A

Project-2 Report

On

**BANK CUSTOMER CHURN PREDICTION USING MACHINE LEARNING**

Submitted in partial fulfilment of the requirements

for the award of the degree of

**Bachelor of Technology**

in

**Electronics & Computer Engineering (ECM)**

By

|  |
| --- |
| **G. SANDILYA SRINIVAS (17311A19E6)** |
| **S. LIKHIT KUMAR (17311A19E7)** |
| **J. PAVAN GANESH (17311A19G8)** |

**Under the Guidance of**

**Mrs. M. Shailaja  
Associate Professor  
Internal Guide**

#### 

**Department of Electronics & Computer Engineering**

**Sreenidhi Institute of Science & Technology (Autonomous)**

**JUNE 2021**

**DEPARTMENT OF ELECTRONICS & COMPUTER ENGINEERING**

**SREENIDHI INSTITUTE OF SCIENCE & TECHNOLOGY (AUTONOMOUS)**

****

**CERTIFICATE**

This is to certify that the project entitled **“BANK CUSTOMER CHURN PREDICTION USING MACHINE LEARNING”**, submitted by **G. SANDILYA SRINIVAS** (**17311A19E6)**, **S. LIKHIT KUMAR (17311A19E7)**, **J. PAVAN GANESH (17311A19G8)** towards partial fulfillment for the award of Bachelor’s Degree in Electronics & Computer Engineering from Sreenidhi Institute of Science & Technology, Ghatkesar, Hyderabad, is a record of bonafide work done by them. The results embodied in the work are not submitted to any other University or Institute for award of any degree or diploma.

**Internal Guide Project Co-ordinator Project Co-ordinator**

**Mrs. M. Shailaja Mrs. N. Swapna Dr. M. Nanda Kumar**

**Associate Professor Associate Professor Assistant Professor**

**Head of Department, ECM External Examiner**

**Dr. D. Mohan**

**DECLARATION**

This is to certify that the work reported in the project entitled **“Bank Customer Churn Prediction Using Machine Learning”** is a record work done by us in the **Department of Electronics and Computer Engineering, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar.**

The report is based on the project work done entirely by us and not copied from any other source.

G. SANDILYA SRINIVAS (17311A19E6)

S. LIKHIT KUMAR (17311A19E7)

J. PAVAN GANESH (17311A19G8)

**ACKNOWLEDGMENT**

We convey our sincere thanks to all the faculties of ECM department, Sreenidhi Institute of Science and Technology, for their continuous help, co-operation, and support to complete this project.

We convey our sincere thanks and gratitude to our internal guide **Mrs. M. Shailaja**, Associate Professor and our project co-ordinators **Mrs. N. Swapna,** Associate Professor and **Dr. M. Nanda Kumar**, Assistant Professor, ECM Department, Sreenidhi Institute of Science and Technology, Ghatkesar for initiating and guiding us all through the project.

We are very thankful to **Dr. D. Mohan**, Head of ECM Department, Sreenidhi Institute of Science and Technology, Ghatkesar for providing an initiative to this project and giving valuable timely suggestions over our project work and for their kind cooperation in the completion of the project.

We convey our sincere thanks to **Dr. T. Ch. Siva Reddy**, Principal and **Dr.** **P. Narsimha Reddy,** Director, Sreenidhi Institute of Science and Technology, Ghatkesar for providing resources to complete this project.

Finally, we extend our sense of gratitude to almighty, our parents, all our friends, teaching and non-teaching staff, who directly or indirectly helped us in this endeavor.

**ABSTRACT**

Customer relations is of utmost importance for industries that directly provide goods and services to the people. The goodwill of the customers is what keeps the company up and running in many of the sectors such as telecom, banking, educational institutions, etc,. Various aspects like competitors, novel and innovative business models and enhanced services are increasing the cost of customer acquisition. In such a fast set-up, service and goods providers have realized the importance of retaining the on-hand customers. Thus, establishing and maintaining good customer relations is the key for success of a company in today’s modern world of business.

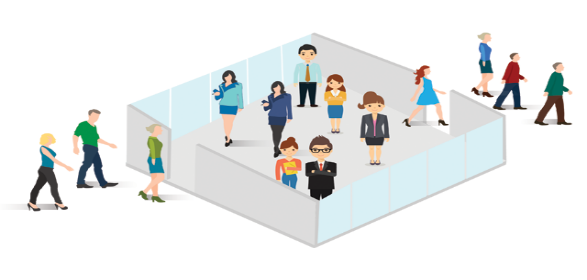
Banking is one of the highly competitive sectors where customer relations is of utmost importance for any bank. Each customer is considered as a customer for life by the banks. Home Loans are typically the longest relationship with any customer for the bank. Statistics and some of the top leading banks prove that customers are more sophisticated, and they need some offers and boosts to continue their relationship with the organization. The term “Customer Churn” refers to the state in which the customer or the subscriber stops involving in business transactions with a company or a service provider. To deal with the concept of customer churn, many organizations both big and small use machine learning to predict at what rate the customer churn might happen and based on the customer churn rate, the company comes up with a scheme or offer to hold on to their on-hand customers.

We predict the customer churn rate, using Machine Learning models which will indicate whether a customer will leave the bank or not based on many factors, this in turn will help the bank in knowing which category of customers generally tend to leave the bank. Further the banks can bring in exciting offers so that it can retain its customers. In this predictive process popular models such as logistic regression, decision trees, random forest and other boosting techniques have to be used to achieve a decent level of accuracy, for the banks to rely on so that they can clearly predict which customer may leave next based on customer data available.

**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 lead by visionary leaders not just in the IT industry but in 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. The 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 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 few other sectors. In intense market scenarios that are prevailing today in every sector, it becomes important for the organizations to keep a track of customer churn and the various reasons causing the customer to stop their transactions with the company. Almost every organization is well-versed with the concept that retention of the existing customers will save a lot of money as trying to acquire new customers will cost five to six times the cost to retain an existing customer. Therefore, each organization out in the market started to understand and analyze the various factors that might be the cause for a customer or client to leave the organization’s business. In fact the company starts 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.



**Fig 1: Customer Churn**

Once importance of customers and customer churn was identified by various leading companies, they started to gather data on customer behavior like how often does a customer purchase a product,etc,. With this the collection, storage and managing of the customer data on a timely basis became really crucial for the companies. Thus, the decision making process shifted from being an event-driven approach to a data-driven approach. New technologies like Big Data, Cloud storage and machine learning supported the companies data gathering, processing and analysis phase. Therefore, the whole process shifted to statistical analysis as oppossed to predictive analysis.

This project report deals with the customer churn in a 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 the rate of customer churn by making use of well-known and popularly used machine learning models.

**2. LITERATURE SURVEY**

**2.1 INTRODUCTION**

Numerous researches have been carried out pertaining to churn prediction by making use of various statistical and machine learning algorithm over the last decade. This chapter deals with the recent and significant publications on prediction of customer churn in the banking sector in recent times.

**2.2 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. As Oyeniyi and Adeyemo (2015) pointed out, “churning is an important problem that has been studied and researched across several areas of interest, such as telecom, insurance, and healthcare. Other fields where the customer churn has been analyzed include online social network churn analysis, and the retail banking industries”. The broad or the generally accepted definition of churn refers to the exiting of customers from an organization’s business activities, thus causing some sort of loss to the company. Eichinger, Nauck, and Klawonn (2006) defined customer churn as when a customer is leaving our company for a competitor. This belief has been supported by Qiasi, Roozbehani, & Minaei-bidgoli (2002) who consider churn when a customer discontinues the use of an organization´s products and services in favor of a competitor´s products and services.

In the banking industry, the radius of the term, churn is wide and is being used in several areas of the business. When a customer abstains from using his/her credit card or on the whole doesn’t carry out any kind of transactions using their account, there are chances of him/her becoming the churned customer in the near future. Similarly, network/internet banking churn might happen if a customer doesn’t involve in an online banking transaction for a specified time window. The bank can look into the account balance of a customer i.e savings, securities and few other kinds of products and services and can define churn in two different ways i.e voluntary and involuntary churn. Voluntary churn deals with those kinds of customers who leave the bank due to being unsatisfied with the products and services offered by the bank. Whereas, involuntarily churned are those customers who are asked to leave by the bank due to various reasons like not maintaining the required balance and numerous other reasons.

The current analysis will be particularly focused towards tracking the behavioral history of customers who have been churned in the past within a specific period of time in order to spot certain patterns that might indicate that a customer is in risk of being churned anytime soon. To gather data about the past churned customers, we’ll have to do some ground work to determine which parameters decide whether a customer is a churner or not, as most of the companies don’t erase their past customer details and save them for further analysis and learning the pattern in which a customer interacts with their business. A customer is termed to be an involuntary churner whenever the relationship is weakned due to poor involvement. Poor involvement can be attributed to one of the following reasons:

* The person is not the primary account holder.
* There is no transaction made by the customer in the past three to four months using his/her account.
* The account balance of the customer is lower than 50,000/- with no credit card, credit loans and financial assets.

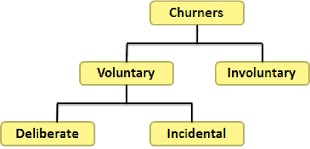
By taking this details into consideration, it is crucial to determine the kind of customers that will be taken into consideration in this project. The aforementioned conditions of poor involvement immediately pose some strenuous challenges to overcome with regards to keeping track of involuntary churn. The fact that a customer is showing signs of being a churner is whenever his balance drops below Rs. 50,000/- states that there could be daily switch between being a churner and non-churner. These kinds of situations will not be addressed in the present study, hence voluntarily churned customers will be the main target focus of this project.

**2.3 OVERVIEW OF CHURN ON RETAIL BANKING INDUSTRY**

This sections aims to deliver a brief description on prevailing state of retail banking industry across the globe and also give an insight into the future events that will impact this industry and in due course have an impact on customer churn in the years to come. After a significant financial crisis a few decades ago, the banking industry came into existence to cater the needs of those difficult times. Since its inception, it has undergone numerous changes in the way it functions and its operating structure, on the whole, has gone through a lot of shaping and structuring. The shift towards digital methods in response to market demand has resulted in the closure of a huge number of branches throughout the country, with the inevitable layoff of thousands of employees. Such events have had an impact on customer satisfaction, as the number of employees to serve the customers at their offline branches reduced significantly.

It is important to note that reigniting pricing competition between financial organizations to win over credit operations, which in turn have an effect on churn as customers are less hesitant to move their business elsewhere if conferred with better pricing proposals.

