PROJECT REPORT

MACHINE LEARNING

TEAM MEMBERS:

PANKAJ BHATIA

SUNIL KUMAR JAYAKRISHNAN

YUKTI KAKKAR

JEEVAN JOT SINGH

**Objective:**

The classification goal is to predict the likelihood of a liability customer buying personal loans. Use different classification models to predict the likelihood of a liability customer buying personal loans. This case is about a bank (Thera Bank) whose management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

**Data Description:**

The file Bank.xls contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan).

**Attributes:**

* ID : Customer ID
* Age : Customer's age in completed years
* Experience : #years of professional experience
* Income : Annual income of the customer ($000)
* ZIP Code : Home Address ZIP code.
* Family : Family size of the customer
* CCAvg : Avg. spending on credit cards per month ($000)
* Education : Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
* Mortgage : Value of house mortgage if any. ($000)
* Personal Loan : Did this customer accept the personal loan offered in the last campaign?
* Securities Account : Does the customer have a securities account with the bank?
* CD Account : Does the customer have a certificate of deposit (CD) account with the bank?
* Online : Does the customer use internet banking facilities?
* Credit card : Does the customer use a credit card issued by Universal Bank?

**Domain:** Banking

**Libraries Used:**

* Pandas
* NumPy
* Seaborn
* Matplotlib
* sklearn
* SciPy

The first step taken by the group was to understand the desired packages for the project.

The group was aware about Pandas and NumPy, yet more information was expected to show the visual representation of our work and keeping in mind of exploratory analysis, we further got to realize that we could require help of Seaborn, Matplotlib and Sklearn.

Train\_test\_split is a capability in Sklearn model choice for dividing information clusters into two subsets: for preparing information and for testing information. With this capability, we do not have to partition the dataset manually.

SciPy is a scientific computation library that uses NumPy underneath. SciPy stands for Scientific Python. It provides more utility functions for optimization, stats and signal processing. Like NumPy, SciPy is open source so we can use it freely.

**Algorithms:**

1. **Linear Regression:**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

1. **Naïve Bayes:**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.It is mainly used in *text classification* that includes a high-dimensional training dataset.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.**It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.

1. **KNN:**

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-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

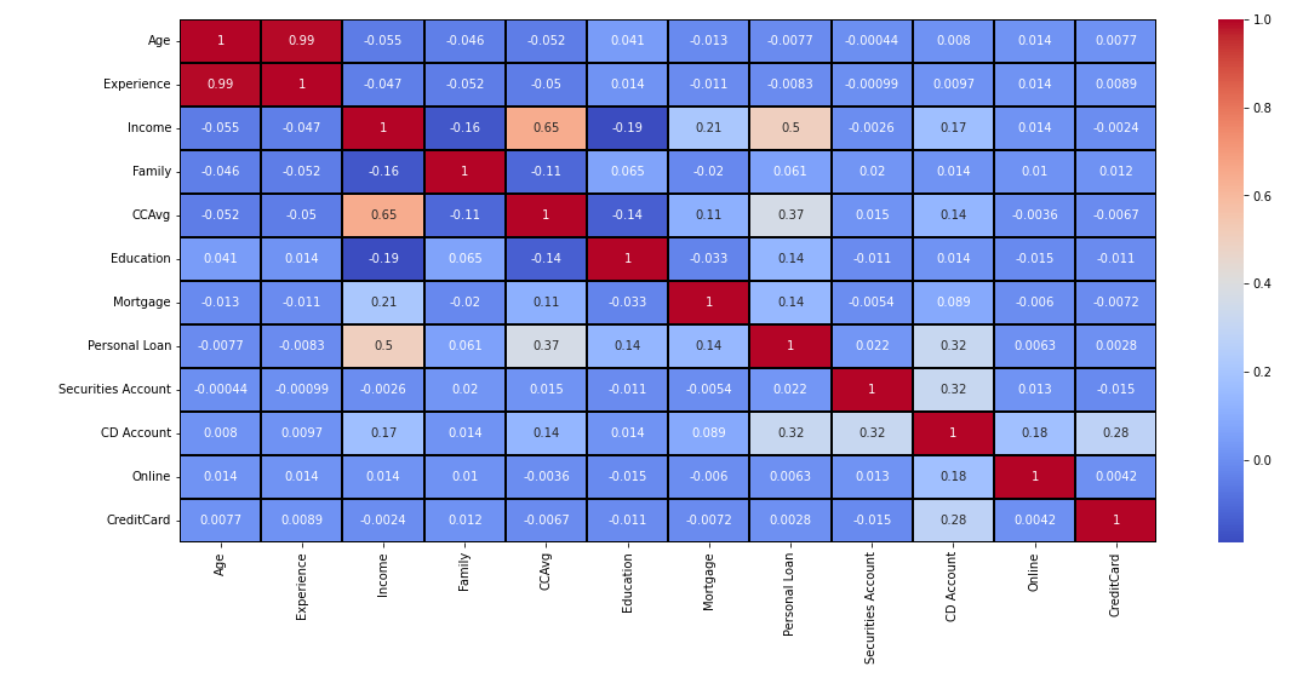
K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.

