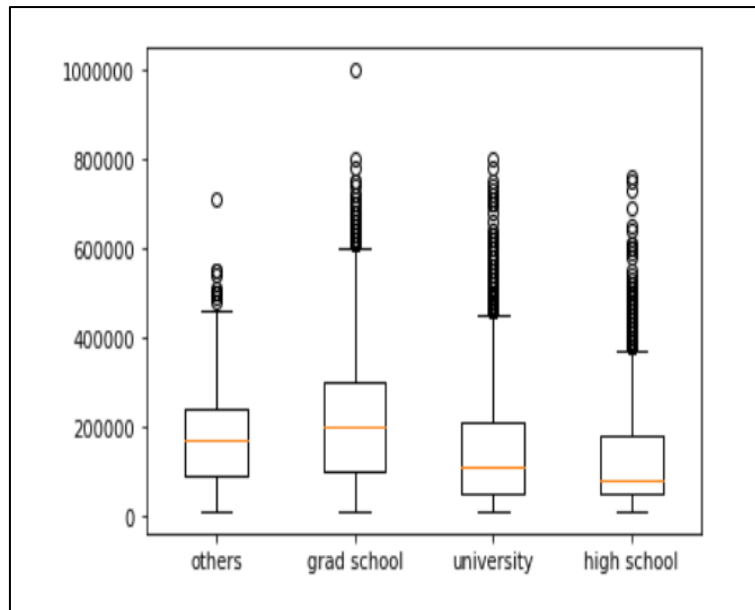
Customers with higher education have less delayed payment.

Figure indicates customers with lower education levels default more. Customers with high school and university educational level have higher default percentages than customers with grad school education. Notice there is an education group “others” which appears to have the least default payment, but this group only has 468 (or 1.56%) customers, and we don’t know what consists of this group.



Customers with a high education level get higher credit limits.

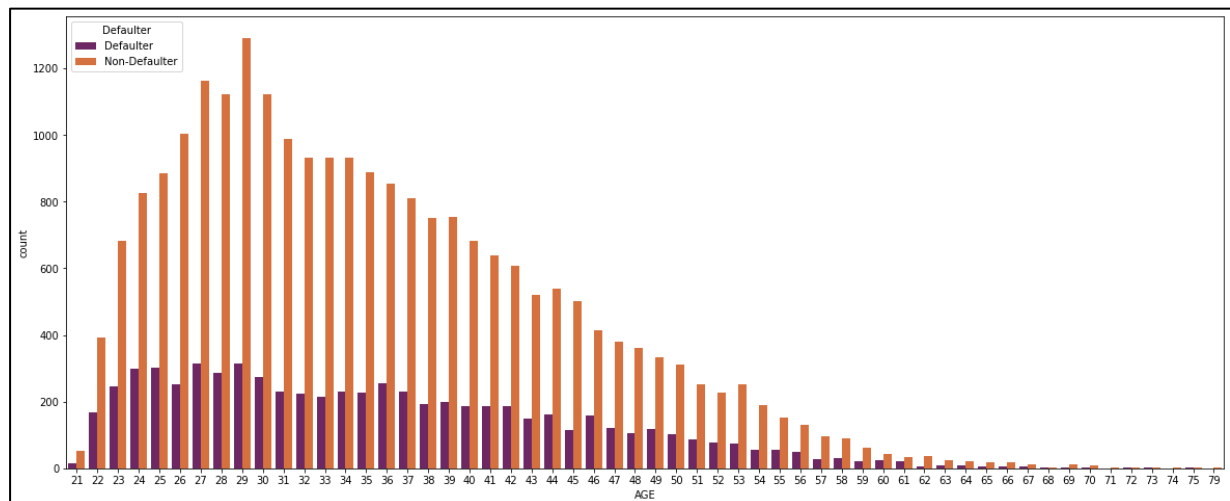
From the boxplot in figure 3, it is obvious that customers with grad school education have the highest 25% percentile, highest median, highest 75th percentile and highest maximum numbers, which suggests customers with higher education levels do get higher credit limits.



3.Age Variable

Middle-aged customers have the lowest default rate.