In addition, banks have countered the effects of lower financial margins by increasing commissions across the various sectors of the business, which directly have an impact on customer satisfaction and thus churn levels. According to an afresh questionnaire conducted by the bank that consisted in contacting past churners to understand the underlying causes for them ceasing their business with the company, results disclosed that churners usually fall into four different classes:



**Fig 2: Types Of Churners**

**3. ANALYSIS**

**3.1 EXISTING SYSTEM**

Many approaches were explored and used in predicting the churn in various industries like telecom, banking,etc,. Most of the approaches have utilized the power of machine learning algorithms and data mining techniques. The significant amount of related work focused on applying only one method of data mining to extract data from appropriate sources, and the others focused on several machine learning algorithms to predict churn rate.

Supervised machine learning approaches have been employed in customer churn prediction problems back in the day with SVM-POLY using AdaBoost as the best possible model (Vafeiadis, Diamantaras, Chatzisavvas & Sarigiannidis, 2015). The most usual approaches applied for forecasting customer churn rate are Decision tree, Random forest, Multilayer perceptron, and SVM.

In existing fact-finding of customer churn prediction problem in almost every possible industry, the researcher Guo-en, X., has employed SVM model for the telecom industry as it can solve the non-linearity, high dimension, and local minimization problems. The model prediction relies on the data structure and the prevailing condition.

**\**Approaches that are widely used to forecast customer churn are neural networks, support vector machines and logistic regression models. Data mining research works suggests that machine learning techniques, such as neural networks should be employed for non-parametric datasets because they often outclass conventional statistical techniques such as linear and quadratic discriminant analysis approaches (Zoric, 2016).

Logistic Regression is a kind of probability statistical classification model primarily used for classification problems (Nie, Rowe, Zhang, Tian & Shi, 2011). This approach can work wonders with a different blend of variables and can assist in predicting the customer churn with higher accuracy rate.

Random Forest is an ensemble learning method for classification, regression problems and employs the bagging approach to generate the results. The default hyperparameters of Random Forest gives good outcomes and it is exceptional at averting overfitting problem (Pretorius, Bierman & Steel, 2016).

**3.2 PROPOSED SYSTEM**

This project proposes the use of machine learning and data mining techniques to differentiate the customer who are in the risk of being churned and customers who are happy and satisfied with the products and services of the bank. The algorithms and various technologies employed to predict the customer churn for a bank are discussed in the following sections in detail. This project makes use of several data pre-processing steps and machine learning models like logistic regression, random forest,etc,. and compares each of their prediction power by taking into consideration the accuracy of each machine learning models and selecting the best model with highest accuracy, to predict the customer churn.

A front-end application which is a simple webpage is provided to the end-user to enter the details of a customer to check whether he/she is at a risk of being churned or is happy and satisfied with the products and services of the bank.

**3.3 PROBLEM STATEMENT**

The aim of this project is to convey the details of the customer who are at the risk of being churned from a bank, for the products and services offered by a competitor. This project aims to predict the customer attrition rate for a bank. Thus, providing them with some early pointers that a customer is about to leave their business for a competitor’s products and services. By knowing this at an early stage, the bank can engage the customer with some kind of special gifts and schemes to continue their relationship with the bank.

* 1. **SOFTWARE REQUIREMENT SPECIFICATION**

The following are the software requirements for building this project and to get the desired level of results:

* **Python Programming:** Python is a high-level, general-purpose and a very well favoured programming language. Python programming language (latest Python 3) is being used in automation, robotics, web development, Machine Learning applications, along with all groundbreaking technology in IT Industry.
* **Jupyter Notebook:** The Jupyter Notebook is an open source web application that you can utilize to design and distribute documents that contain live code, equations, visualizations, and text. Jupyter Notebook is conserved by the various members of Project Jupyter.
* **Visual Studio Code IDE:** Visual Studio Code is a lightweight but robust source code editor which runs on your desktop or laptop and is accessible for Windows, macOS and Linux. It has built-in foundation for JavaScript, TypeScript and Node.js and has a rich environment of extensions for other languages (such as C++, C#, Java, Python, PHP, Go) and runtimes (such as .NET and Unity).
  1. **REQUIRED PYTHON LIBRARIES**

**3.5.1 PANDAS:**

Pandas is a software component developed for the python language for data manipulation and analysis. In particular, it provides data structures and operations for handling numerical tables and time series data. It is a free software released under the three-clause BSD license. The name is acquired from the term "panel data", an econometrics term for data sources that include findings over multiple time periods for the same individuals. Its name is a play on the expression "Python data analysis" itself. Wes Mckinney started developing what would be called pandas at AQR capital while he was a scientist there from 2007 to 2010.

Various intresting and fascinating features of pandas library are as follows:

* DataFrame objects can be created with indexing.
* Tool for reading and writing data from some of the widely used storage locations.
* Pivoting and reshaping of data sets acquired.
* Data filtration capability.
* The library is highly stable with majority of its code written in Cython or C programming language.

****

**Fig 3: Pandas**

**3.5.2 NUMPY**

**NumPy** is a library for the Python programming language, providing support for huge, multi-dimensional arrays and matrices, along with a large set of high-level mathematical functions to utilize on these arrays. The ancestor of NumPy, Numeric, was originally developed by Jim Hugunin with contributions from few other developers. In 2005, Travis Oliphant developed NumPy by embedding features of the competing Numarray into Numeric, with vast modifications. NumPy is open-source software and has numerous contributors.

NumPy focuses on the CPython reference implementation of Python, which is a non-optimizing bytecode interpreter. Mathematical Algorithms and functions designed for this version of Python often run much slower than compiled identicals. NumPy addresses the slowness issue partly by providing multi-dimensional arrays and functions and operators that work efficiently on arrays; using these demands for rewriting some code, in most of the cases inner loops, using NumPy.

Using NumPy in Python gives ability to it to be comparable to MATLAB since they are both interpreted, and they both permit the users in writing code in short period of time as long as most operations work on arrays or matrices instead of scalars. In contrast, MATLAB boasts a huge number of additional tools, remarkably Simulink, whereas NumPy is intrinsically integrated with Python, a more user-friendly and complete programming language. Moreover, supporting Python packages are available; SciPy is a library that gives much more MATLAB-like functionality and Matplotlib is a plotting package that provides MATLAB-like plotting functionality. Internally, both MATLAB and NumPy depend on BLAS and LAPACK for methodical linear algebra calculations.

Python bindings of the vastly used computer vision library OpenCV uses NumPy arrays to capture and operate on data. Since pictures with multiple channels are simply depicted as three-dimensional arrays, indexing, slicing or masking with other arrays are very structured ways to access particular pixels of a picture. The NumPy array as ubiquitous data structure in OpenCV for pictures, drawn out feature points, filter kernels and many more majorly simplify the programming workflow and debugging.



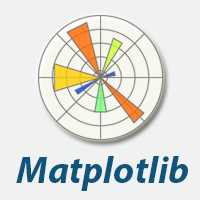
**Fig 4: NumPy**

**3.5.3 MATPLOTLIB**

Matplotlib is a graphical plotting library for the Python programming language and its arithmetical mathematics add-on NumPy. It gives an object-oriented API for implanting plots into applications using all-purpose GUI tools like Tkinter, wxPython, Qt, or GTK. There is also a methodical "pylab" interface based on a state machine (like OpenGL), created to nearly match that of MATLAB, though its purpose is demoralized. SciPy utilizes Matplotlib.

Matplotlib in the first place was developed by John D. Hunter. Since then it has an agile development community and is given out under a BSD-style license. Before John Hunter's demise in August 2012, Michael Droettboom was proposed as matplotlib's lead developer and was further accompanied by Thomas Caswell.Several tools are available that enhance the functionality of the matplotlib. Few are independent downloads and few come along with the matplotlib source and have external dependencies:

* Basemap: Helps in plotting maps with numerous map projections, coastlines and polotical boundaries.
* Cartopy: a mapping component that supports object-oriented map projection definitions.
* Excel tool: Supports exchange of data between excel and some other source.
* Mplot3D: Gives the ability to plot 3-D graphs.



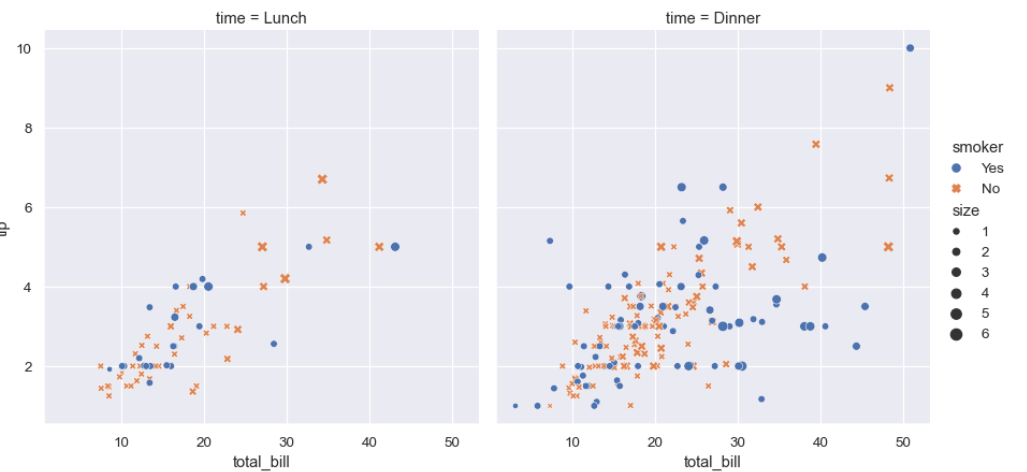
**Fig 5: Matplotlib**

**3.5.4 SEABORN**

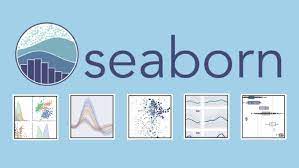
Seaborn is a independent library for designing statistical graphics in Python. It has its foundation on top of matplotlib and can work in close relation with pandas data structures.

Seaborn supports you to explore and understand your data. Its graph plotting functions operate on dataframes and arrays consisting of the whole datasets and internally perform the required semantic correlation and statistical aggregation to generate descrptive plots. Its dataset-oriented, explanatory API lets you concentrate on what the various aspects of your plots mean, rather than on the particulars of how to plot them.

