# Customer Churn Prediction:Bank

Using Machine Learning Technique

Mohammed Afzal Mohiuddin
#ID:220920526
December 2022

**Abstract**

In every business, the number of service suppliers is rising quickly. Customers have a wide range of options these days when deciding where to invest their money in the banking industry. As a result, customer turnover and engagement have emerged as major challenges for the majority of banks. Customer churn is the rate at which customers stop doing business with a company. In the machine learning domain, churn prediction is a common case. This project aims to develop a method to predict customer churn in a bank using machine learning techniques, which is a branch of artificial intelligence. The research promotes the exploration of the likelihood of churn by analyzing customer behavior. The random forest, k-nearest neighbors, and logistic regression classifier are used to predict the churn for the banks. The experimentation was conducted on the churn modeling data set provided by the institution. To identify appropriate models with greater precision and predictability, the results are compared. As a result, the accuracy of the Random Forest model is higher than that of other models. Machine learning is a useful approach for predicting customer churn.

**Index Terms**— customer churn in the bank,k-Nearest neighbor, random forest, logistic regression,

# Introduction

## Machine Learning

Machine learning (ML) has proven to be one of the most game-changing technological advancements of the past decade. Python is the most potent and most usable programming language. In 2022,[1] the most well-liked languages for programming data analysis will be machine learning (ML) and deep learning (neural networks). The scripts can be created by using an IDE such as Jupyter, Spyder, or interactive notebooks that are Google Colab, which is used to perform the coding of the given tasks. The packages in Python include matplotlib and seaborn, which are used for plotting the visualizations. Sklearn, a machine learning (ML) tool, is also available. Supervised and unsupervised learning are the most commonly used types of machine learning (ML). One of the fundamental types of machine learning (ML) is supervised learning.Supervised learning is further divided into two types: classification and regression. A classification separates the data, and regression fits the data. On the other hand, unsupervised learning involves algorithms trained on unlabeled data, which means that human labor is not required to make the data set machine-readable. It is classification-supervised machine learning. Customer churn prediction using machine learning will help you identify risky customers and understand why they are willing to leave.

## Churn

Churn is a measure of the percentage of accounts that cancel their subscriptions or choose not to renew. A high churn rate can negatively impact your monthly recurring revenue (MRR) and can also indicate dissatisfaction with a product or service. In any banking sector, customers are the most valuable resources because they are viewed as the primary source of revenue. Companies nowadays are aware that they need to work hard to win over new clients and keep the ones they already have. Churners are people who switch companies for a variety of reasons. The organization must be able to accurately predict customer behavior, draw links between customer attrition, and maintain elements within its control in order to reduce customer churn. The binary classification challenge of churn prediction separates churners from non-churners.

### Churn Management System

As we know that churn is a combination of the words "change" and "turn." This phrase refers to a circumstance in which a client wishes to move service providers but that transfer must be prevented (switch). Churn, in a phrase, is the shifting or loss of clients. It costs more to get new consumers than it does to keep your current ones. Churn Management deals with the challenge of re-engaging consumers who have terminated their contracts or are considering changing their purchase or use patterns. Some studies have indicated that it costs six to seven times more to recruit a new client than it does to maintain an existing one.

### Objective of Churn Management

There are three main objective of churn management

1. Identifying customers at risk of migration
2. Reduced churn rate
3. Recovering old customer through active management relationship

These are objectives that should play an important role in every company. Churn Management has only been the focus of conceptual and empirical research knowing which clients are most likely to depart or stop using your service is known as customer churn prediction. This forecast is significant for many banks. This is due to the fact that it is sometimes more expensive to acquire new clients than to keep old ones. When a client is at risk of leaving, you need to know precisely what marketing efforts to make with them to increase the possibility that they will stay.Customers terminate their memberships for a variety of behaviors, preferences, and causes. To retain each of them on your customer list, it is crucial to engage in active communication. You must understand which marketing initiatives and at what times are most successful for specific clients.

### Importance monitoring of Customer Churn Prediction

It is very important step to analyze the the importance of the customer churn prediction because it costs more to attract new customers than it does to sell to existing ones, customer turnover is crucial. This measure determines whether a firm succeeds or fails. Successful customer retention raises the average lifetime value of the customer, increasing the value of all subsequent sales and boosting unit profits.

### Advantages of Customer Churn Monitoring

The advantages are

1. Increases profit
2. Customer experience Improved
3. Optimization of services and products
4. Customer Retention

# Problem Statement

Churn prediction is generally considered a major use case in banking business. 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.The main issue to resolve is to predict if a customer would be credit defaulter or not depending upon the previous data of the customer.

### Aim of Project

The project's goal is to forecast bank churn. The purpose of this project is to develop a machine learning model that can reliably predict which customers will quit a company, allowing the business owner to make informed marketing choices. The strategy used to tackle this challenge involved first evaluating the data, then extracting insights from the given dataset, and then using a machine learning process.There will be three models utilised. Predict the churn from our sample with ease using these three machine learning models. Additionally, it is believed that this effort will demonstrate how same graphical analysis approaches may be used to explain customer calls in any analogous social phenomena, adding to the body of information already known about data analytics.

