**BANK CUSTOMER CHURN PREDICTION**

**1. INTRODUCTION**

In an era where the markets have become mature, new technologies have been on the rise, and change in demand has caused the uprise of some of the finest companies led by visionary leaders not just in the IT industry but in a few other major sectors such as banking and telecom. In such scenarios, it becomes fundamental for almost every organization in any business sector to manage relationships with their customer base in order to maintain the levels of revenue and also the standards of business for the customers to rely on. In business studies, this concept is known as “Customer Relationship Management (CRM),” and it speaks volumes about customer satisfaction with respect to the business. Organizations that strictly follow the core concepts of CRM nearly always improve their customer retention power. That is, they manage to maintain a relationship with the acquired customer, and hence the probability of the customer leaving their organization is significantly reduced.

Customer churn, also known as customer attrition, is the term coined for the probability of whether an existing customer continues his/her transactions with the organization or not. The probability factor of this parameter depends on numerous other factors in various industries like banking, telecom, and a few other sectors. In intense market scenarios that are prevailing today in every sector, it becomes important for organizations to keep track of customer churn and the various reasons causing the customer to stop their transactions with the company. Almost every organization is well-versed in the concept that retention of existing customers will save a lot of money, as trying to acquire new customers will cost five to six times the cost of retaining an existing customer. Therefore, each organization out in the market started to understand and analyze the various factors that might cause a customer or client to leave the organization’s business. In fact, the company has started to roll out some special gifts and offers for customers who are on the verge of leaving the company’s business in order to keep them engaged with the organization.

This project report deals with customer churn in the banking sector and highlights the various factors that affect the attrition of customers from the bank. It also sheds some light on how a bank can predict customer churn rate by using well-known and popularly used machine learning models.

## **Definition of Churn in Banking Industry**

The term churn, also known as attrition, turnover or defection, is a widely known concept in almost every industry. (Rosa, 2019)Explained the importance of customer retention, where a better offer from other companies influences customers' behavior in today's high competition market. Therefore, it is vital that firms closely monitor their clients and identify any signs of customer turnover. The broad or generally accepted definition of churn refers to the departure of customers from an organization's business activities, which results in a loss to the company. (Eichinger et al., 2006)Customer churn was defined as the departure of a customer from the company to a competitor.(Rosa, 2019) further stated that customer churn is when a customer withdraws his or her use of an enterprise's goods and services in favor of those of its competitors.

**2. RELATED WORK:**

Some of the major references related Bank Customer Churn Prediction are:

* (Bilal Zorić, 2016) A new framework for analyzing and anticipating customer turnover in a Portuguese commercial bank has been proposed in the paper. The purpose of the study is to get a set of prediction variables through analysis of past churners' behavior and patterns using data mining techniques. To train the predictive models, these variables are applied in conjunction with networks of neurons. In a sample of existing customers who are at risk for churning, the models' performance shall then be evaluated.
* (Chayjan et al., 2021) The article examines the use of artificial intelligence techniques for predicting a customer's churn in banking. To train and test machine learning models, in particular decision trees as well as logistic regressions, the study is based on data from a Portuguese retail bank. The accuracy and predictability of these models shall be considered when evaluating their performance. These results demonstrate that machine Learning techniques can accurately estimate the customer's churn rate, while decision trees are outpacing logistic regression models. The importance of customer loyalty in the banking industry and the potential for machine learning techniques to develop effective marketing retention strategies are highlighted. The study found that banks may benefit from machine learning models to identify at risk clients and undertake targeted retention campaigns to reduce customer turnover.
* (Eichinger et al., 2006) This article looks at the use of sequence mining methods for predicting customer behavior in banking. The study is based on data from a German retail bank and uses a sequential pattern mining algorithm to identify common customer behavior patterns that can predict future churn. In addition, the accuracy of the prediction model is examined in detail by examining the impact of various sequence mining parameters. Results indicate that it is possible to accurately predict customer behavior with a high degree of precision and recall rates using sequence mining techniques. The article emphasizes the potential that sequence mining techniques can provide in identifying at risk customers and establishing focused retention strategies. This study has concluded that sequence mining techniques can be an important tool for improving customer retention in the banking sector.
* (M.A.H. et al., n.d.) In order to make it more transparent for the management of customer relationships and CRMs, The article proposes a combination approach in which rules are extracted from Support VectorMachineSVM. The proposed approach involves three phases: (i) SVM-recursive feature elimination (SVM-RFE) to reduce the feature set; (ii) extracting support vectors from the reduced feature dataset to obtain SVM model; (iii) generating rules using Naive Bayes Tree (NBTree) in the final phase.
* (Tsai & Lu, 2008) A framework for churn prediction, based on artificial neural networks and logistic regression models, is proposed in the article. The study uses a Taiwanese telecommunications company's data, with an objective of examining the accuracy and effectiveness of these models. The authors have compared the performance of different models and methods of selecting features, and have found that the neural network model with the principal component analysis PCA outperforms the others. The Article goes on to state that telecommunications operators could gain valuable insight into developing targeted marketing campaigns for retaining customers by proposing such a framework.
* (Xie et al., 2019) The proposed approach consists of two main steps. First, based on the customer's transaction behavior, a clustering algorithm is used to segment the customer into different groups. Secondly, for each cluster a Decision Tree is set up with the aim of identifying the main factors contributing to churn prediction. Accuracy, precision, recall and F to measure metrics shall be used to evaluate the performance of the proposed approach.