Here is an example of how seaborn does what it does:



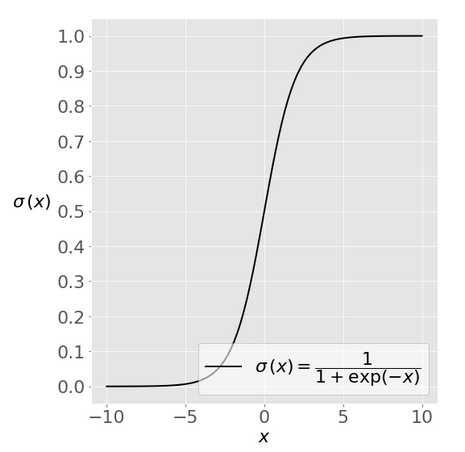
**Fig 6: Seaborn Plot Example**



**Fig 7: Seaborn**

**3.5.5 LOGISTIC REGRESSION**

Supervised machine learning algorithms to interpret models that understand relationships among data. A field of supervised machine learning that attempts to forecast which group or category some unit belongs to, on the basis of its features is called Classification. For example, one might inspect the employees of some organization and try to define a dependence on the features or variables, such as the educational background, experience in the current position, age, salary, probability for being promoted, and so on. The group of data related to a single employee is one observation. You can utilize the classification concept in many areas of science and technology. For example, text classification algorithms are employed to categorize emails into legitimate and spam, as well as comments into positive and negative ones. Additional examples involve medical applications, biological classification, credit scoring, and many more.Logistic regression is a rudimentary classification approach. It falls under the category of linear classifiers and is somewhat identical to polynomial and [**linear regression**](https://realpython.com/linear-regression-in-python/). Logistic regression is quick and comparatively straightforward, and it is appropriate for you to predict the results. Although it’s necessarily a method for binary classification, multiclass problems can also be solved using this concept.



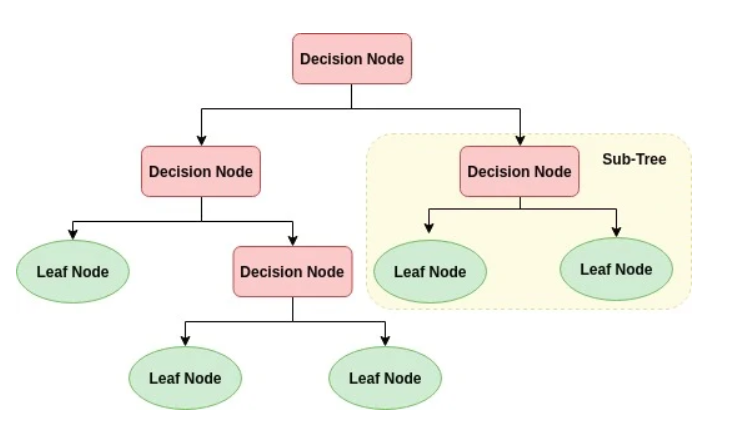
**Fig 8: Logistic Regression (Sigmoid Function)**

**3.5.6 DECISION TREE CLASSIFIER**

A decision tree is a flowchart kind of tree arrangement like a binary tree in data strucures and algorithms, where an internal node serves as a feature(or attribute), the branch denotes a decision rule that is to be satisfied, and each leaf node represents the final result. The foremost node in a decision tree is known as the root node. It ascertains to segregate on the basis of the attribute value. It splits the tree in recursive fashion call recursive partitioning. This flowchart-like structure aids you in decision making. It's visualization such as a flowchart diagram which easily resembles the human standards of thinking. This is why decision trees are simple to understand and interpret.

Decision Tree is a widely used ML algorithm. It shares internal decision-making logic, which is unavailable in the black box kind of algorithms like Neural Network. Its training period is less compared to the neural network algorithm. The time complexity of decision trees can be determined by the number of records and number of attributes in the data available. The decision tree is a distribution-free or non-parametric classifier, which do not rely upon the probability distribution presumptions. Decision trees can withstand high dimensional data providing good accuracy in its predictions.The fundamental idea behind any decision tree algorithm is as mentioned under:

1. Select the appropriate attribute using Attribute Selection Measures (ASM) to divide the records.
2. Consider that attribute as a decision node and split the dataset into small-scale subsets.
3. Start the tree construction by repeating this sequence of steps recursively for every child until one of the condition will suit:
   * All the tuples map to the same attribute values.
   * There are no more left behind attributes.
   * All the instances are covered.



**Fig 9: Decision Tree**

**3.5.7 RANDOM FOREST CLASSIFIER**

Random forest is a supervised learning algorithm. Both classification and regression can be performed by using this classifier. Random forest is also flexible and most easy to use. A forest consists of trees. It is assumed that the more trees that are available, the more powerful a forest turns out to be. Random forests designs decision trees on arbitrarily choosen data samples, gets forecast from every single tree and chooses the best possible solution utilizing the voting concept. A pretty good indicator of the feature importance is also provided by random forest.

There are numerous applications of random forest, such as recommendation engines, image classification and feature selection. It can be employed to split loyal loan applicants from fraud applicants, recognize fraudulent activities and predict diseases at an early stage. It is the foundation of the Boruta algorithm, which identifies important features from a dataset.

Random forest is basically an ensemble method (based on the divide-and-conquer approach) of decision trees established on a arbitrarily split dataset. This cluster of decision tree classifiers is popularly known as the forest. The independent decision trees are designed using an attribute selection parameters such as information gain, gain ratio, and Gini index for every attribute in the dataset. Each tree relies on an stand-alone arbitrary sample. In a classification issue, each tree votes and the most recurring class is elected as the final result. The average of all the tree outputs is taken into account as the final result in case of regression. It is elementary and more robust compared to the other non-linear classification algorithms.It functions in the following four steps:

1. Select arbitrary samples from a dataset available.
2. Generate a decision tree for each and every sample and get a forecast result from each decision tree.
3. Vote for each predicted outcome.
4. Choose the prediction outcome with the highest votes as the final prediction.

**Advantages:**

* Because of the numerous decision trees getting involved in the process, Random forests is contemplated as a highly accurate and robust classifier.
* It does not tolerate the overfitting problem. The main cause is that it considers the average of all the predictions, which nullifies the biases.
* Both classification and regression problems can be solved by using this algorithm.

**Disadvantages:**

* Random forests is slow in producing predictions because it has numerous decision trees. Whenever it makes a prediction, all the trees in the forest have to predict for the particular given input and then go ahead and voting for the best. This entire process is time-taking.
* The model is hard to understand compared to a decision tree, where you can make a decision with ease by following the branch in the tree.

**3.5.8 K- NEIGHBORS CLASSIFIER**

The K-nearest neighbors (KNN) algorithm is a kind of supervised machine learning algorithms. KNN is extremely simple to accomplish in its utmost basic form, and yet carries out quite composite classification activities. Since it doesn't have an exclusive training phase, it is a lazy learning algorithm. Rather, it uses the entire data for training while classifying a fresh data point or instance. KNN doesn't assume anything about the rudimentary data and thus is termed as a non-parametric learning algorithm. This is an exceptionally convenient feature as majority of the actual world data doesn't really adhere to any theoretical presumptions e.g. linear-separability, uniform distribution, etc. The foreknowledge behind the KNN algorithm is straightforward among all the supervised machine learning algorithms. It just determines the distance of a new data point to the remaining training data points. There are multiple kinds of distance e.g Euclidean or Manhattan etc. It then chooses the K-nearest data points, where K can be any possible integer. Finally the data point is put into the class to which the majority of the K data points belong.

**Pros and Cons of KNN Algorithm:**

**Pros:**

* It is exceptionally simple to implement.
* As mentioned earlier, it is lazy learning algorithm and therefore needs no training prior to predicting in real time scenarios. Thus making the KNN algorithm much faster compared to other algorithms that require training e.g SVM, linear regression, etc.
* New data can be added seamlessly as the algorithm does not require any prior data for training.

**Cons:**

* The KNN algorithm doesn't perform well with high dimensional data because with huge number of dimensions, it turns out that it is strenuous for the algorithm to determine the distance for each dimension.
* Finally, the KNN algorithm doesn't perform well with categorical attributes since it is strenuous to determine the distance between dimensions with categorical features.

**3.5.9 ADABOOST CLASSIFIER**

In present years, boosting algorithms have been gaining popularity and fame among data science or machine learning enthusiasts. These enthusiasts prefer using boosting algorithms in order to win some competitions as they give high accuracy. The Data science projects provide a platform for learning, exploring and providing feasible solutions for numerous business and government issues. Boosting algorithms get together numerous low accuracy(or weak) models to generate a high accuracy(or strong) models. It can be put to use in various sectors such as credit, insurance, marketing, and sales. Boosting algorithms like AdaBoost, Gradient Boosting, and XGBoost are prominently used machine learning algorithms to build data science projects. Yoav Freund and Robert Schapire in 1996 proposed one of the ensemble boosting classifier called Ada-boost or Adaptive Boosting. It puts together multiple classifiers to improve the accuracy of classifiers. AdaBoost is an iterative ensemble technique. AdaBoost classifier designs a robust classifier by cascading multiple badly performing classifiers so that you will get high accuracy powerful classifier. The fundamental concept that drives Adaboost is setting the weights of classifiers and training the sample data in each step such that it makes sure the accurate forecasts of unusual data points. Any machine learning algorithm can be employed as base classifier if it takes on weights on the training set. The following two conditions should be meet by AdaBoost:

1. Various weighted training examples are required to train the classifier interactively.
2. The model tries to generate an excellent fit for these examples by minimizing training error at each iterative step.

The following steps depicts the way in which the classifier functions:

1. Initially, Adaboost chooses a training subset arbitrarily.
2. Based on the accurate prediction of the last training, it iteratively trains the AdaBoost machine learning classifier by choosing the training set.
3. Higher weights are assigned to wrongly classified data points so that in upcoming iterations those data points will have higher chance for proper classification.
4. Weights are also assigned to classifiers in each iteration according to the prediction accuracy of the classifier. Higher the accuracy, higher the weight assigned.
5. This process continues until the entire training data fits without any error or until we’ve reached to the specified limit of estimators.
6. Carry out a voting mechanism across learning algorithms you built to classify the observations.

**3.5.10 GRADIENT BOOSTING CLASSIFIER**

Gradient boosting classifiers are a bunch of machine learning algorithms that bring together many weak learning models together to produce a powerful predictive model. Decision trees are commonly used when using gradient boosting. Gradient boosting models are gaining popularity because of their productivity at classifying compound datasets, and have been used lately to win many Kaggle data science competitions by enthusiasts.The thought behind "gradient boosting" is to take a poorly performing hypothesis or weak learning algorithm and make a sequence of modifications to it that will enhance the power of the hypothesis/learner. Such kind of Hypothesis Boosting is relies on the concept of Probability Approximately Correct Learning (PAC). This PAC learning technique investigates machine learning issues to understand how complex they are, and a closely related technique is employed to Hypothesis Boosting.