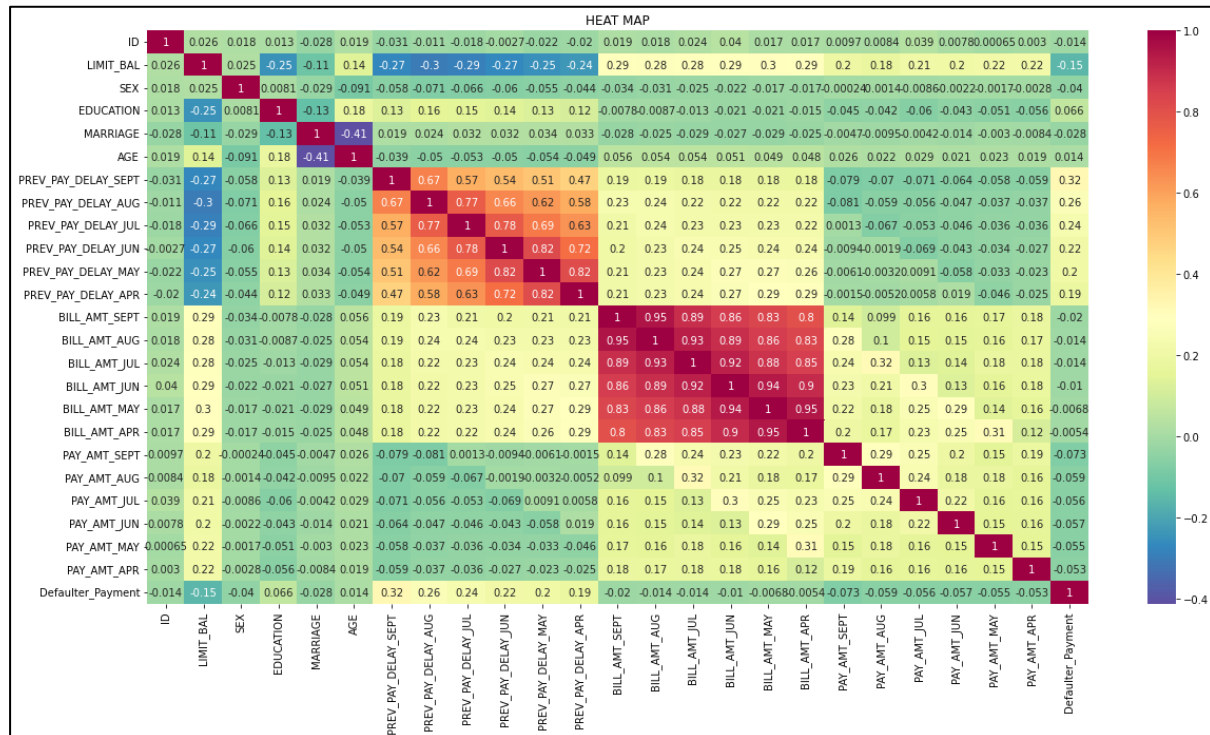
The count plot shows the default probability increases for customers younger than 30 and older than 70. Customers aged between 30 and 50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates. This aligns with social reality that customers aged 30-50 typically have the strongest earning power. We also notice the delayed rate drops slightly again in customers older than 70. This is understandable because elder customers' consumption tends to decrease.



4. Correlation Analysis: -

Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship. A correlation plot typically contains a number of numerical variables, with each variable represented by a column. The rows represent the relationship between each pair of variables. The values in the cells indicate the strength of the

relationship, with positive values indicating a positive relationship and negative values indicating a negative relationship. Correlation heatmaps can be used to find potential relationships between variables and to understand the strength of these relationships. In addition, correlation plots can be used to identify outliers and to detect linear and nonlinear relationships. The color-coding of the cells makes it easy to identify relationships between variables at a glance. Correlation heatmaps can be used to find both linear and nonlinear relationships between variables.



From the above heat map, we conclude that,

- There is a negative correlation in age, limit_bal and marriage
- There is high positive correlation in delay payment and bill column

5. Multicollinearity

- The **variance inflation factor (VIF)** identifies correlation between independent variables and the strength of that correlation. Using Variance Inflation Factor- VIF- we can determine if two independent variables are collinear with each other.
- A **variance inflation factor (VIF)** detects multicollinearity and it affects the performance of regression and classification model's analysis. Multicollinearity is when there's correlation between predictors (i.e., independent variables) in a model; it's presence can adversely affect your results.

