**1. Introduction**

**1.1 Abstract:-**

The Problem is based on the domain of the Banking sector where the bank wants to predict the Churn of a customer depending upon the previous data of the customer. By churn it is meant that the bank wants to predict if a customer would be a defaulter in the next quarter depending upon its previous credit history.

**What is the problem?**

The main problem is to predict if a customer would be credit defaulter or not depending upon the previous data of the customer.

**Why is it important?**

It is important from a bank’s perspective in order to maintain business and customer relationship/ Apart from that if someone could be predicted as a defaulter then primitive measures can be taken in order to ensure that such violations do not happen.

**What is your basic approach?**

The basic approach of solving this problem was first studying the data , then bringing out insights from the dataset and after that I have followed a machine learning pipeline in order to solve the problem.

The ML Pipeline that I have followed is :

* Importing the necessary libraries and the dataset
* Performing Data Preprocessing (Exploratory Data Analysis and Data Manipulation)
* Modelling using Logistic Regression, KNN and Random Forest
* Performing Prediction
* Visualization in between Actual and predicted Values

**The environment used was python 3.7 and the libraries such as numpy, pandas, matplotlib , Standard Scaler and Scikit Learn module were used for Scientific computations.**

**2. Problem Definition and Algorithm**

**The Problem Statement is as follows :-**

**Customer Churn Prediction:**

A Bank wants to take care of customer retention for its product: savings accounts. The bank wants you to identify customers likely to churn balances below the minimum balance. You have the customers information such as age, gender, demographics along with their transactions with the bank.

Your task as a data scientist would be to predict the propensity to churn for each customer.

**Data Dictionary**

There are multiple variables in the dataset which can be cleanly divided into 3 categories:

I. Demographic information about customers

* **customer\_id** - Customer id
* **vintage** - Vintage of the customer with the bank in a number of days
* **age** - Age of customer
* **gender** - Gender of customer
* **dependents** - Number of dependents
* **occupation** - Occupation of the customer
* **city** - City of the customer (anonymized)

II. Customer Bank Relationship

* **customer\_nw\_category** - Net worth of customer (3: Low 2: Medium 1: High)
* **branch\_code** - Branch Code for a customer account
* **days\_since\_last\_transaction** - No of Days Since Last Credit in Last 1 year

III. Transactional Information

* **current\_balance** - Balance as of today
* **previous\_month\_end\_balance** - End of Month Balance of previous month
* **average\_monthly\_balance\_prevQ** - Average monthly balances (AMB) in Previous Quarter
* **average\_monthly\_balance\_prevQ2** - Average monthly balances (AMB) in previous to the previous quarter
* **current\_month\_credit** - Total Credit Amount current month
* **previous\_month\_credit** - Total Credit Amount previous month
* **current\_month\_debit** - Total Debit Amount current month
* **previous\_month\_debit** - Total Debit Amount previous month
* **current\_month\_balance** - Average Balance of current month
* **previous\_month\_balance** - Average Balance of previous month
* **churn** - Average balance of customer falls below minimum balance in the next quarter (1/0)

**2.1 Task Definition**

The task was to predict those customers likely to churn balances below the minimum balance.The customers information such as age, gender, demographics along with their transactions with the bank.The task as a data scientist was to predict the propensity to churn for each customer.

The ML Pipeline that I have followed is :

* **Importing the necessary libraries and the dataset:**  
  Here, the libraries such as numpy, pandas and matplotlib were called. **Numpy** is known to be Numerical Python which is responsible for performing all the Numerical tasks in this project whereas **Pandas** would make the data frame and **Matplotlib** was used for visualization..
* **Performing Data Preprocessing (Exploratory Data Analysis and Data Manipulation:-**In this step the data was thoroughly analysed and the steps such as Univariate Analysis and Bivariate Analysis were performed. Univariate Analysis would be the analysis of a data if we are studying/ analysing a particular data majorly depending upon its type (for e.g Continuous or Categorical Type )

If the data is said to be a continuous data then in order to do a univariate analysis we check the distribution of the data i.e, we check if the data is normally distributed or not and if the data is said to be a categorical data then we would check the bar plot of that data..In order to perform a Bivariate Analysis (The study of two data at an instance) we would have to plot the scatter plot in order to check the relationship in between them.

* **Modelling using Logistic Regression, KNN and Random Forest**

The modelling was done by using Logistic Regression , Random Forest and KNN and at the end it was ensemble by using VotingClassifier.

The Following are the algorithms that were used for performing the predictions

**2.2.1 Logistic Regression:-**

Classification techniques are an essential part of machine learning and data mining applications. Approximately 70% of problems in Data Science are classification problems. There are lots of classification problems that are available, but the logistics regression is common and is a useful regression method for solving the binary classification problem. Another category of classification is Multinomial classification, which handles the issues where multiple classes are present in the target variable. For example, IRIS dataset a very famous example of multi-class classification. Other examples are classifying article/blog/document category.

Logistic Regression can be used for various classification problems such as spam detection. Diabetes prediction, if a given customer will purchase a particular product or will they churn another competitor, whether the user will click on a given advertisement link or not, and many more examples are in the bucket.