In hypothesis boosting, you consider all the data points that the machine learning algorithm is given training upon, and the observations that the machine learning method successfully classified are left behind, separating out all the other observations. Creation of a new poor learner takes place and it is tested on the set of data points that were poorly classified and the ones classified successfully are kept gradient boosting classifiers are the AdaBoosting technique joined with weighted minimization, after which the classifiers and weighted inputs are recalculated and reassigned. The goal of Gradient Boosting classifiers is to minimize the failure, or the difference between the actual class data of the training example and the predicted category value. It isn't required to understand the procedure for minimizing the classifier's loss, but it functions similarly to gradient descent in a neural network.

In the case of Gradient Boosting Machines, each time a fresh poor learner is put to the model, the weights of the earlier learners are maintained the same or cemented in place, left unmodified as the latest layers are introduced. This is different from the methods used in AdaBoosting where the values are modified when new learners are added.The fact that gradient boosting algorithms can be used on more than binary classification problems makes it more powerful. They can be employed for multi-class classification problem and also for regression problems.

**3.2.11 XGBOOST CLASSIFIER**

**XGBoost** is an open-source software library and it facilitates a regularizing gradient boosting framework supported by C++, Java, Python, R, Julia, Perl, and Scala. It functions on Linux, Windows, and macOS. Based on project interpretation, its goal is to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Component". It executes on a single machine, and also runs fine on the distributed processing frameworks like Apache Hadoop, Apache Spark, and Apache Flink. It has become popular and has grabbed the attention recently as the first-class algorithm for many winning teams of machine learning competitions. Major features of XGBoost which makes it different and unique from rest of the gradient boosting algorithms are as follows:

* Intelligent penalization of trees
* A proportional contracting of leaf nodes
* Newton Boosting
* Additional randomization parameter
* Contraption on single, distributed systems and out-of-core computation
* Pre-programmed Feature selection.

**3.5.12 PICKLE**

For serializing and de-serializing Python object structures, Pickle is utilized. It is also known as marshalling or flattening. Serialization depicts the procedure of transforming an object in memory to a byte format that can be held/saved on disk or shared with others over a network of systems. In future, this character stream can then be recovered and de-serialized back into a Python file. Pickling is a bit different when compared to compression. The former is the conversion of an object from one format (data in Random Access Memory (RAM)) to another (text on disk), while the latter involves the procedure of encoding data with fewer bits, so as to save disk space on our system.

Pickling is useful for purposes where you need some level of resolution in your data. Your program's state information can be stored on to a disk, so you can continue working on it later on. It can also be employed to transmit data over a Transmission Control Protocol (TCP) or socket connection, or to hold python objects in a database. Pickle comes very handy for when you're building and configuring machine learning algorithms, where you want to store them to be able to make new predictions at some point of time in near future, without having to rephrase everything or provide training to the model from the scratch.Pickle is not advisable when we want to work with data across multiple different programming languages. Its rules are specific to Python, thus, inter-language compatibility is not guaranteed. The same applies for various versions of Python itself. Unpickling a file that was pickled in some other version of Python may not always operate perfectly, so you have to ensure that you're using the same version and perform an upgrade if it is required. You should also avoid unpickling information from an unreliable source. Spiteful code present inside the file might be executed upon unpickling and could harm your local system.

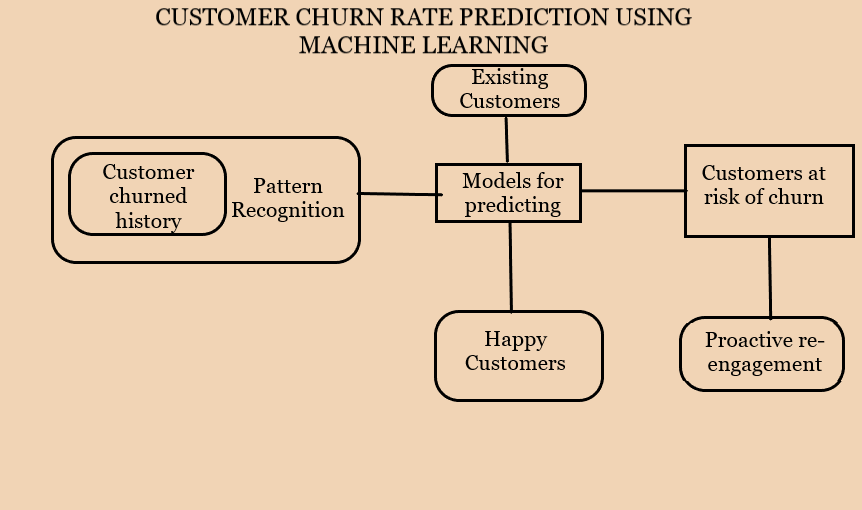
Objects with the following data types can undergo pickling:

* Booleans,
* Integers,
* Floats,
* Complex numbers,
* Strings,
* Tuples,
* Lists,
* Sets, and
* Dictionaries that contain pickable objects.

Classes and functions can also be pickled, for instance, if they are defined at the top level of a module.

**4.DESIGN**

**4.1 BLOCK DIAGRAM**



**Fig 10: Block Diagram**

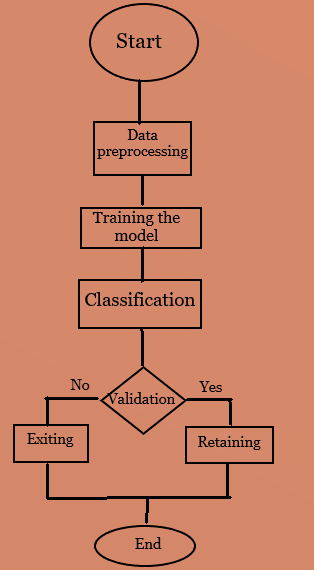
As the volume of data is huge and is constantly changing it is intimidating job for the analysts to work on the data continuously for time being. That’s when customer churn prediction using machine learning come into the picture which play a major part. The prediction takes place step by step and every step has its own importance. Initially the data is collected from the valid sources, it’s pre processed and is transformed into a model which is used for prediction. The next stage is modelling, testing and finally deploying the model. Accuracy rate when compared with the traditional approach is fast for machine learning techniques. As per the block diagram the whole process is based on the customers are with the bank or exiting. Based on the earlier records of the customers the process of pattern recognition is done, it is then mapped into building the models which are used for predicting purpose.

The prediction divides into two ways which are happy customer and churned customers. Happy customers are the one’s who are satisfied with the necessary features provided by the bank, the customers who are at risk bank try solve their issues to get them back which is nothing but the proactive re-engagement. The customer strategies should be updated so that its profitable and create a hype for the people to engage more.

The proactive support depends on the capability of bank to forecast the customers wants before they experience. There are certain ways of creating proactive engagement opportunities. Based on the surveys customer retention and acquisition are available online, but getting new customer is better than retaining existing customers. There is correlation between churn rate and increased acquisition spend. Customer churned history deals with past records of people churning i.e. separating from the bank. Hence will have an idea of what type of customers need to be targeted and let them stay with bank for long time.

Pattern recognition from the diagram describes the process of recognizing the patterns by using machine learning algorithms. It can be identified based the knowledge obtained by extracting various patterns. A label is allotted to a pattern based on the extraction that are generated using a set of training patterns. Hence the whole procedure is to generate a best accuracy rate which will help the banks to retained the customers based on prediction.

**4.2 FLOWCHART**



**Fig 11: Customer Churn Flowchart**

1. The dataset considered here is obtained from Kaggle and name of the dataset is “Bank customer churn data”. The dataset consists of 1000 records and 14 features.

2. Data pre-processing is the most basic and important step in any machine learning project. Data processing involve certain steps they are:

-Data cleaning

-Data imputation

-Dealing with outliers

- Data transformation

-Data visualization

3. The next step is training the model and classification is done using different algorithms.

4. Validation provides us with the information whether a customer is churned or not.

5. It involves two cases based on the input values entered, the prediction is made in such a way that customers are retained or exited.

6. If its **Yes** the customer are retained and he/she is with the bank.

7. In case of **No** customer is exited which is nothing but the customer is churned.

8. We intend to make a front end(website) were the values entered give the output which help the company know whether the customer/client is staying or leaving the organization/bank.

9.Our project intends to predict the best accuracy rate by comparing different algorithms and the output is displayed on the website.

1. **IMPLEMENTATION**

**5.1 DATA COLLECTION**

The Data comprehension period deals with assembling and exploring the data to get the insights of the whole data. The dataset can be numeric, categorical etc. It’s basically gathering information from various sources. The data provides us information regarding the past events, based on that we can find recuring patterns. Patterns can be used for building models with the help of various algorithms which are used for future prediction. The coverage of data collection is based on a recent surveys and past data of the customer. Data is laying across several servers and the data should be gathered at one place for ease of access. Data is present in various formats therefore it should be changed to particular format for data collection. JSON contains data of chat servers, data of business applications is generally tabular. If both kinds of data need to used it should be converted to either JSON or whole data to CSV or xlsx. Data is also present in HTML text it should be cleaned for further usage purpose. There are some methods for data collection. Acquisition of appropriate data for training the models is the problem to overcome as machine learning is widely used and we don’t have enough labelled data. If the count of datasets are increasing rapidly, searching for the right ones itself becomes a challenge. Next method involve data labelling which is necessary in all supervised learning applications. As manual labelling is expensive, there are some scalable techniques 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.

**5.2 PRE-PROCESSING**

Data processing is something where the raw data is prepared and is made in a suitable form where it can further used for machine learning model. As the data is not clean, pre-processing play a vital role in removing unwanted missing values and unusable format of data which effect the model building. The task of pre-processing the data increases the accuracy and efficiency of the model. Data pre-processing is a place where the data is encoded or transformed, to guide to that state so that parsing is done easily and the features can be interpreted by the algorithm. Data pre-processing involve several stages they are:

* Data Inspection
* Data Filtering
* Data Cleaning
* Feature selection
* Data interpolation
* Data splitting into train and test sets
* Feature scaling

**5.2.1 DATA INSPECTION**

Once the data is imported, it should be checked so that there are any missing values and various other checks for ensuring consistency in data. In such cases domain knowledge come in a handy manner. For the missing data we checkout the rows and columns which are having no or null data. If there are any cases found the decisions are made based on the scenarios or instincts. Domain knowledge help in deciding the need of some columns, if there are 40% of missing data then that column is discarded. In a rare case if the data is less than 40% various interpolation and replacement techniques are used to fill the space of missing data. Most commonly used are replacement of nulls by central tendency which is nothing but mean, mode, median. During model building statistical significance tests can also be used to determine which column to need be used.