**3. Data Description:**

There are several methods for gathering data. The method of collecting data includes data labeling, which is required in all supervised learning applications. Because manual labeling is costly, some scalable techniques based on semi-supervised learning, crowdsourcing, and weak supervision have been proposed. Finally, instead of training from scratch, one can improve the quality of existing data or use transfer learning to reuse existing models.

The dataset used in this research project consists of

**CustomerId:** Unique identifier for each customer.

**Surname:** Customer's last name.

**CreditScore:** The credit score of the customer.

**Geography:** The country of the customer.

**Gender:** Customer's gender.

**Age:** The age of the customer.

**Tenure:** Number of years the customer has been with the bank.

**Balance:** The current balance in the customer's account.

**NumOfProducts:** Number of products held by the customer, such as insurance loans and credit cards.

**HasCrCard:** Whether the customer has a credit card with the bank or not.

**IsActiveMember:** Whether the customer is an active member of the bank or not.

**Estimated Salary:** The estimated salary of the customer.

**Exited:** Whether the customer has churned (i.e., closed their account) or not.

**4. Methodology:**

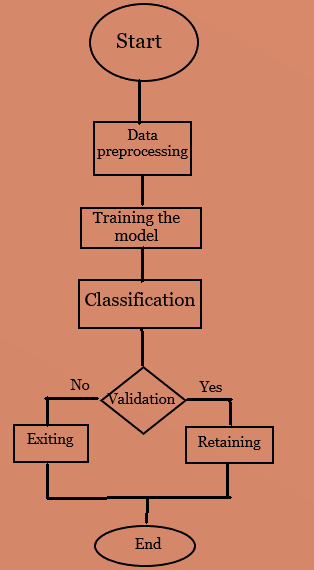


Fig 1: Flow Chart

**4.1 Data Collection:**  
The Data comprehension period deals with assembling and exploring the data to get insights into the whole data. The dataset can be numeric, categorical, etc. It’s gathering information from various sources. The data provides us with information regarding past events, and based on that, we can find recurring patterns. Patterns can be used for building models with the help of various algorithms for future prediction. The coverage of data collection is based on recent surveys and past data of the customer. Data lies across several servers, and the data should be gathered in one place for easy access. Data is present in various formats; therefore, it should be changed to a particular format for data collection. JSON contains data from chat servers, and data from business applications is generally tabular. If both kinds of data need to be used, it should be converted to JSON or whole data to CSV or xlsx. Data is also present in HTML text. It should be cleaned for further useful purposes.

There are some methods for data collection. If the number of datasets is increasing rapidly, searching for the right ones becomes challenging. Acquisition of appropriate data for training the models is a problem to overcome as machine learning is widely used, and we don’t have enough labeled data. The next method involves data labeling, which is necessary for all supervised learning applications. As manual labeling is expensive, some scalable techniques are proposed using semi-supervised learning, crowdsourcing, and weak supervision. At last, one can also improve the quality of existing data or use transfer learning to re-use existing models instead of training from scratch. The present project dataset is taken from Kaggle, which consists of 1000 records and 14 features.

**4.2 Data Preprocessing:**

Data is preprocessed so that incorrect, inaccurate, and irrelevant data is modified, replaced, or deleted whenever needed. It is considered as the fundamental element in machine learning. When it comes to the real-time scenario, the data may be inconsistent, or it has missing values. If the data is corrupted, then it may obstruct the process the output will be inaccurate. There are some examples of data cleaning. Person X is a general manager of a company. The company collects data on people visiting the place and buying the product from the company. As the information is available to the company, they will have insights into which products people are more attracted to, and the company will try to increase the production of products. One drawback is that the complete analysis will go wrong if any values of company data need to be added. The model is based on how the data is clean, and based on that, processing is done. Data cleaning involves imputation or handling because some of the chosen algorithms can handle missing data.

