**Bank Personal Loan Modelling Using Random Forest Classification**

1. **Introduction:**

The Random forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees.  
The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically *a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.*

1. **Problem Statement:**

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).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success.

This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

1. **Objective:**

* Cleaning Dataset
* Dataset exploration using various types of data visualization.
* Identifying suitable patterns for modeling (Pattern Recognition)
* Build Random Forest models that can predict personal loan customers.

1. **Inputs:**

We have taken Bank\_Personal\_Loan.csv file as our input. We have taken following features as our input:

* Age
* Experience
* Income
* Family
* CCAvg Education
* Mortgage
* Personal Loan
* Securities Account
* CD Account
* Online

1. **Output:**

* CreditCard

Our Output Predicts whether the customer will accept a personal loan?

(0:No & 1:Yes): [0]

1. **Data Set Info:**

This dataset have 5000 row and 14 columns and personal Loan is target. other features are:  
• id : Customer ID  
• age : Customer's age in completed years  
• experience : years of professional experience  
• income : Annual income of the customer  
• zipcode : Home Address ZIP code.

• family : Family size of the customer

• ccavg : Avg. spending on credit cards per month

• education : Education Level. Undergrad Graduate Advanced/Professional

• mortgage : Value of house mortgage if any.

• personalloan : Did this customer accept the personal loan offered in the last campaign?  
• securitiesaccount : Does the customer have a securities account with the bank?

• cdaccount : Does the customer have a certificate of deposit (CD) account with the bank?  
• online : Does the customer use internet banking facilities?  
• creditcard : Will customer purchase a credit card?

1. **Candidate algorithms**
   1. **K-Nearest Neighbour (KNN) Algorithm :**

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

K nearest neighbors is a simple algorithm that stores all available cases and classifies new data points on a similarity measure (distance function).

A new data point is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the new data point is assigned to the class of its nearest neighbor.

Chart, scatter chart

Description automatically generated

Figure 1 : KNN Model Diagram

* 1. **Random Forest Algorithm :**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. In this paper it is used for classification. Random forest algorithm is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. In simpler words, it creates decision trees on randomly selected data samples from the data set, gets predictions from each tree and selects the best possible solution by the means of voting.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Diagram

Description automatically generated

Figure 2: Random Forest Model Diagram

* 1. **Support Vector Machines (SVM) :**

SVM is a ML algorithm which belongs to a class of classification algorithms. The original form of SVM was invented on the year 1963,but in the year 1992 a new way was invented to create a nonlinear classifier These classes of algorithms are used for Classification algorithms are used to predict the binary values such as true or false, male or women, emails as spam or not, faulty or not faulty in our case More formally a SVM creates a hyper plane to separate linearly separable points and according to that it gives the prediction result.

SVM is a supervise ML method which is used both classification and regression problems. SVM creates a hyperplane in multidimensional space to divide different classes. SVM finds a max hyper plane or line that segregates the dataset into classes.

* Vectors — the vectors which are nearest to the hyperplane.
* Hyperplane —It is a subspace with dimension 1 lower than its ambient space.
* Margin — the range between the nearest data point and the hyperplane from either side.
* Kernel — It is a mathematical function used to transform input data into a different form. Common kernel functions are included linear, nonlinear, polynomial, etc.

**The mathematical equation of the hyper plane** :

Here, **W** is the normal vector

**X** is data points

**b** is linear coefficient

Diagram

Description automatically generated

Figure 3: SVM Model Diagram

1. **Libraries Used**

**8.1. Pandas**

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

**8.2. NumPy**

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It is open-source software. It contains various features including these important ones:

* A powerful N-dimensional array object
* Sophisticated (broadcasting) functions
* Tools for integrating C/C++ and Fortran code
* Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

**8.3. Scikit-learn**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

1. **Performance measure used**

Classification methods Random Forest has been used for this work. For a classification problem the following outcomes are possible:

* True positives: the prediction belons to the class that it really belongs to.
* True negatives: prediction does not belong to the class that it really belons to
* False positives: prediction belong to a class that it does not belong to
* False negatives: prediction does not belong to a class when it should.

The outcomes can be visualised in a confusion matrix.

The main metrices for the evaluation of a classification model can also be called accuracy precision and recall.

* Accuracy: It is the percentage of correct predictions for the test data.

Accuracy =

* Precision: It is defined as the ratio to the true positives to all which were predicted to belong in a certain class

Precision=

* Recall: It is defined ratio of the predictions that belong to a classification to the ones that really belong in that classification.

recall =

1. **Advantages of Random Forest Classification:**

* It reduces overfitting in decision trees and helps to improve the accuracy
* It is flexible to both classification and regression problems
* It works well with both categorical and continuous values
* It automates missing values present in the data
* Normalising of data is not required as it uses a rule-based approach.

1. **Disadvantages of Random Forest Classification:**

* It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
* It also requires much time for training as it combines a lot of decision trees to determine the class.
* Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.