6. Data Preparation

Data preparation includes feature engineering and feature selection. In feature engineering we converted categorical features such as Marriage, Education and Sex and into binary form. Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.

- One hot encoding is a process by which categorical variables are converted into a numerical variable that could be provided to ML algorithms to do a better job in prediction.
- Here we perform one hot encoding on 'EDUCATION', 'MARRIAGE', and 'SEX'.
- **get dummies:** - It converts categorical data into dummy or indicator variables.

VI. Model Fitting

For modeling we tried various classification models such as-

- 1) Logistic Regression
- 2) K-Neighbors Classifier
- 3) Support Vector Classifier
- 4) DecisionTree Classifier
- 5) RandomForest Classifier
- 6) XG Boosting

Usually, larger part of data is needed to teach the models and so 80% of the final data is utilized for model training and the remaining 20% of the data is used for testing purpose. We started with building a simple logistic regression model and evaluated its performance using evaluation metrics. Next, we build the K-Nearest Neighbors model on the training set. Using GridSearchCV we got the best performing n-neighbors value as 5. We built the Support Vector Classifier and evaluated the model using various evaluation metrics.

Using GridSearchCV again we built a Decision tree classifier to find the best hyperparameters. We built Random Forest and XGboosting algorithms using GridSearchCV method to tune the hyperparameters of algorithms as well as to get the best hyperparameters. Tuning the hyperparameters is necessary to get better model performance and to prevent overfitting in case of tree-based models like Decision Tree, Random Forest, XGBoosting. We plotted the confusion matrix and roc curve to evaluate models.

We also plotted variable importance plots for tree-based models to determine the features that were most important while model building.

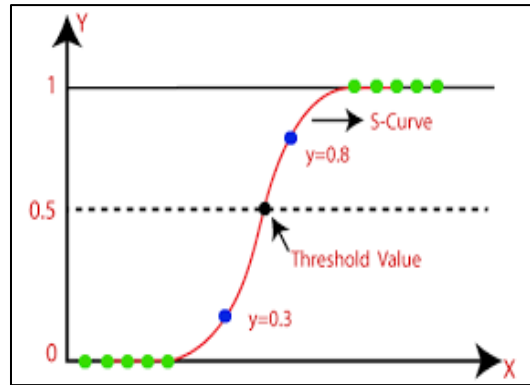
VII. Classification Analysis

Classification is a technique for determining which class the dependent belongs to based on one or more independent variables.

1. Algorithms-

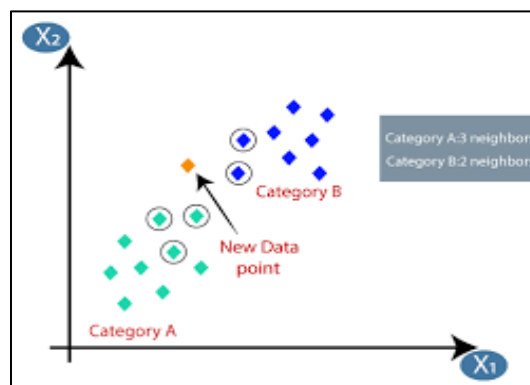
1. Logistic Regression:

Logistic regression is kind of like linear regression, but is used when the dependent variable is not continuous but categorical (e.g., a "yes/no" response). It's called regression but performs classification based on the regression and it classifies the dependent variable into either of the classes. Logistic Regression is used when the dependent variable(target) is categorical. For example, to predict whether an email is spam (1) or not (0).



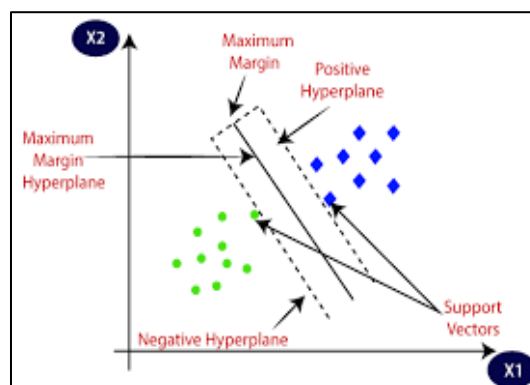
2. K-Neighbors Classifier:

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. K-NN is a non-parametric, lazy learning algorithm. It classifies new cases based on a similarity measure (i.e., distance functions). K-NN works well with a small number of input variables, but struggles when the number of inputs is very large.