Feature selection is one of the steps of data pre-processing. Fundamental notions in machine learning contribute considerably to the model's performance. Data cleaning and feature selection are essential steps for model designing. In feature selection, features are selected manually or automatically, contributing to the prediction. It is essential to have relevant features, as irrelevant features decrease the model's accuracy, which creates a massive blunder in the output. It is applied to the dataset with many dependent variables in the given dataset. It trains the model faster and reduces the complexity. There are three benefits of selecting features before the modeling is done: reduced overfitting, improved accuracy, and reduced training time.

**1. Reduces Overfitting:** As there is less redundant data, we have less opportunity to make decisions based on noise.

**2. Improves accuracy:** As there is no misleading data, it automatically increases the accuracy.

**3. Reduces Training time:** Algorithm complexity is reduced due to few data points.

**4.3 Dependent and Independent variables:**

For the purpose of the study, the data is split into Dependent variable (variable of interest) and Independent variables. “Exited” which tells Whether the customer has churned (i.e., closed their account) or not. And “CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary” which us in the prediction process.

Then the data is split into train and test for training the model and validating the predicted values.

1. Splitting the Predictors(X) and the Target variable(y).

2. Splitting the X and y further. That is the test-train split. Now, we have 4 variables. X\_train, X\_test ; y\_train, y\_test. The train data is 80% of the complete data whereas the test data is 20% of the complete data sampled randomly.

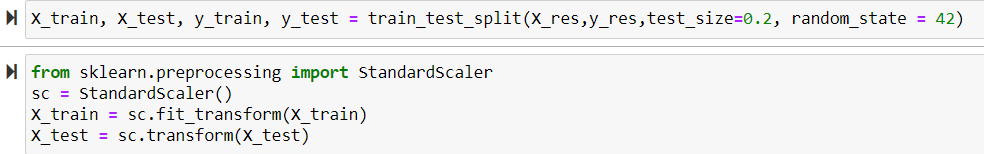
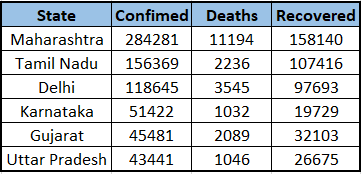
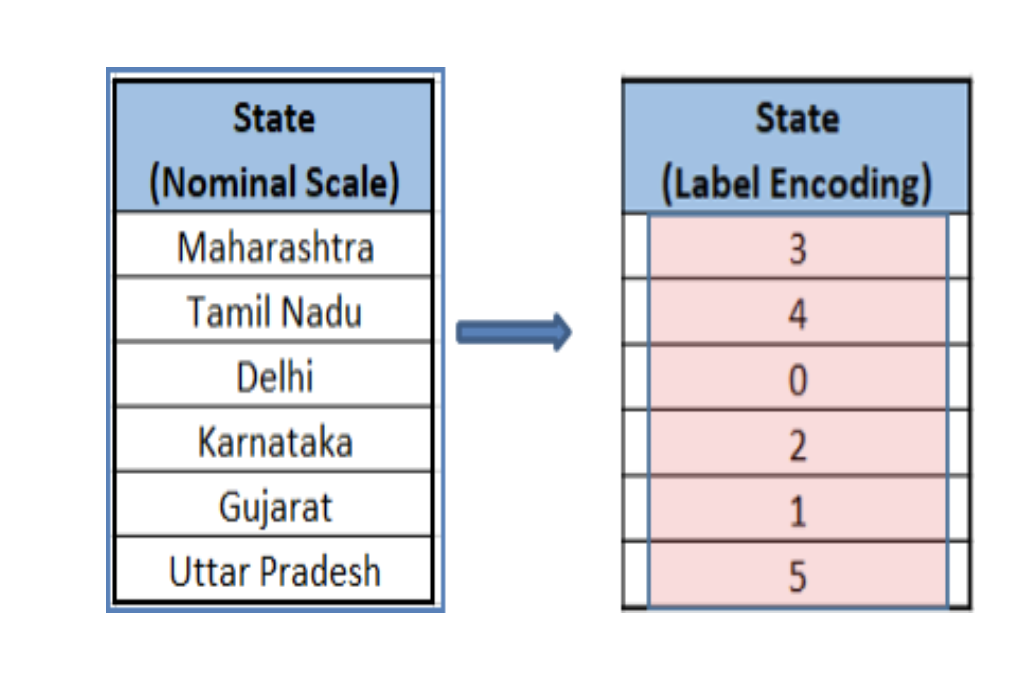
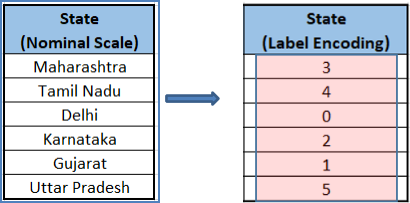


