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

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital. This can be proven by industry business reports and statistical data. For example, "The Federal Reserve

Bank of New York measures credit card delinquencies based on the percent of balances that are at least 90 days late. For the third quarter of 2019, that rate was about 8%, about the same level as in the previous quarter." Thus, assessing, detecting and managing default risk is the key factor in 2 generating revenue and reducing loss for the banking and credit card industry.

Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting. The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally underbanked with no credit history or new immigrants who have repaying power but lack credit history. According to a literature review report on analysing credit risk using machine and deep learning models, "credit risk management problems researched have been around credit scoring; it would go a long way to research how machine learning can be applied to quantitative areas for better computations of credit risk exposure by predicting probabilities of default."

The purpose of this project is to conduct quantitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision-making process.

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.

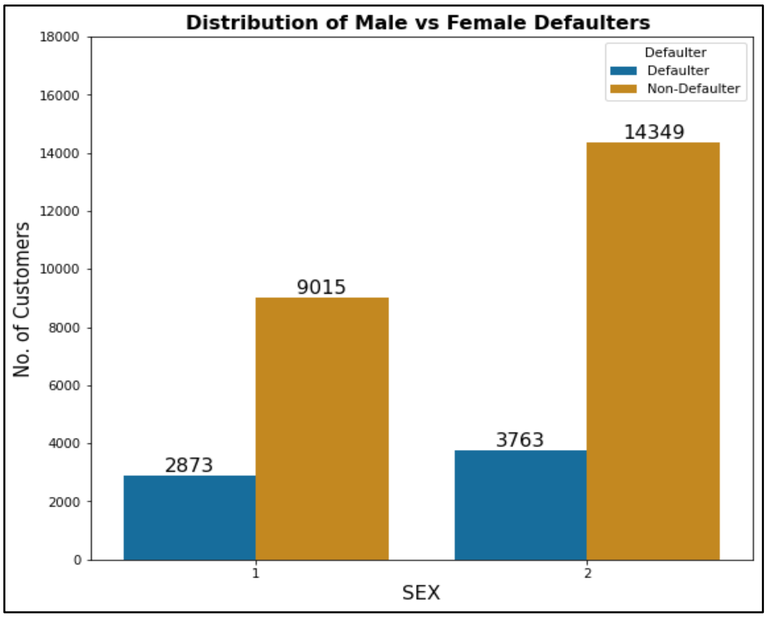
The main findings from exploratory analysis are as following:

- 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

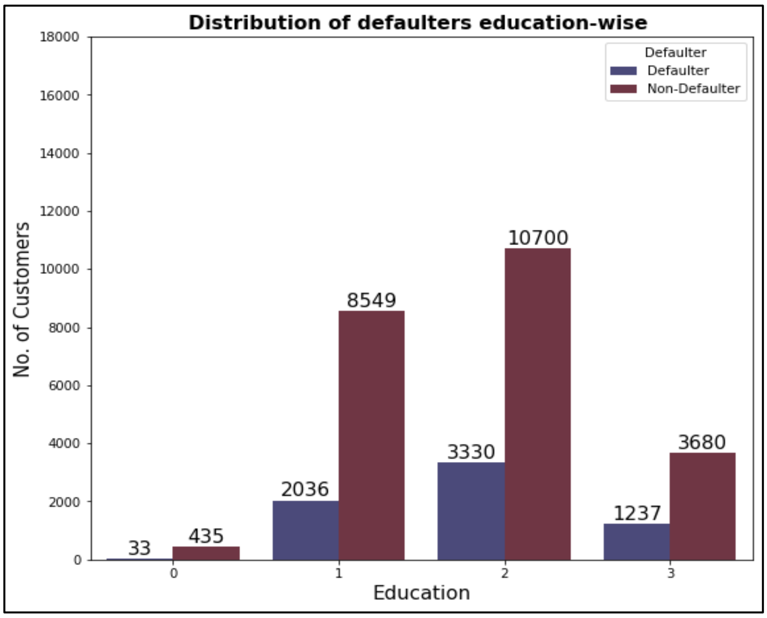
Females have more delayed payments than males in this dataset.

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.