**5.2.2 DATA FILTERING**

The process of data filtering deals with selecting a part of dataset and that subset of data is used for analysis purpose. It is temporary process as the whole data is kept separate and some part is used for calculation purpose. It can be used for seeing the results for particular period of time, calculating the results for particular group of interests and train, validate statistical models. Filtering need to be stated with certain rules to identify the instances which are included in analysis.

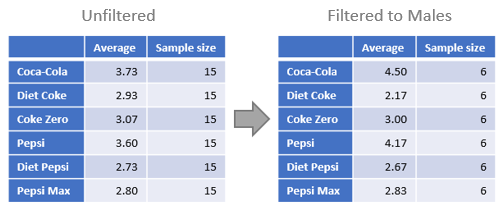
Filtering the data involve:

1. Rules for the observation needed.

2. Select the observations that fit the rule.

3. Conducting the analysis using information contained in the selected observations.

Filtering is carried out implicitly sometimes and the best example is in surveys the columns of the crosstab correspond to a special case of filtering where the results are separately computed for each column and they are shown side by side.



**Fig 12: Data Filtering**

**5.2.3 DATA CLEANING**

Data cleaning is the process where incorrect, inaccurate, and irrelevant data is modified, replaced or deleted whenever needed. It is considered as the fundamental element in machine learning. Data is very precious thing for analysis and Machine learning as it is used in computing or business data and in other departments as well. 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 then output will be inaccurate. There are some examples of data cleaning. A person X is a general manager of a company. The company collects data of people visiting the place and buying the product from the company. As the information is available with the company, they will have the insights of which products people are more attracted to and the company will try to increase the production of product. One drawback here is, if there are any missing values of company data then the whole analysis will go wrong. Model is based on how the data is clean and based on that processing is done. Data cleaning involve imputation or handling because some of the chosen algorithms which can handle missing data.

**5.2.4 FEATURE SELECTION**

Feature selection is one of the key notions in machine learning which has the huge contribution for the performance of the model. Data cleaning and feature selection are the most important steps for model designing. In feature selection features are selected manually or automatically which contribute to the prediction. It is important to have relevant features as irrelevant features decrease the accuracy of the model which create a huge blunder in the output. It is applied to the dataset where there are huge number of dependent variables in the given dataset. It trains the model faster and reduces the complexity. There are three benefits of selecting features before the modelling is done, they are reduce overfitting, improves accuracy, reduces 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 which automatically increases the accuracy.

3. Reduces Training time: Algorithm complexity is reduced as there are few data points.

There are three Feature selection methods:

1. Univariate selection

2. Feature importance

3. Correlation Matrix with heatmap

1. Univariate selection: This method works by selecting the best features based on the statistical tests. Each feature is compared with target variable to check whether they are statistically significant or not. This process is called as analysis of variance (ANOVA). It is a pre-processing step to an estimator. The scikit library provides the class called SelectKBest which is used with the suite of various statistical tests to select the specific number of features.

2. Feature importance: It gives the score for each feature of the data, high is the score more important are the features towards the output variable. Feature importance can be achieved for each feature of the dataset using feature importance property of the model. It is an in-built class which come with the tree based classifiers.

3. Correlation matrix with heatmap: This state how the features are related with the target variable. Correlation can be positive where increase in value of one feature increases the value of target variable when negative increase in value of feature decreases the value of target variable. Heatmap makes is simpler to recognize which features are related to the target variable based on that a heatmap plot is drawn using the correlated features using the seaborn library.

**5.2.5 DATA INTERPOLATION**

It is the process where known data values are used to estimate unknown data values. Different interpolation techniques are often used in the atmospheric sciences. One of the easiest methods is linear interpolation which requires knowledge of two points and the constant rate of change between them. Data interpolation can be used for adding missing values to the columns with cells having missing values. There are various procedures which can be used in interpolation, and the most prominent are average interpolation, KNN- interpolation etc.

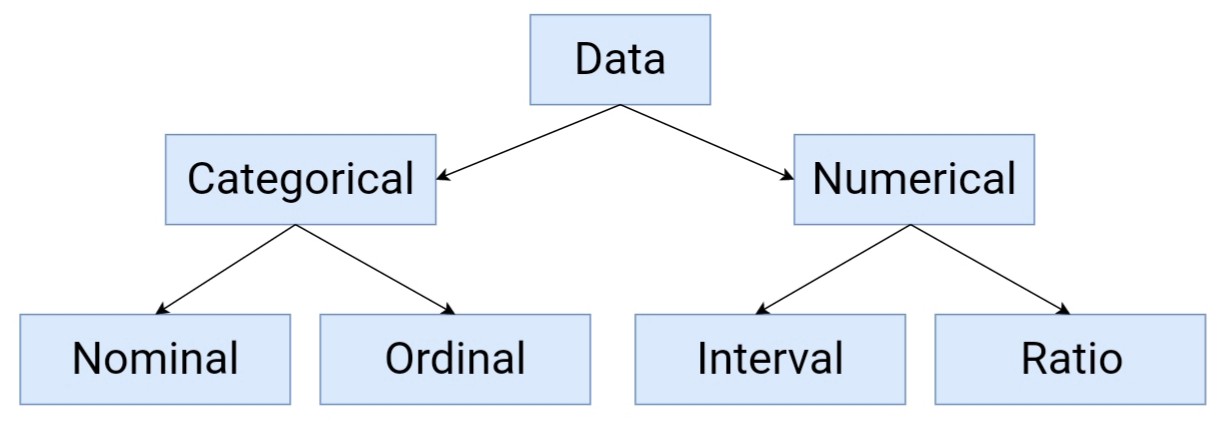
**5.2.6 DATA SPLITTING**

Data is split into test and validation before it is fed to machine learning. This procedure is used to estimate the performance of machine learning algorithms during prediction on data not used during the training the model. Sklearn library of python provides a special function to train-test-split for it. It is appropriate when there is a large dataset and it can be used for classification or regression problems and can also be used for any supervised learning algorithm. The whole dataset is divided into two subsets. First subset is used to fit the model and it is called as training dataset. The second subset are used to make predictions and are compared with the expected values. It is also called as the test dataset. The goal is to guess the performance of machine learning model on new data. The percentage of data can be specified into train or test sets. The process returns four arguments which are training independent variable, training dependent variable, testing independent variables and testing dependent variable.

**5.2.7 FEATURE SCALING**

Feature scaling is a technique to standardize the independent features present in the data with in a fixed range. It’s a standard normalization of data. This is done so that independent variable has more importance. If there is no feature scaling the machine learning algorithms tend to weigh greater values. The columns here are normalized individually so that there is same distribution. Feature scaling is the final step in data pre-processing.

**5.2.8 FEATURES IN MACHINE LEARNING**



**Fig 13: Statistical Data Types**

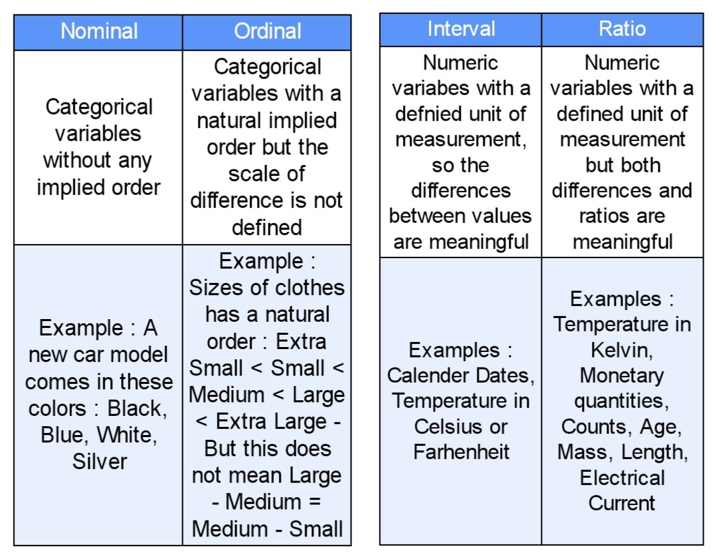
They are of two types:

1. Categorical

The variables which are of two or more categories and there is no particular order. It is also called as nominal variable. Let us consider an example where a binary variable is a categorical variable with two categories yes or no and there is no order for that categories. Hair colour also come’s under categorical variable as it has various categories such as blonde, brown, red etc. A nominal variable is something that allows you to assign categories but we cannot order them. If the variable has an order, then it is considered as ordinal variable. Categorical variables are of two types nominal and ordinal.

2. Numerical:

The variables which take a value within a finite or infinite interval. Numerical variables are of two types interval and ratio. Features for numerical variables are continuous or integer valued. These variables are represented by numbers and they have the properties of numbers. For example, the number of steps a person walk in a day, speed of a car etc these all are calculated numerically,



**Fig 14: Table Showing Types of Categorical and Numerical Variables**

**5.2.9 MISSING VALUES:**

It is a major task of handling missing values as the algorithms of machine learning doesn’t support the data having missing values. Dealing with the missing data is huge task in data preparation phase as there is better way of dealing with them. To know what to do with the missing values first we should identify the type of missing values present in the dataset. There are various factors for missing values present in a dataset. One reason is that data was not gathered properly or inappropriate data. The variables which contain more than 60% missing values can be removed from the dataset. The continuous variables can take any values it maybe minimum or maximum and the missing values are imputed by mean values from 2% to 30%. The reasonable estimate for randomly selected observations is mean from a normal distribution. Missing values create many problems and most of the time is consumed resolving them as it reduces the statistical power which cause wrong evaluation of hypothesis. Missing values will decrease the representativeness of the sample and will complicate the whole analysis of data. Some algorithms don’t work with missing data. The labels present in categorical variables of missing values are imputed use final observations and they are pushed forward where most of the techniques for imputing missing values. The values where imputed which occurred most of the time.

* Eliminate rows with missing data:  
  It is a simple and an effective strategy. It fails when many objects have missing values. When the feature has at most missing values then that particular feature can be removed.
* Estimate missing values:  
  If some appropriate numbers of values are missing, then simple interpolation methods can be used for filling the missing values. Most common method for dealing with the missing values is mean, median, mode.