### Objective of Project

The project's objectives will be to: Precisely identify the above-mentioned problem

1. Understanding a problem and final goal
2. Data collection
3. Data Pre-processing
4. Data Splitting and testing
5. Model Implementation

These are the main objective of the project, after fulfilling these major objectives we accomplish our goals and able to predict the churn form the banks.

## Data Set Analysis

## Data set Collection

In machine learning is a collection of data points that a computer may process for analysis and prediction as a single unit. This means that since machines don't see data in the same way that people do, the data collected should be standard and intelligible. The dataset was given by university.

### Information of Data Set

This is the data set Bank , which contains the **1000** rows and **12** columns. Each sample contains 11 features and 1 targeted variable "churn" which indicates the class of the sample.In 10,000 line sample there are **7963** samples belong to class "no (0)" and **2037** samples belongs to class "yes (1).

### Description of Data Set

There are 11 input features and one targeted feature are

| Column Name | Datatype of Column | Data set Description |
|---|---|---|
| customer id | Numerical(integer) | contain ids numbers |
| credit score | Numerical(Int) | credit score of customer |
| country | String(object) | residences country |
| gender | String(object) | customer genders |
| age | Numerical(int) | Customer ages |
| tenure | Numerical (int) | Time period of customers |
| balance | Numerical(float) | Customer balance |
| products number | Numerical(int) | Products purchased |
| credit card | Numerical(int) | customer has a credit card |
| active member | Numerical(int) | active users |
| estimated salary | Numerical(int) | Customer estimated Salary |
| churn | Numerical(int) | Target variable |

Table 1: Data Set Description

# Methodology

Details on the project's strategy for anticipating customer churn for banks are provided in this section. The following step to be followed for accomplish the prediction of the churn these are collecting the accurate data for the prediction. The good and balanced data have a better results and good prediction percentage.There are 4 major steps, these are

1. Collection of Data set
2. Data Preparation
3. Data Pre-preprocessing
4. Model Implementations

### 1.Collection of Data set

It is 1st step for predicting the churn, that are collecting the dataset, a good and clean dataset always provide a accurate accuracy of the prediction.

### 2.Data Preparation

Data preparations is 2nd most important step to be follow to predict the churn for bank.Data preparation, which is required to clean the data and make it acceptable for the model, increases the accuracy and efficacy of a machine learning model. It includes the EDA (Exploratory Data Analysis) of the dataset.Its

contains the loading, reading, information,description of data set and also Visualizations of the dataset. The visualization is design by using the basic pythons libaries.(Mathplot,seaborn) etc.

### 3.Data Pre-Processing

Data is prepared for primary processing and further exploration during the program's significant pre - processing stage. It is a fundamental and crucial phase in the creation of a machine learning model. Every time we implemented a machine learning model, we needed clean, accurate data.Followings steps includes,data preparation and EDA steps.Furthermore, Data preprocessing includes encoding and splitting dataset into training and testing sets.Encoding is a technique of converting categorical variables into numerical values so that it could be easily fitted to a machine learning model[2].There are two steps of the converting the categorical data into numeric data these involves label encoding and other one is One-hot Encoding. It is an important pre-processing step for the structured dataset in supervised learning.

### 4.Model Implementation

A model implementation is the output of one or more model developers who, using their engineering skills, integrate the model's abstract idea into the model speciation's implementation. Depending on the target infrastructures, this implementation may be on any number of computer platforms. A machine learning model is created by learning from training data and extrapolating it. The customer churn prediction for banking system uses three supervised machine learning models.In supervised machine learning model the classifier models are implemented to predict the churn of banking system. These are following

1. Random Forest Classifier
2. Logistic Regression Classifier
3. K-Nearest Neighbour Classifier

# Results and Conclusion

# Results

In this section, we discuss the results of the model we implemented in the code, for the prediction of the churn. There are three model implemented for the prediction of churn to get the better results. The churn prediction is classified using machine learning classifiers.. Three machine learning algorithms logistic regression, K-nearest neighbor classifier, and random forest will be analysed and evaluated.

### Function of evaluation

Accuracy, precision, recall, and error rate can all be used to assess a classification model's performance. Model of the confusion matrix for performance characteristic evaluation was selected. Using group test data for which the real values are known, a confusion matrix is used to assess the effectiveness of a classification model. The values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) a Afterwards, use the value in the following equation to calculate the model's performance. It is used to evaluate how well the machine learning model is working

### Confusion matrix

In the below table,

| **Classifiers** | **TP** | **TN** | **FP** | **FN** |
|---|---|---|---|---|
| RFC | 2565 | 307 | 336 | 92 |
| KNN | 2488 | 64 | 579 | 169 |
| LR | 2657 | 0 | 643 | 0 |

Table 2: Confusion matrix values

The above table shows the confusion matrix of the algorithm we implement. the best result we achieve from the random forest algorithm because it has least number of false positive and false negative. So, we have good accuracy from these results.