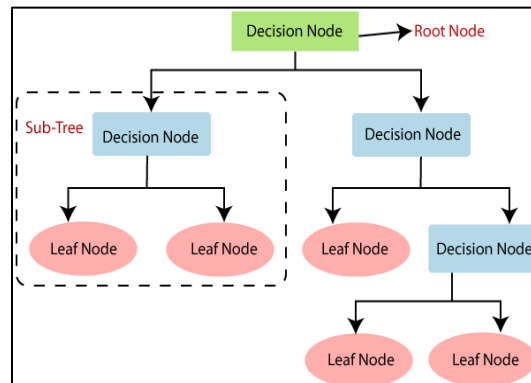
3.Support Vector Classifier:

Support vector is used for both regression and classification. It is based on the concept of decision planes that define decision boundaries. A decision plane (hyperplane) is one that separates between a set of objects having different class memberships. It performs classification by finding the hyperplane that maximizes the margin between two classes with the help of support vectors.



4. Decision Tree Classifier:

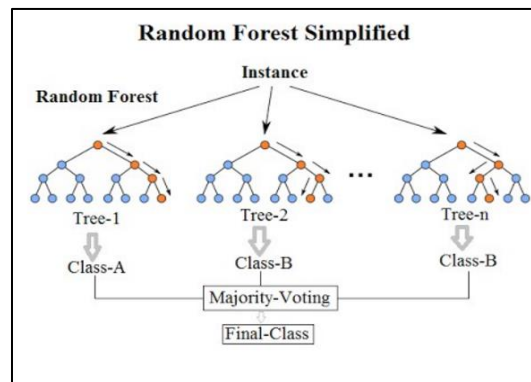
Decision tree build classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. It follows the Iterative Dichotomised 3(ID3) algorithm structure for determining the split.



5. Random Forest Classifier: -

Random forest classifier is an ensemble algorithm based on bagging i.e bootstrap aggregation. Ensemble methods combine more than one algorithm of the same or different kind for classifying objects.

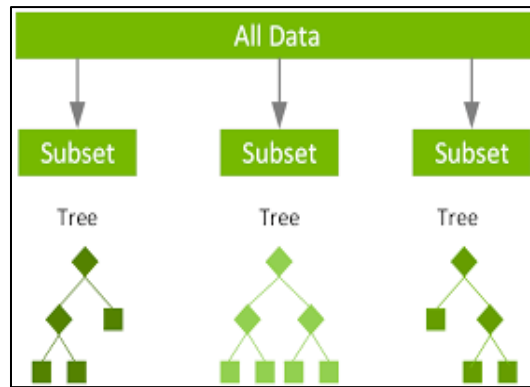
Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction.



6. XGboosting

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. It's vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting.

Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset's features.



2. Hyperparameter tuning

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Searching for the best hyperparameter can be tedious, hence search algorithms like grid search and random search are used.

RandomSearch CV has the same purpose as GridSearchCV as they both are designed to find the best parameters to improve the model. However, in random search not all parameters are tested. Rather, the search is randomized and all the other parameters are held constant while the parameters we are testing are varied. Practically, the implementation of RandomSearch CV is very similar to that of the GridSearchCV. The main difference between the practical implementation of the two methods is that we can use `n_iter` to specify how many parameters values we want to sample and test.

3. Model Performance-

Evaluation Metrics:

There are different metrics for supervised algorithms (classification and regression) and unsupervised algorithms. Let's start exploring various Evaluation metrics we used.

- 1) Accuracy-** The accuracy of a classifier is calculated as the ratio of the total number of correctly predicted samples by the total number of samples. Accuracy metric should not be used when the data set is imbalanced.

$$Accuracy = \frac{\text{Total number of correctly predicted samples}}{\text{Total number of samples}}$$

- 2) Confusion Matrix-** A confusion matrix is an N dimensional square matrix, where N represents the total number of target classes or categories. Confusion matrix can be used to evaluate a classifier whenever the data set is imbalanced.