There are three types of missing data:

1. Missing at Random (MAR)

2. Missing Completely at Random (MCAR)

3. Missing Not at Random (MNAR)



**Fig 15: Types of Missing Values**

1. Missing a random (MAR):

It takes place when missingness is not random but it can be fully considered for variables with the total information. It is an assumption which is not possible to prove statistically. MAR data has the standard relationship, it is the tendency of missing values and the observed data but not the missing data.

2. Missing completely at random:

This describes that there is no relationship between the missingness of data and values which are considered are being missing. This type of missing values is easy to understand. The main point here is that the data is missing has no link with the observed and non-observed data, and there is no sense in it. Missing data is just the random subset of the data.

3. Missing not at random (MNAR):

It is defined as there is connection between the readiness of a value and missing value. We should know the values of missing data to tell whether is it’s a MNAR. The reality here points to the data missing and is directly related to the unobserved data which doesn’t exist, the missingness is connected with the things that we aren’t aware of. The best way to know about the missing data is to understand the data collection process, and there are methods to know whether the given data is MCAR or MAR using statistical methods.

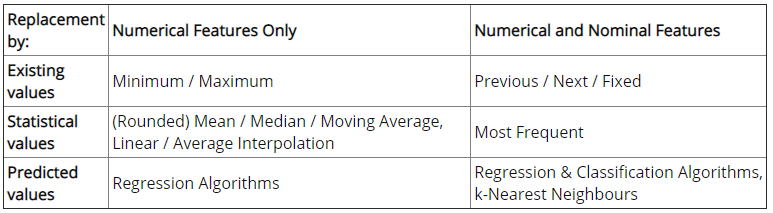
**5.2.9.1 IMPUTATION AND ITS TYPES:**

It is the replacement of missing data with the substituted values. When all the missing values are imputed, the dataset is analysed using various standard techniques for completing the data. Basically, the columns present in the dataset with numeric continuous values are replaced with mean, median or mode and this will supress the loss of data. There are certain issues that missing data can initiate substantial amount of bias, it makes the analysis of data difficult and reduces the efficiency.

Imputation maintains everything by replacing missing data with some unknown value using the known information. When all the missing values are imputed, the dataset is analysed using various standard techniques to complete the data. There are two types of imputation methods single and multiple imputation.

1. SINGLE IMPUTATION:

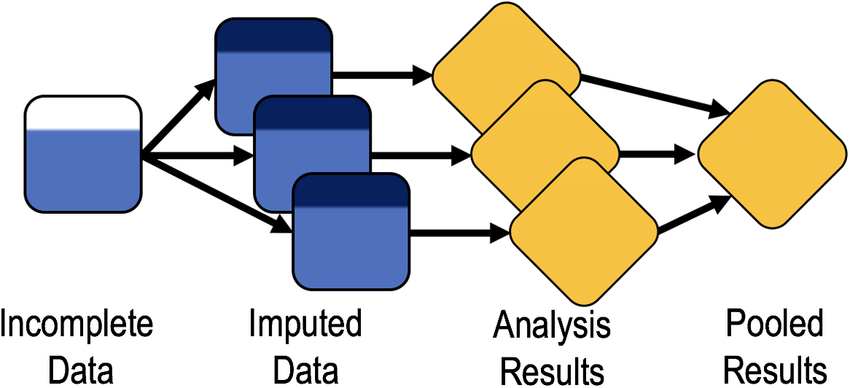
In single imputation, missing observations are generated using a single value. The imputed values are considered as the real value by avoiding the known fact that no method of imputation can generate the correct and exact value. Single imputation doesn’t return the missing values uncertainty. Most of the methods are of single imputation and they are associated with three things which are replacement by existing values, replacement by statistical values, replacement by predicted values. Based on the values used with the help of those schemes they work on both numerical and nominal columns.



**Fig 16: Table Showing Single Imputation Methods for Numerical Features and Nominal Features, Based on Existing Values, Statistical Measures, and Predicted Values**

2. Multiple Imputation

Multiple imputation is an imputation approach derived from statistics, the single imputation strategies have the disadvantage that they do not take into account the uncertainty of the imputed values, that is, they recognize the imputed values ​​as real values ​​without taking into account the standard error of the biases associated with it the approach that solves this problem is to use multiple imputations if there are none, but multiple imputations are created for each missing value. To do this, the missing values ​​must be filled in several data records of. multiple imputation. A well-known algorithm MICE (Multiple Chained Equation imputation) has been developed for multiple imputation. MICE is a robust, informative method for dealing with missing values in datasets.

**Fig 17: Multiple Imputation**

**5.2.10 OUTLIERS**

An outlier is defined as a data point which is different from other points. It’s due to variability in the measurement or may indicate experimental errors. The outliers must be prohibited from the data set as they detect abnormal instances which are very difficult and typical. Data outliers damage and mislead the training process as it takes longer training times than usual by generating less accurate models and ultimately poorer results. The outliers can be dealt in three ways:

1. Univariate method: This technique checks for data points with extreme values on one variable. One of the basic methods for identifying outliers is box plot. A box plot is a graphical exhibit for relating the distributions of the data. Box plot make use of median, lower and upper quartiles. The famous Tukey's method describes an outlier as the values of a variable that fall far from the central point nothing but a median. The maximum distance from the centre of the data to be enclosed is called cleaning parameter. If the cleaning parameter is extensive, the test becomes less sensitive to outliers. If the parameter values are small, many values are detected as outliers.

2. Multivariate method:

There is no need of extreme values for outliers. As we know with respect to point B, the univariate method does not always work well. Then multivariate method plays an important role by building a model using all the data available, and then cleaning those instances with errors above a given value.

3. Minkowski error:

The Minkowski error is a misfortune list which is more obtuse toward exceptions than the standard total squared mistake. The aggregate squared error raises each case mistake to the square, making a too enormous commitment of exceptions to the absolute mistake. The Minkowski error addresses that by raising each occurrence mistake to a number more modest than 2, for example 1.5. This diminishes the commitment of anomalies to get error. For example, if an anomaly has a mistake of 10, the squared error for that case will be 100, while the Minkowski error will be 31.62. The formula are given below which are used in the above examples.

**Mean squared error** = ∑ (outputs−targets) ^ 2 / instances number

**Minkowski error** =∑(outputs−targets) ^ minkowski-parameter / instances number

**5.3 DATA TRANSFORMATION**

It is the process of changing the format, structure or the data values. According to the machine learning projects data is transformed at two stages in the data pipeline. In ETL (extract, transform, load) process transformation play a crucial role where it involves data integration, data migration, data warehousing etc. It may be constructive which include (adding, copying, and replicating data), destructive which include (deleting fields and records) and structural which include (renaming, moving, and combining columns in a database). There are certain benefits and challenges of data transformation.

**BENEFITS**

1. The Data is modified so that it can be easy for the people who are working with it using transformation.

2. The data is made in format and validated to improve the data quality and protect the them from potential landmines such as null values, unexpected duplicates, incorrect indexing, and incompatible formats.

3. Information change works with similarity between applications, frameworks, and sorts of information. Information utilized for various reasons should be changed in an unexpected way.

**CHALLENGES**

1. Information change can be costly. The expense is reliant upon the particular foundation, programming, and instruments used to deal with information. Costs may incorporate those identified with authorizing, processing assets, and employing essential faculty.

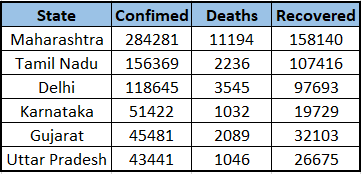
2. Absence of skill and indiscretion can be present issues during the change. Information investigators without proper knowledge ability are less inclined to see grammatical errors i.e., they are less acquainted with the scope of precise and passable qualities. For instance, somebody chipping away at clinical information who is new to applicable terms may neglect to signal illness names that ought to be planned to a solitary worth or notice incorrect spellings.

**5.3.1 ENCODING AND ITS TYPES:**

The input and the output variables need to be numeric according to the machine learning models. It is nothing but if the data consists of categorical data, we use 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 is assigned with numbers and things which are used to measure it may be dimension, temperature, humidity, area and volume etc. Qualitative data is assigned with the characteristics which cannot be measured such as smells, tastes, textures, attractiveness, and colour. There are three types of encoding:

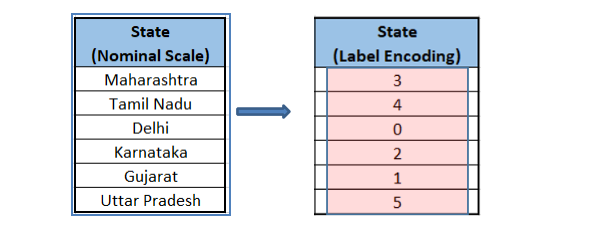
1. Label encoding:

In this encoding method, the unmitigated information is doled out a worth from 1 to N (where N is the number for various classifications present in the information). This sort of an encoding strategy is applied to the ordinal information. The doling out of the worth from 1 to N happens either in an expanding or a diminishing request. Once if the request is picked to be rising or plunging it is fixed all through for every one of the qualities in the section and can't be changed arbitrarily or in the middle. The solitary limitation that accompanies the ordinal information is the authoritative request should be in either expanding or diminishing request. Label encoding is explained in brief with an example of covid-19 cases.



**Fig 18: Table of Covid-19 cases**

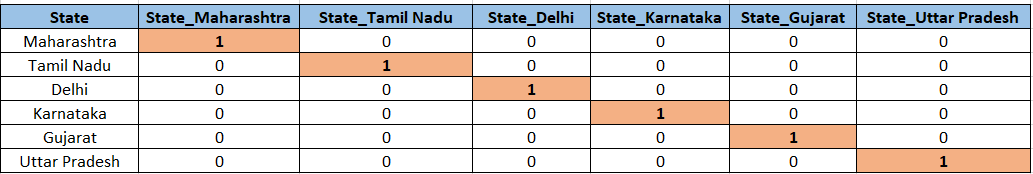
From the above table state column consists of categorical variables, we use label encoding here to convert to numerical data. As there are several states, each state is assigned with numeric value starting from “ 0 ” based on the alphabetical order. The process of assigning numerical value to categorical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



**Fig 19: After Label Encoding**

2. One-hot encoding:

In one-hot encoding for each type of feature a new column is created called dummy variable with binary encoding i.e., 0 or 1 to represent the row is relevant to particular category. It involves mapping of various categories present in the features depending on the presence or absence of feature. It is normally applied to nominal data present in the dataset. Based on the previous example from the fig 5.3.1.1 a new state table is created with six columns starting from Maharashtra to Uttar Pradesh. The value 1 is allotted to a specific line that has a place with the classification, and 0 is allotted to the remaining columns that doesn't have a place with this class.