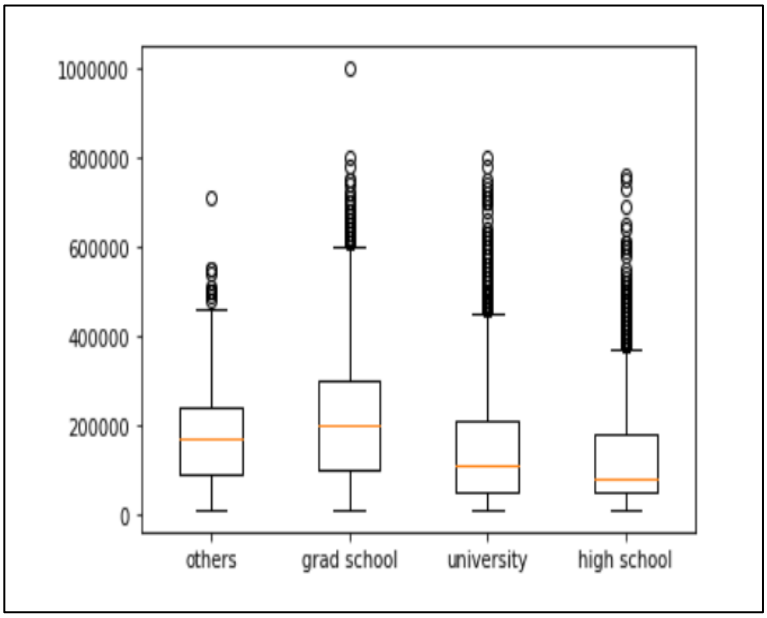
2. Education Variable

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.

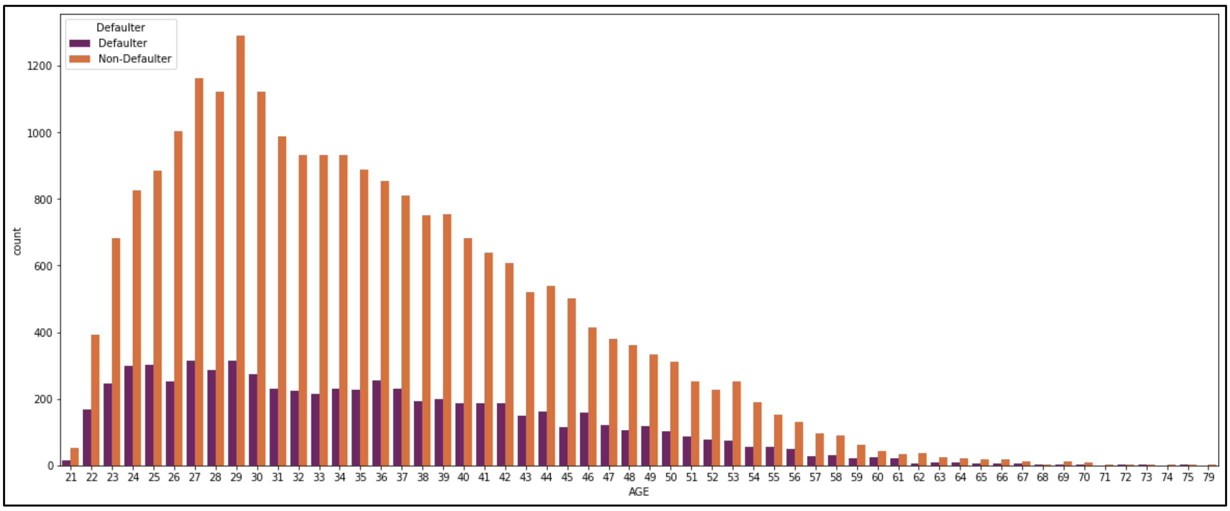
From the boxplot in figure 3, it is obvious that customers with grad school education have the highest 25% percentile, highest median, highest 75th percentile and highest maximum numbers, which suggests customers with higher education levels do get higher credit limits.