	Predicted: NO	Predicted: YES
Actual: NO	True Negative (TN)	False Positive (FP)
Actual: YES	False Negative (FN)	True Positive (TP)

There are four important terms in a confusion matrix

True Positives (TP): These are the cases where the predicted “Yes” actually belonged to class “Yes”.

True Negatives (TN): These are the cases where the predicted “No” actually belonged to class “No”.

False Positives (FP): These are the cases where the predicted “Yes” actually belonged to class “No”.

False Negatives (FN): These are the cases where the predicted “No” actually belonged to class “Yes”.

3) Precision- Precision is the ratio of true positives (TP) by the sum of true positives (TP) and false positives (FP).

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

4) Recall- Recall is the ratio of true positives (TP) by the sum of true positives (TP) and false negatives (FN).

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

5) F1 Score- F1 score should be used when both precision and recall are important for the use case. F1 score is the harmonic mean of precision and recall. It lies between [0,1].

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

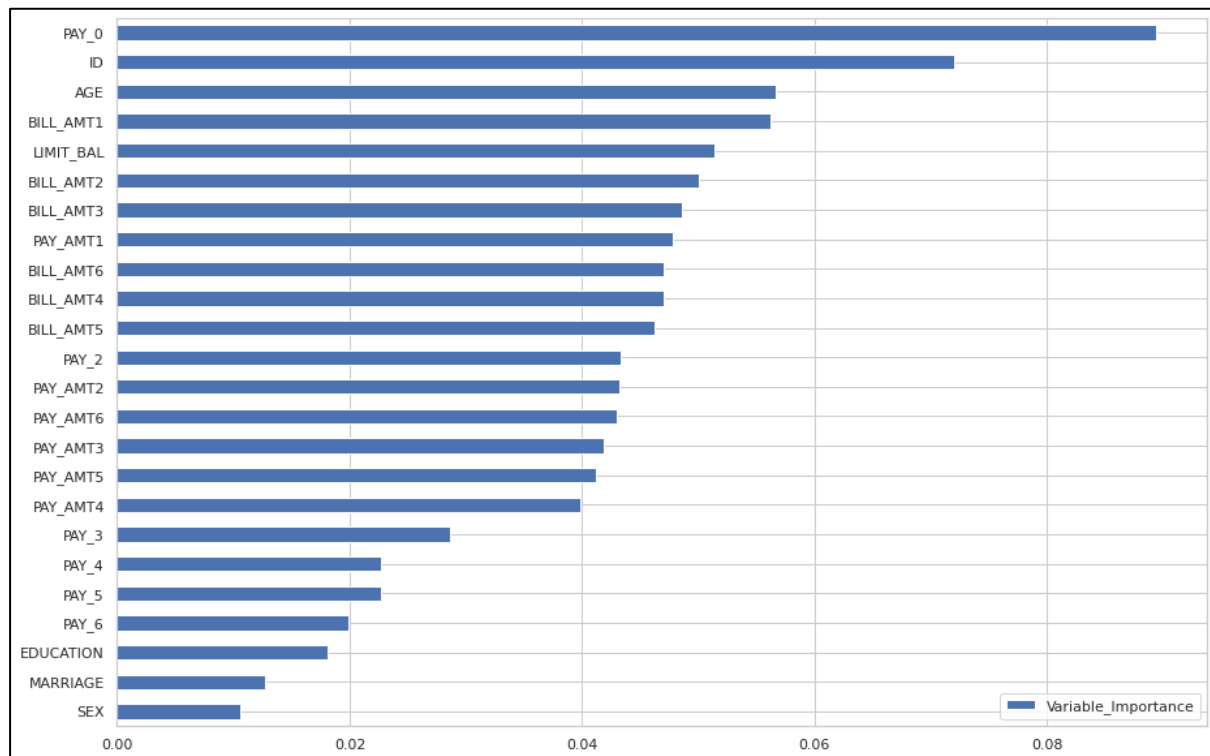
6) AUC-ROC Curve-

AUC-ROC Curve is a performance metric that is used to measure the performance for the classification model at different threshold values. ROC is Receiver Operating Characteristic Curve and AUC is Area Under Curve. The higher the value of AUC (Area under the curve), the better is our classifier in predicting the classes. AUC-ROC is mostly used in binary classification problems. The ROC curve is plotted between True Positive Rate (TPR) and False Positive Rate (FPR) i.e., TPR on the y-axis and FPR on the x-axis. AUC is the area under the ROC curve. An excellent classifier has an AUC value near 1, whereas a poor-performing classifier has an AOC value near 0. A classifier with an AOC score of 0.5 doesn't have any class separation capacity.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Feature Importance:

Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Feature importance scores play an important role in a predictive modeling project, including providing insight into the data, insight into the model, and the basis for dimensionality reduction and feature selection that can improve the efficiency and effectiveness of a predictive model on the problem.



- **Pay_0 has most feature importance in our dataset.**

Result: -

After training the models with their best hyperparameters, the performance of each of the classification models were evaluated using the testing set with evaluation metrics. The model which provides the highest Accuracy is the best one. Since our dataset was highly imbalanced so we will not consider accuracy to evaluate models.

Model_Name	Accuracy_Score	Precision_Score	Recall_Score	F1_Score	RUC_AUC_Score
Logistics regression	0.6863440587788193	0.44777986241400874	0.6685340802987861	0.5363295880149813	0.6807552181744277