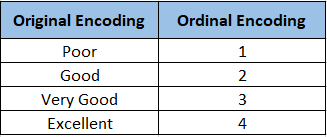


**Fig 20: One Hot Encoding Table**

One-hot encoding is the way of making extra sections, 1 for every special value in the arrangement of the downright characteristic we'd prefer to encode. To keep it basic, one-hot encoding is a useful asset, yet it is just a material for all information that have a low number of kind qualities. There is a drawback for this method as it increases the dimensionality of the dataset. It is a most powerful mechanism which is applicable for categorical data that have a low number of unique values. Dummy variables are created to form the redundancy to the dataset. If there are 3 categories, there should be two dummy variables because, if an observation is neither one or two it must be the third one. It is generally referred as the dummy variable trap, which is best way to remove one dummy variable column from such an encoding.

3. Ordinal Encoding:

An Ordinal encoding is utilized to encode unmitigated highlights into an ordinal mathematical value. This methodology changes clear cut value to mathematical worth in arranged sets. This encoding strategy shows up practically like Label encoding but it will not consider a variable as ordinal, however in case of ordinal encoding, it will allocate a succession of numerical values according to the information. We should make an example ordinal downright information identified with the client criticism overview, and afterward we will apply the Ordinal Encoder strategy. For this situation, suppose the information is gathered utilizing a Likert scale in which mathematical code 1 is allotted to Poor, 2 for Good, 3 for Very Good, 4 for Excellent. We can notice one thing here that 5 is better than 4 and much superior to 3, yet taking the distinction somewhere in the range of 5 and 2 is trivial.

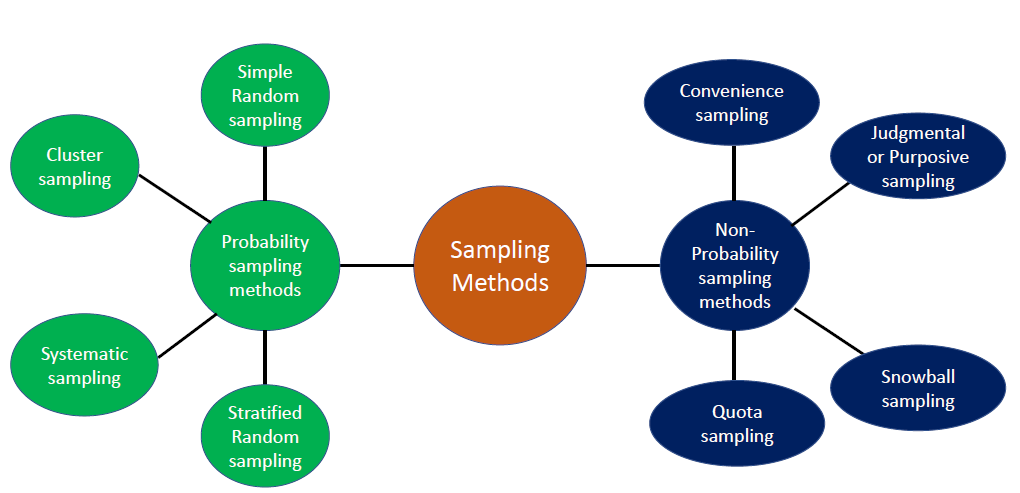


**Fig 21: Ordinal Encoding**

**5.3.2 SAMPLING AND IT’S TYPES**

Sampling is a process of selecting individual members or a whole subset to make statistical inferences from them and estimate characteristics of the whole population. Sampling is done to know about conclusions of populations from samples, and this helps us determine the characteristics by directly observing only a part of the population. The way of selecting a sample requires less time than selecting every item in a population. Sample selection is a cost-efficient method.

There are different types of sampling techniques shown in the below figure.



**Fig 22:Types of Sampling Techniques**

In probability sampling each and every component in the population has equal probability of getting selected. This provides the better way to create a sample that is truly representative of the population. Where as in non-probability sampling there is no equal probability of getting selected and it involves certain risk of non-representative sample which does not produce generalized output.

Types of Probability sampling:

1. Simple Random sampling:

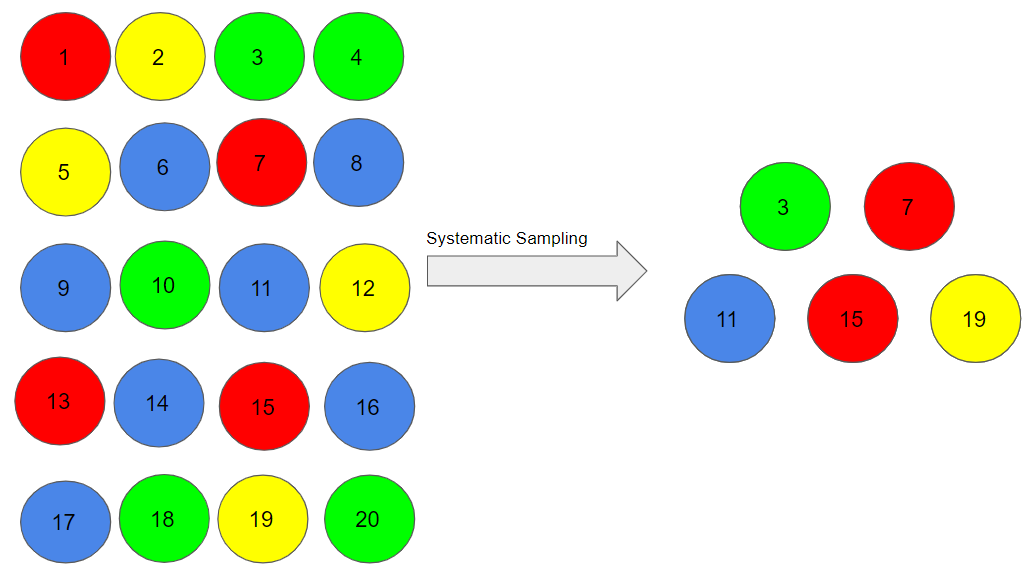
In this method all the individuals are selected based on the chance and every person of the population has the equal chance of getting selected. The advantage of this method is, it is a direct probability sampling. As it comes with a caveat – it’s not possible to choose individuals with the characteristics of interest. To estimate the unknown parameters Monte Carlo methods, use repeated random sampling for the estimation.



**Fig 23: Simple Random Sampling**

2. Systematic Sampling:

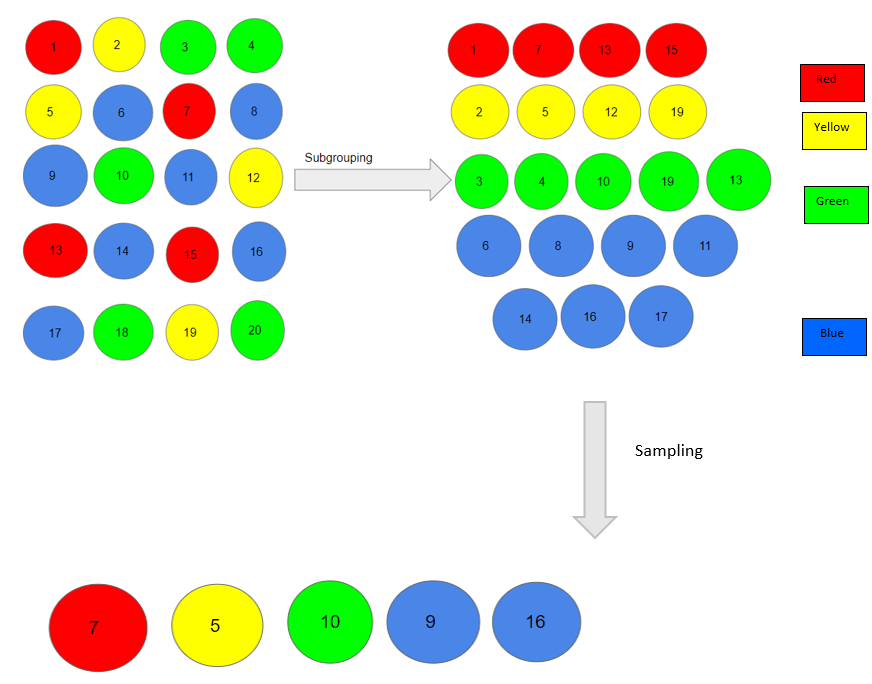
In this type of sampling, the first individual is selected randomly and others are selected using a fixed ‘sampling interval’. Let’s take a simple example to understand this. Say our population size is x and we have to select a sample size of n. Then, the next individual that we will select would be x/nth intervals away from the first individual. We can select the rest in the same way. Systematic sampling is more convenient than simple random sampling. However, it might also lead to bias if there is an underlying pattern in which we are selecting items from the population (though the chances of that happening are quite rare).



**Fig 24: Systematic Sampling**

3. Stratified sampling:

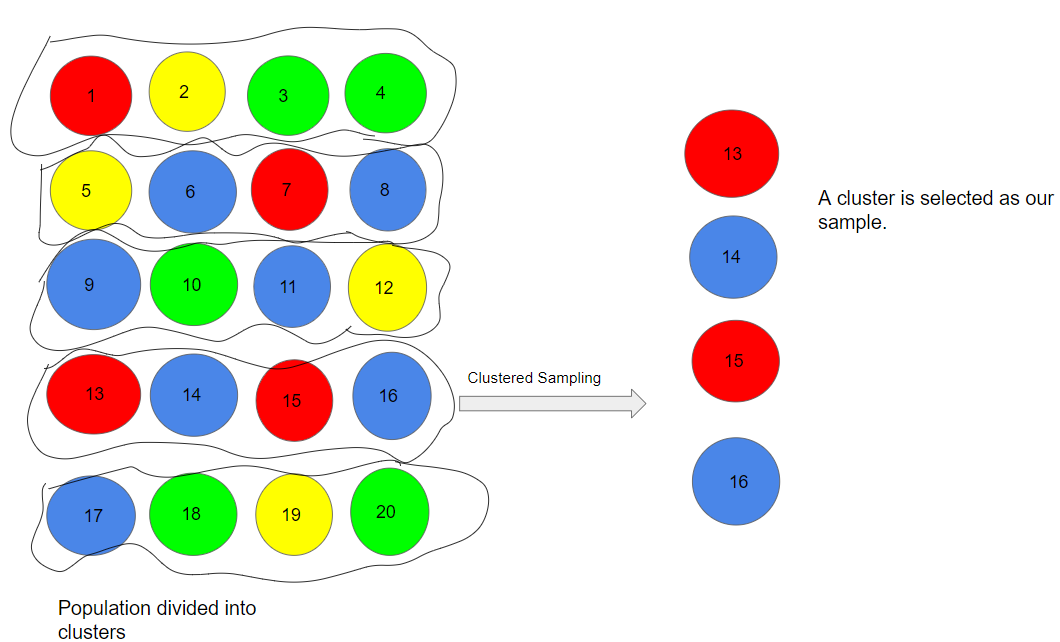
In this kind of sampling, we partition the population into subgroups (called layers) in view of various qualities like sex, class etc and then we select the samples from these subgroups. The initial partitioned population is divided into subgroups depending on various shades of red, yellow, green and blue. At that point, from each tone, we chose a person in the extent of their numbers in the population. We utilize this sort of sampling when we need portrayal from all the subgroups of the population.