3.Age Variable

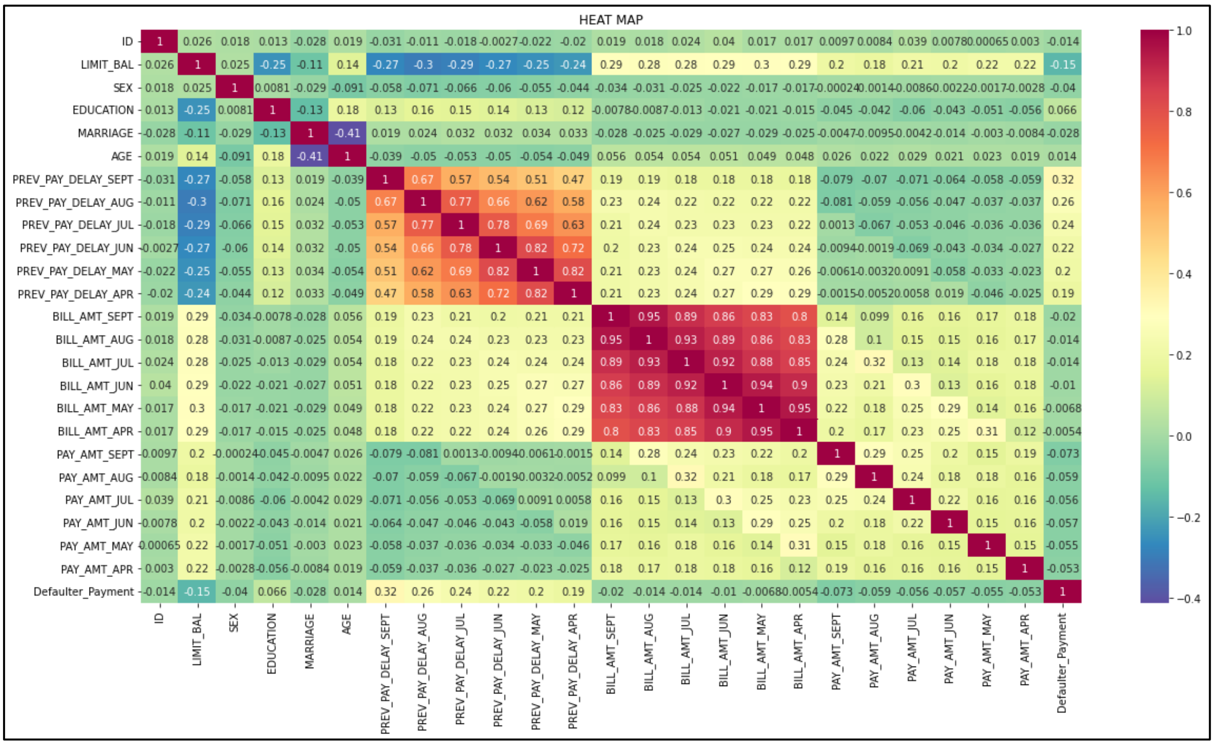
Middle-aged customers have the lowest default rate.

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.