## Model Results

Three models are implemented K-nearest neighbor classifier,Random forest classifier and Logistic regression classifier.

### 1.K-nearest neighbor classifier

KNN (K-Nearest Neighbor) is a simple supervised classification algorithm we can use to assign a class to new data point.Imported from the SK learn library

```
========== K-Nearest Neighbour Classifier ==========

Accuracy on validation set: 77.3333

Output:
              precision    recall  f1-score   support

           0       0.94      0.81      0.87      3067
           1       0.10      0.27      0.15       233

    accuracy                           0.77      3300
   macro avg       0.52      0.54      0.51      3300
weighted avg       0.88      0.77      0.82      3300

Confusion Matrix :
 [[2488  579]
 [ 169   64]]
```

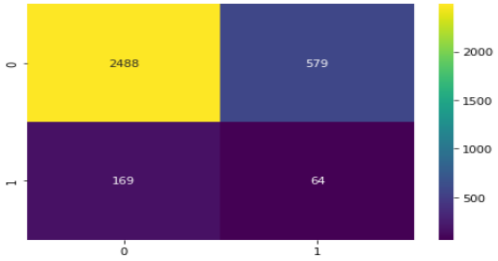


Figure 1: KNN classifier

## 2. Logistic regression classifier

The Second classifier, logistic regression, models the dependent variable using a logistic function.When the data sets are linearly separable, it works well.

```
========== logistic Regression ==========

Accuracy on validation set: 80.5152

Output:
              precision    recall  f1-score   support

           0       1.00      0.81      0.89      3300
           1       0.00      0.00      0.00         0

    accuracy                           0.81      3300
   macro avg       0.50      0.40      0.45      3300
weighted avg       1.00      0.81      0.89      3300

Confusion Matrix :
 [[2657  643]
 [   0    0]]
```

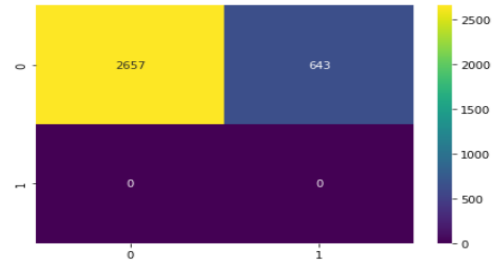


Figure 2: LR classifier

## 3. Random forest Classifier

The Sklearn package also allows for the import of the random forest classifier. Like the classifiers,it is a supervised learning method that may be used for both classification and regression. The algorithm is the most adaptable and user-friendly.

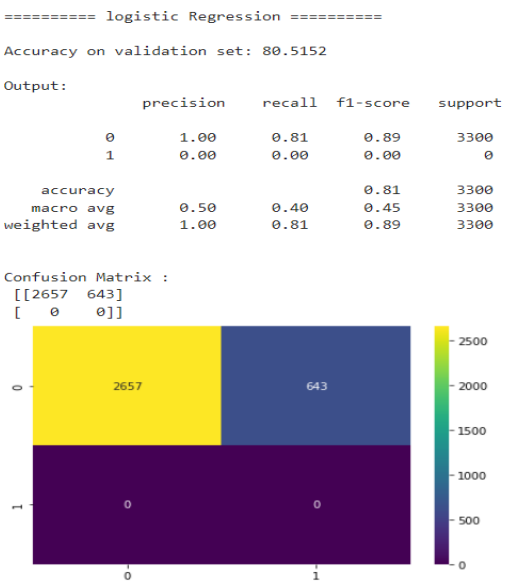


Figure 3: RFC classifier

## Classifiers Output report

In the classification report we get the results of the accuracy's, precision and recall score of the classifiers are used to predict the customer churn. After implementing the algorithms on the data sets we have following results

| Classifiers | Accuracy | Precision(0,1) | Recall(0,1) | F1-Score(0,1) |
|---|---|---|---|---|
| KNN | 77.33 | (0.94,0.10) | (0.81,0.27) | (0.87,0.15) |
| LR | 80.51 | (1.0,0.0) | (0.81,0.0) | (0.89,0.0) |
| RFC | 87.0 | (0.97,0.47) | (0.88,0.77) | (0.92,0.59) |

Table 3: Classification Report

According to this Classification report, the best model is **Random Forest Classifier**

# Conclusion

Customer churn analysis has become a major concern.The model developed will help banks identify clients who are likely to be churners and develop appropriate marketing actions to retain their valuable clients.The supervised machine learning algorithm is implemented to predict the customer churn prediction.Random forest classifier with the 87 percent accuracy is best among the two other classifiers of supervised machine learning algorithms that is implemented to predict the result of customer churn in banking system. The precision score and recall score is high among the two other models implemented. With the 87 percent random forest classifier gain the higher weighed average churn in the model.In order to properly comprehend the facts, we must conduct more research and obtain context.

# Bibliography

[1] Pierre Carbonnelle. PYPL PopularitY of Programming Language, kuddos, 2022.

[2] tkhan kiit. Different Types of Encoding, ai ml analytics, 2022.