Fig 2: Splitting data into training and testing

**4.4 ENCODING:**

According to the machine learning models, the input and output variables must be numeric. It is nothing, but if the data consists of categorical data, we use the encoding process to encode it to numeric form before evaluating the model. Encoding is the must-pre-processing stride while working with categorical data using machine learning algorithms. Data is of two types quantitative and qualitative. Quantitative data are assigned with numbers, and things used to measure it may be dimension, temperature, humidity, area, volume, etc. Qualitative data is assigned with characteristics that cannot be measured, such as smells, tastes, textures, attractiveness, and color. **Fig 3: Table of Covid-19 cases**

****

**Fig 4: Encoding**

From the above table state column consist of categorical variables, we use label encoding here to convert to numerical data. As there are several states, each state is assigned with numeric value starting from “ 0 ” based on the alphabetical order. The process of assigning numerical value to categorical data is called label encoding. The below table show the label encoding of state column. In our project we used Label Encoder class usingscikit-learn library.

**4.5 Model Builing:**

Before the model is constructed there are a lot of pre processing steps that need to be done in order for the model to be properly constructed.

The Machine learning algorithms used in model construction are:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. AdaBoost Classifier
5. Gradient Boosting Classifier.

**1. Logistic Regression:** Logistic regression is a basic classification technique. It belongs to the linear classifiers group and is similar to polynomial and linear regression. *Logistic regression* is a simple and rapid method for predicting results and is ideal for you to use. Although it is primarily a binary classification approach, it may also be used to address multiclass issues such as this concept.

**2. Decision Tree Classifier:** A decision tree is a popular machine learning method. Internal decision-making logic is shared, unavailable in black-box algorithms like Neural Networks. It takes less time to train than the neural network approach. The number of records and attributes in the data supplied can impact the temporal complexity of decision trees. The decision tree is a non-parametric or distribution-free classifier that does not rely on probability distribution assumptions. Decision trees can handle large amounts of data and yet make accurate predictions.

**3. Random Forest Classifier:** Random-forest is a learning method that is supervised. This classifier may be used for both classification and regression. Random forest is also very adaptable and simple to utilize. Trees make up a forest. It is considered that the greater the number of trees available, the stronger a forest becomes. Random forests create decision trees based on randomly selected data samples, obtain forecasts from each tree, and use the voting concept to select the best potential answer. The random forest also provides a fairly decent measure of feature value.

**4. AdaBoost Classifier Boosting:** Is algorithms have recently become more popular and well-known among data science and machine learning enthusiasts. These enthusiasts want to use boosting algorithms to win tournaments since they provide excellent accuracy. Boosting algorithms combine many low-accuracy (or weak) models to produce a high-accuracy (or robust) model. The data science projects provide a platform for learning, researching, and developing practical solutions to various corporate and government problems.

**5. Gradient Boosting:**

Gradient boosting classifiers are a collection of machine learning algorithms that combine a number of weak learning models into a powerful prediction model. When applying gradient boosting, decision trees are frequently employed. Gradient boosting models are gaining popularity as a result of their efficiency in categorising compound datasets, and have recently been utilised by enthusiasts to win a number of Kaggle data science challenges. The idea behind "gradient boosting" is to take a bad hypothesis or a bad learning algorithm and make a series of changes to it that will improve the hypothesis's strength.

**5. Results:**

This project intended to form a basis to address churn in a bank presently not utilizing its robust database and analytic applications to solve this critical problem. The initial step consisted of gathering a dataset of customers for a certain period who would probably be churners during that period. The aim is to keep track of these customers' behavior during the time given in the dataset, which would possibly represent the risk of being churned shortly. As such, the choice of variables depended completely on using dummy variables that depicted a reduction in the level of association with the bank, meaning customers owned very few financial goods at one point in time compared to some other point in time. The pre-processing task was carried out with the intent of outlier elimination and data transformation. It was finalized that the utmost significant variables to train the proposed predictive models were customer gender, age, credit score, geography, tenure, and account balance. The machine learning models used in this project are logistic regression, decision tree, random forest classifier, AdaBoost classifier, and Gradient boosting classifier. Each model was trained on the same dataset of 10,000 customer records and 14 features. Upon comparing the accuracy of these classifiers, the gradient boosting classifier faired well enough with an accuracy of almost 86%. The Logistics Regression, Decision Tree classifier, Random Forest, AdaBoost and Gradient Boostinghave an accuracy of 77%,79%,85%,82% and 84% respectively.

In the end, this project served its purpose to put forth a reliable and effective alternative to predict and have a timely check on customer churn behavior as opposed to the existing reactive approach employed by the banking sector, which consists of creating marketing strategies aimed at regaining past customers who were churned out of the bank. In light of the positive results achieved in this work, the current procedure carried out throughout this project could prove to be a beneficial tool to estimate churn in a company that has yet to make full use of the Business Intelligence tools at its disposal to solve this issue.