	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PREV_PAY_DELAY_SEPT	PREV_PAY_DELAY_AUG	PREV_PAY_DELAY_JUL	PREV_PAY_DELAY_JUN	PREV_PAY_DELAY_MAY	PREV_PAY_DELAY_APR	BILL_AMT_SEPT	BILL_AMT_AUG	BILL_AMT_JUL	BILL_AMT_JUN	BILL_AMT_MAY	BILL_AMT_APR	PAY_AMT_SEPT	PAY_AMT_AUG	PAY_AMT_JUL	PAY_AMT_JUN	PAY_AMT_MAY	PAY_AMT_APR	Defaulter_Payment
ID	1	0.026	0.018	0.013	-0.028	0.019	-0.031	-0.011	-0.018	-0.0027	-0.022	-0.02	0.019	0.018	0.024	0.04	0.017	0.017	0.0097	0.0084	0.039	0.0078	0.00065	0.003	-0.014
LIMIT_BAL	0.026	1	0.025	-0.25	-0.11	0.14	-0.27	-0.3	-0.29	-0.27	-0.25	-0.24	0.29	0.28	0.28	0.29	0.3	0.29	0.2	0.18	0.21	0.2	0.22	0.22	-0.15
SEX	0.018	0.025	1	0.0081	-0.029	-0.091	-0.058	-0.071	-0.066	-0.06	-0.055	-0.044	-0.034	-0.031	-0.025	-0.022	-0.017	-0.017	-0.00024	-0.0014	-0.0086	-0.0022	-0.0017	-0.0028	-0.04
EDUCATION	0.013	-0.25	0.0081	1	-0.13	0.18	0.13	0.16	0.15	0.14	0.13	0.12	-0.0078	-0.0087	-0.013	-0.021	-0.021	-0.015	-0.045	-0.042	-0.06	-0.043	-0.051	-0.056	0.066
MARRIAGE	-0.028	-0.11	-0.029	-0.13	1	-0.41	0.019	0.024	0.032	0.032	0.034	0.033	-0.028	-0.025	-0.029	-0.027	-0.029	-0.025	-0.0047	-0.0095	-0.0042	-0.014	-0.003	-0.0084	-0.028
AGE	0.019	0.14	-0.091	0.18	-0.41	1	-0.039	-0.05	-0.053	-0.05	-0.054	-0.049	0.056	0.054	0.054	0.051	0.049	0.048	0.026	0.022	0.029	0.021	0.023	0.019	0.014
PREV_PAY_DELAY_SEPT	-0.031	-0.27	-0.058	0.13	0.019	-0.039	1	0.67	0.57	0.54	0.51	0.47	0.19	0.19	0.18	0.18	0.18	0.18	-0.079	-0.07	-0.071	-0.064	-0.058	-0.059	0.32
PREV_PAY_DELAY_AUG	-0.011	-0.3	-0.071	0.16	0.024	-0.05	0.67	1	0.77	0.66	0.62	0.58	0.23	0.24	0.22	0.22	0.22	0.22	-0.081	-0.059	-0.056	-0.047	-0.037	-0.037	0.26
PREV_PAY_DELAY_JUL	-0.018	-0.29	-0.066	0.15	0.032	-0.053	0.57	0.77	1	0.78	0.69	0.63	0.21	0.24	0.23	0.23	0.23	0.22	0.0013	-0.067	-0.053	-0.046	-0.036	-0.036	0.24
PREV_PAY_DELAY_JUN	-0.0027	-0.27	-0.06	0.14	0.032	-0.05	0.54	0.66	0.78	1	0.82	0.72	0.2	0.23	0.24	0.25	0.24	0.24	-0.0094	-0.0019	-0.069	-0.043	-0.034	-0.027	0.22
PREV_PAY_DELAY_MAY	-0.022	-0.25	-0.055	0.13	0.034	-0.054	0.51	0.62	0.69	0.82	1	0.82	0.21	0.23	0.24	0.27	0.27	0.26	-0.0061	-0.0032	0.0091	-0.058	-0.033	-0.023	0.2
PREV_PAY_DELAY_APR	-0.02	-0.24	-0.044	0.12	0.033	-0.049	0.47	0.58	0.63	0.72	0.82	1	0.21	0.23	0.24	0.27	0.29	0.29	-0.0015	-0.0052	0.0058	0.019	-0.046	-0.025	0.19
BILL_AMT_SEPT	0.019	0.29	-0.034	-0.0078	-0.028	0.056	0.19	0.23	0.21	0.2	0.21	0.21	1	0.95	0.89	0.86	0.83	0.8	0.14	0.099	0.16	0.16	0.17	0.18	-0.02
BILL_AMT_AUG	0.018	0.28	-0.031	-0.0087	-0.025	0.054	0.19	0.24	0.24	0.23	0.23	0.23	0.95	1	0.93	0.89	0.86	0.83	0.28	0.1	0.15	0.15	0.16	0.17	-0.014
BILL_AMT_JUL	0.024	0.28	-0.025	-0.013	-0.029	0.054	0.18	0.22	0.23	0.24	0.24	0.24	0.89	0.93	1	0.92	0.88	0.85	0.24	0.32	0.13	0.14	0.18	0.18	-0.014
BILL_AMT_JUN	0.04	0.29	-0.022	-0.021	-0.027	0.051	0.18	0.22	0.23	0.25	0.27	0.27	0.86	0.89	0.92	1	0.94	0.9	0.23	0.21	0.3	0.13	0.16	0.18	-0.01
BILL_AMT_MAY	0.017	0.3	-0.017	-0.021	-0.029	0.049	0.18	0.22	0.23	0.24	0.27	0.29	0.83	0.86	0.88	0.94	1	0.95	0.22	0.18	0.25	0.29	0.14	0.16	-0.0068
BILL_AMT_APR	0.017	0.29	-0.017	-0.015	-0.025	0.048	0.18	0.22	0.22	0.24	0.26	0.29	0.8	0.83	0.85	0.9	0.95	1	0.2	0.17	0.23	0.25	0.31	0.12	-0.0054
PAY_AMT_SEPT	0.0097	0.2	-0.00024	-0.045	-0.0047	0.026	-0.079	-0.081	0.0013	-0.0094	-0.0061	-0.0015	0.14	0.28	0.24	0.23	0.22	0.2	1	0.29	0.25	0.2	0.15	0.19	-0.073
PAY_AMT_AUG	0.0084	0.18	-0.0014	-0.042	-0.0095	0.022	-0.07	-0.059	-0.067	-0.0019	-0.0032	-0.0052	0.099	0.1	0.32	0.21	0.18	0.17	0.29	1	0.24	0.18	0.18	0.16	-0.059
PAY_AMT_JUL	0.039	0.21	-0.0086	-0.06	-0.0042	0.029	-0.071	-0.056	-0.053	-0.069	0.0091	0.0058	0.16	0.15	0.13	0.3	0.25	0.23	0.25	0.24	1	0.22	0.16	0.16	-0.056
PAY_AMT_JUN	0.0078	0.2	-0.0022	-0.043	-0.014	0.021	-0.064	-0.047	-0.046	-0.043	-0.058	0.019	0.16	0.15	0.14	0.13	0.29	0.25	0.2	0.18	0.22	1	0.15	0.16	-0.057
PAY_AMT_MAY	0.00065	0.22	-0.0017	-0.051	-0.003	0.023	-0.058	-0.037	-0.036	-0.034	-0.033	-0.046	0.17	0.16	0.18	0.16	0.14	0.31	0.15	0.18	0.16	0.15	1	0.15	-0.055
PAY_AMT_APR	0.003	0.22	-0.0028	-0.056	-0.0084	0.019	-0.059	-0.037	-0.036	-0.027	-0.023	-0.025	0.18	0.17	0.18	0.18	0.16	0.12	0.19	0.16	0.16	0.16	0.15	1	-0.053
Defaulter_Payment	-0.014	-0.15	-0.04	0.066	-0.028	0.014	0.32	0.26	0.24	0.22	0.2	0.19	-0.02	-0.014	-0.014	-0.01	-0.0068	-0.0054	-0.073	-0.059	-0.056	-0.057	-0.055	-0.053	1