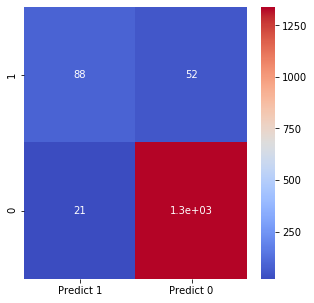
It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

**Outputs:**

****

Heat map shows that Age and Experience are highly correlated. Personal loan is related to (Income, CCAvg,Mortgage,CD Account). Income influences CCAvg, Mortgage, Personal Loan and CD Account. CD Account is influences (Income, CCAvg, Personal Loan, Securities Account, online and credit card).



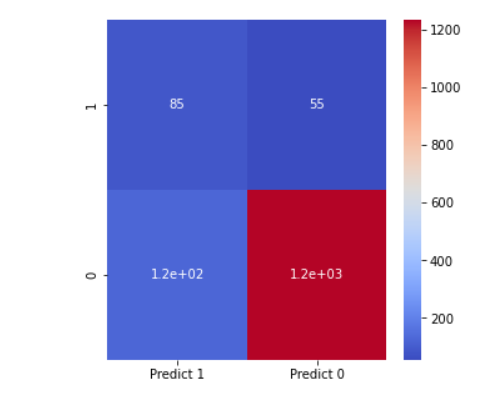
True Positive (TP): our model is correctly predicting 88 from testing data

True Negative (TN): our model correctly predicts that 1339 will not take loan

False Positive (FP): our model is incorrectly predicting that they will take loan (Type 1 error) 21

False Negative (FP): our model is incorrectly predicting that they will not take loan (Type 2 error) 52

Confusion Matrix for Logistic Regression



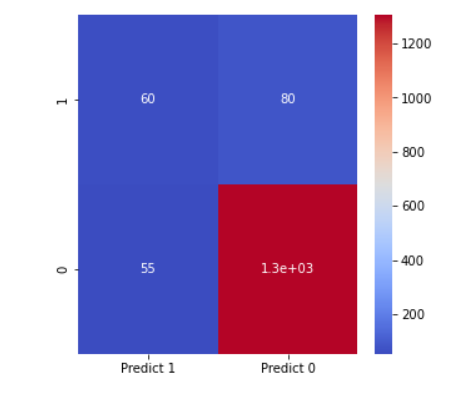
True Positive (TP): our model is correctly predicting 85 from testing data

True Negative (TN): our model correctly predicts that 1235 will not take loan

False Positive (FP): our model is incorrectly predicting that they will take loan (Type 1 error) 125

False Negative (FP): our model is incorrectly predicting that they will not take loan (Type 2 error) 55

Confusion Matrix for Naïve Bayes



True Positive (TP): our model is correctly predicting 60 from testing data

True Negative (TN): our model correctly predicts that 1305 will not take loan

False Positive (FP): our model is incorrectly predicting that they will take loan (Type 1 error) 55

False Negative (FP): our model is incorrectly predicting that they will not take loan (Type 2 error) 80

Confusion Matrix for KNN

**Conclusion:**



The three models were trained in three different algorithms. The Logistic Regression has the best accuracy and lower type 1 and type 2 error values. So the Linear Regression model is the best model of all three models.

**References:**

* Data analytics use cases in Banking
* Machine Learning for Financial Marketing
* <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>
* https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c