Random Forest	0.8816822903470991	0.7710951526032316	0.8020541549953315	0.7862700228832953	0.8566946713780552
K-Neighbor Classifier	0.793767418292374	0.5762611275964392	0.9066293183940243	0.704644412191582	0.8291839220621023
Support Vector Machine	0.7745122878135292	0.5708692247454973	0.680672268907563	0.6209540034071551	0.7450649244398733
XG Boosting	0.7719787180136812	0.5796831314072693	0.580765639589169	0.580223880597015	0.7119753093634302
Decision Tree Classifier	0.7342285279959463	0.5088	0.5938375350140056	0.5480396380870315	0.6901732876739012

- From the above table we get to know that the **Random Forest Classifier** performs best with accuracy score of 88.16% as compared to all other models implemented, Because of high precision and low recall which results in higher value of F1 score. As you can also see that RFC Occupies higher area as 85.66% as compared to other models
- **K- Neighbor Classifier** Performs the second best with accuracy score of 79.37% after random forest classifier.
- **Support Vector Classifier and XG Boosting** give almost same accuracy score of 77 %.
- **Logistic Regression** performs worst as compared to all other models with a least accuracy score of 68% because of low precision and high recall value

VIII. Conclusion

- Dataset consists of 30000 observations and 25 features in which there are 3 categorical features 'SEX' , ' EDUCATION' ' MARITAL STATUS' and the rest are numerical features. There are no null values nor any duplicate values.
- By plotting the distribution plot categorical feature 'EDUCATION' and 'MARRIAGE' we get to see that there were more university customers and more single customers in our dataset.
- There is a positive correlation between PRE_PAY_DELAY_X to PAY_AMT_X.
- There is positive high correlation between bill amount and total month of delay.
- Customers with high credit limits tend to have higher 'No-Default' rate.
- Using the count plot, we get to know that there are less non-defaulters as compared to defaulters in our dataset.
- For AGE = 29 there are more numbers of non-defaulters and for AGE = 21 and after 59 there are least defaulters in our dataset.
- We have used a heat map plot to get to know the correlation between Dependent and Independent variables.
- Using a **Logistic Regression classifier**, we can predict an accuracy of 68.63%.
- Using **Random Forest Classifier**, we can predict an accuracy of 88.16%.
- Using **K-Neighbor Classifier**, we can predict an accuracy of 79.37%.
- Using **Support Vector Machine Classifier**, we can predict an accuracy of 77.45%.
- Using **XG Boosting**, we can predict an accuracy of 77.19%.
- **Random Forest Classifier** performs best among all models.
- **Logistic Regression** is not giving good precision score
- Top 3 models are **Random Forest , K-Neighbor Classifier and Support Vector Machine** that gives best Precision, Recall ,ROC_AUC and F1 score.