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 models 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.

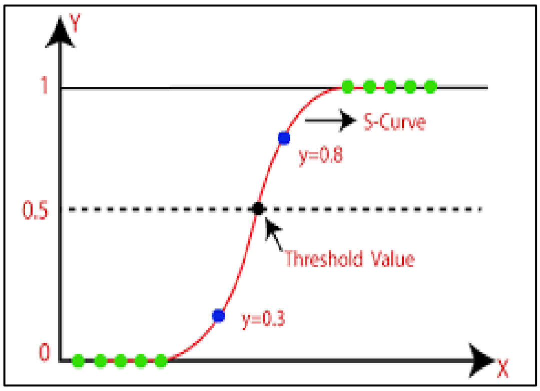
VII. Classification Analysis

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

1. Algorithms-

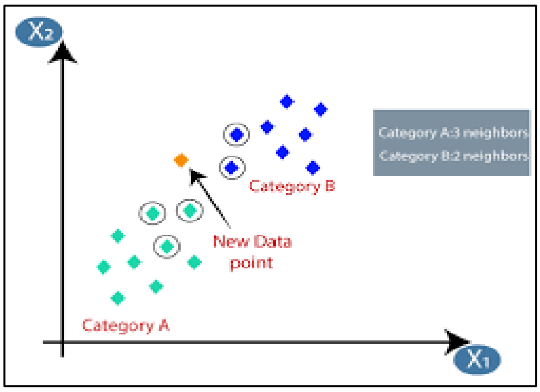
1. Logistic Regression:

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).



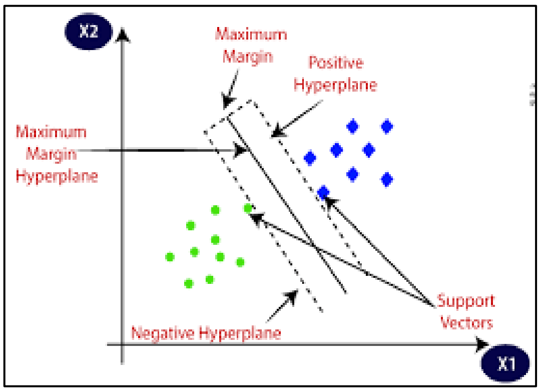
2. K-Neighbors Classifier:

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.



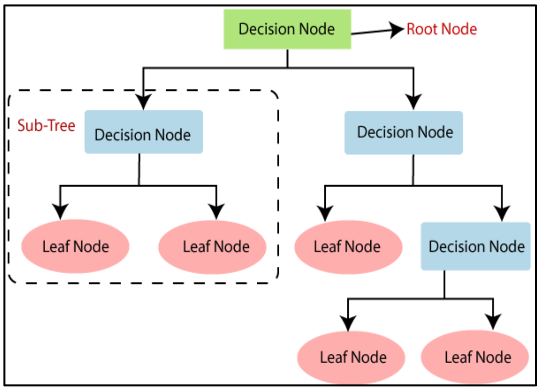
3.Support Vector Classifier:

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.DecisionTree Classifier:

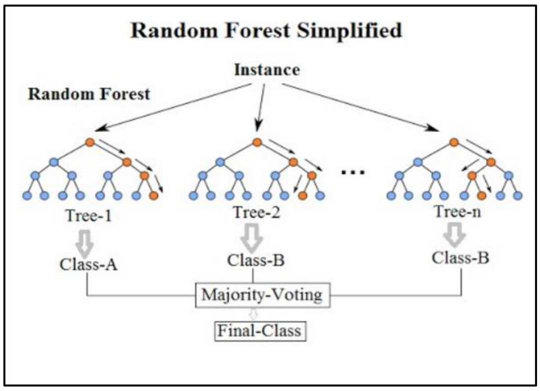
Decision tree build classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. It follows the 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.

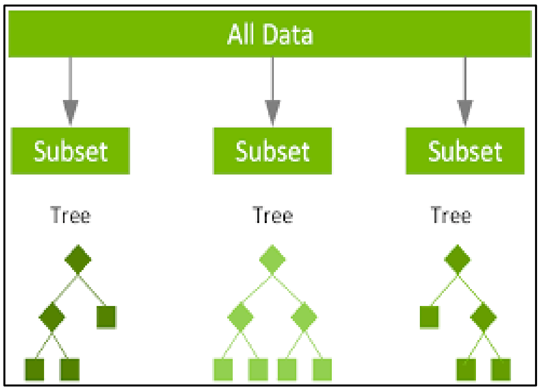
Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model’s prediction.



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.

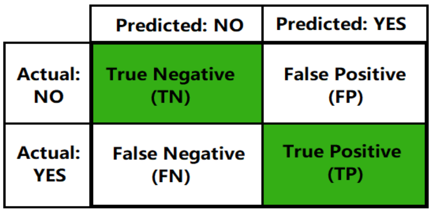
3. Model Performance-

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.

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

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).

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

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

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

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].

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

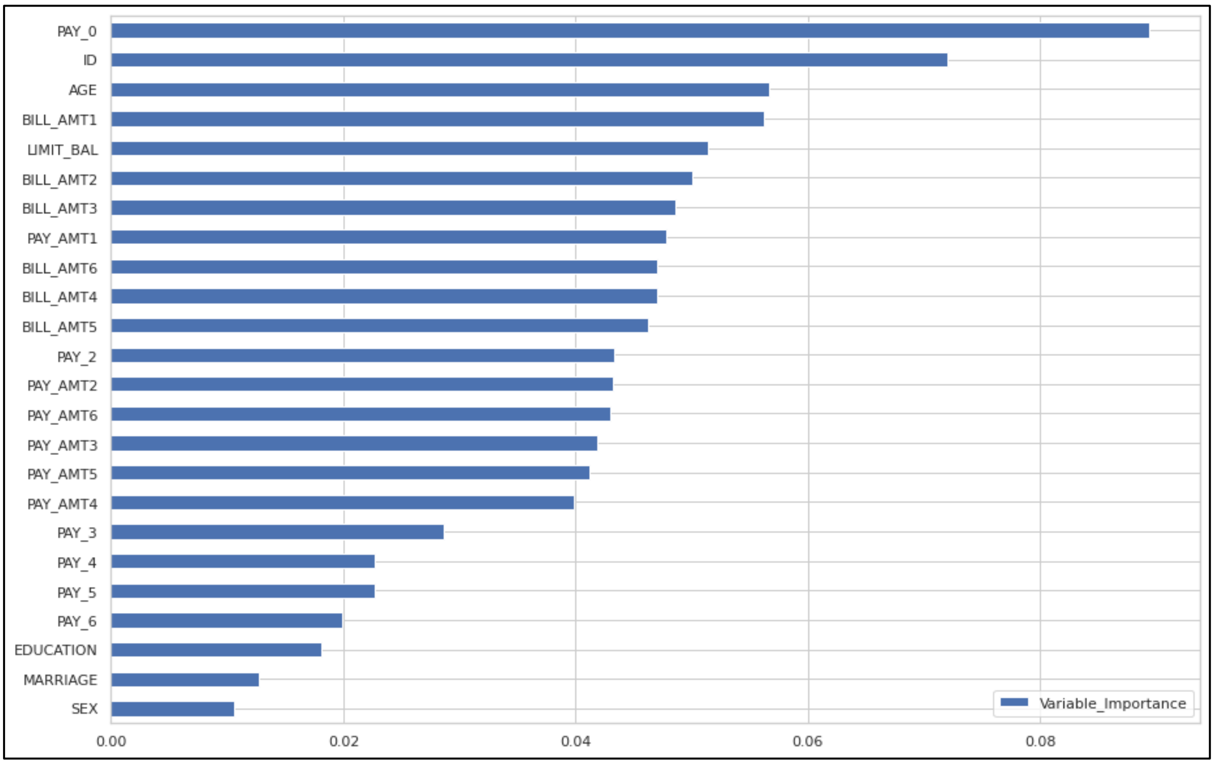
6) **AUC-ROC Curve-**

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.

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

Feature Importance:

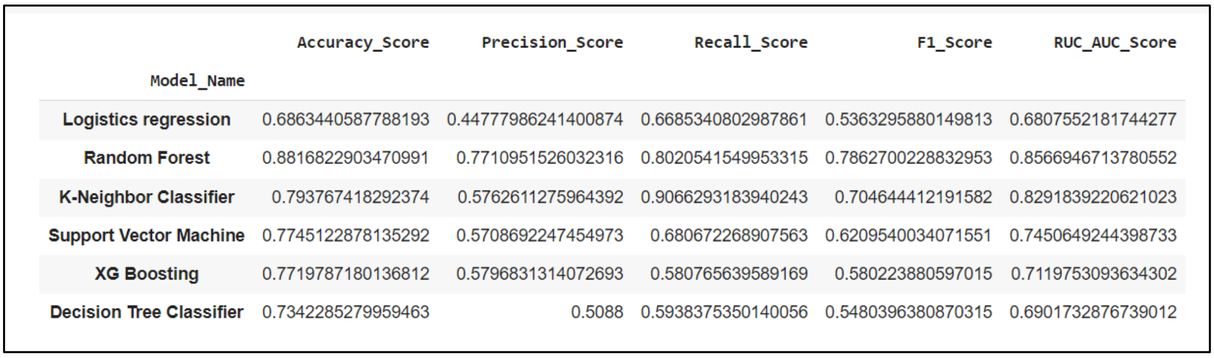
Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Feature importance scores play an important role in a predictive modeling project, including providing insight into the data, insight into the model, and the basis 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.



	Accuracy_Score	Precision_Score	Recall_Score	F1_Score	RUC_AUC_Score
Model_Name					
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

- There is neither null nor duplicate values in our dataset.
- We rename the column for better understanding.
- We check the distribution of defaulter vs non defaulter and we see that around 78% are non-defaulter and 22% are defaulter.
- We have found the proportion of defaulters with respect to Marriage, Education, Sex feature and we found that:
  - Most of the defaulters are Female
  - Most of the defaulters are from university
  - Marital status is Single
- We have used the boxplot to detect the outliers and we see that there are so many outliers in the data so we apply IQR (Inter Quartile Range) which is one of the techniques to remove outliers.
- We plot heat map to see the correlation between variable gives the graphical representation between all the variables and we see that age and marriage are highly negatively correlated to each other.
- After that we build the six models Logistic Regression, Random Forest, XG Boosting, Support Vector Classifier, K-Neighbor Classifier, Decision Tree Classifier and we obtained the best accuracy from Random Forest Classifier.
- Using a **Logistic Regression classifier**, we can predict an accuracy of 69% .
- Using **Random Forest Classifier**, we can predict an accuracy of 88. % .
- Using **K-Neighbor Classifier**, we can predict an accuracy of 79% .
- Using **Support Vector Machine Classifier**, we can predict an accuracy of 78% .
- Using **XG Boosting**, we can predict an accuracy of 77%.
- Using **Decision Tree Classifier** , we can predict an accuracy of 77%.
- **Random Forest Classifier** performs best among all models.
- **Logistic Regression** are not giving good precision score.
- Top 3 models are **Random Forest , K-Neighbor Classifier and Support Vector Classifier** that gives best Precision, Recall, ROC_AUC and F1 score.