**Fig 25: Stratified Sampling**

4. Cluster Sampling:

In clustered sampling, we utilize the subgroups of the population as the sampling unit other than individuals. The population is separated into subgroups, known as clusters, and an entire group is chosen to be remembered for the study: The example is partitioned into 5 groups or clusters and each cluster comprises of 4 people and we have taken the fourth cluster in our sample. We can incorporate more clusters according to our sample size.

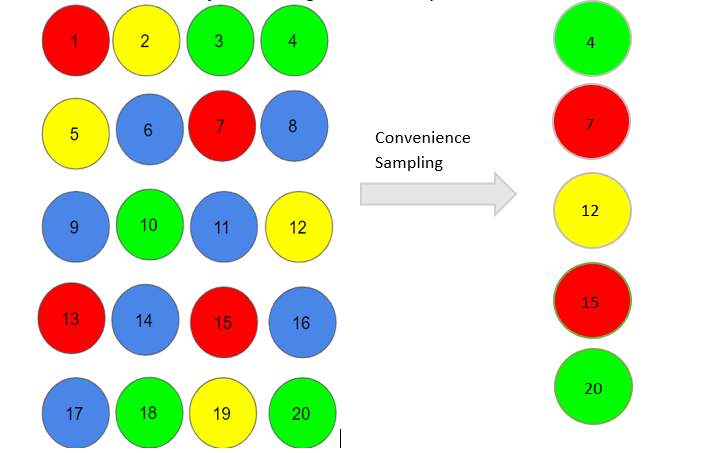


**Fig 26: Cluster Sampling**

Types of Non-probability sampling:

1. Convenience Sampling:

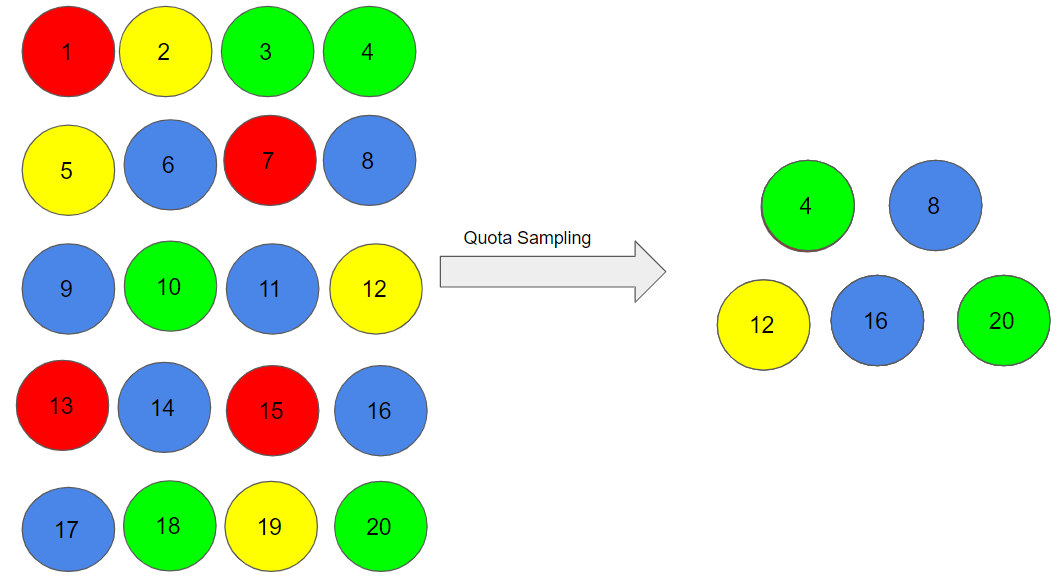
This is maybe the most effortless strategy for sampling since people are chosen depending on their accessibility and eagerness to participate. Here, suppose people numbered 4, 7, 12, 15 and 20 need to be essential for our sample, and consequently, we will remember them for the sample. Convenience sampling is inclined to significant bias, on that grounds the sample may not be the portrayal of the particular attributes like religion or, say the sexual orientation, of the population.



**Fig 27: Convenience Sampling**

2. Quota Sampling:

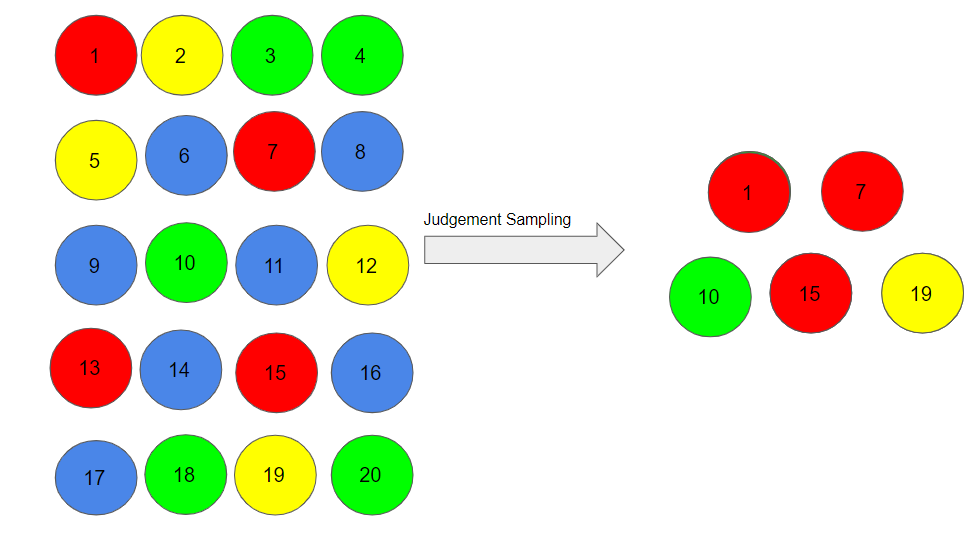
In this sort of sampling, we pick things depending on foreordained qualities of the population. Consider that we need to choose people having a number in products of four for our sample. Therefore, the people numbered 4, 8, 12, 16, and 20 are as of now held for our sample. In quota sampling, the chosen test probably won't be the best portrayal of the characteristics of the population which aren't under consideration.



**Fig 28: Quota Sampling**

3. Judgement Sampling:

Judgement sampling also known as selective sampling and it depends on the judgment of the experts for participation. If the experts accept that people 1, 7, 10, 15, and 19 are reviewed for sampling and it also help to deduce the population in a better way.



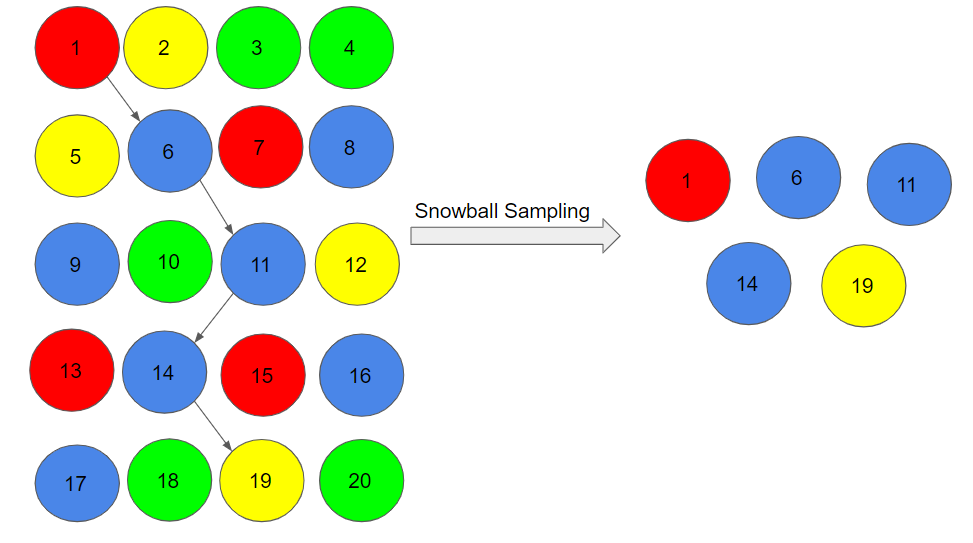
**Fig 29: Judgement sampling**

4. Snowball Sampling:

In snowball sampling the **existing persons were told to nominate people so that the size of the snowball increases in size.** This process of sampling is productive when a sampling frame is difficult to understand. Let a random person 1 is chosen for the sample and then he/she suggested person 6, and person 6 suggested person 11 and this continues.

1->6->11->14->19

There is particular risk of selection bias in snowball sampling, as the individuals will share common traits with the person who recommends them.



**Fig 30: Snowball Sampling**

**5.3.3 SMOTE FOR IMBALANCED CLASSIFICATION**

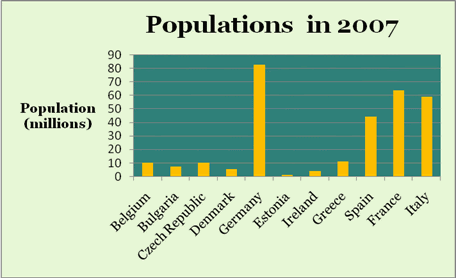
Imbalanced classification includes creating predictive models on classification datasets that have a serious class imbalance. The main challenge of working with imbalanced datasets is that most AI methods will disregard, and they lack performance wise, therefore minority class is generally significant. One way to deal with addressing imbalanced datasets is to over sample the minority class. The easiest methodology includes duplicating examples in the minority class, as these examples don't add any new data to the model. All things being equal, new examples can be integrated from the current examples. This sort of information increase in the minority class and is alluded to as the Synthetic Minority Oversampling Technique or SMOTE. One approach to take care of this issue is to over sample the models in the minority class. This can be accomplished by basically duplicating examples from the minority class in the training dataset before fitting a model. This can adjust the class conveyance however doesn't give any extra data to the model. An enhancement for duplicating examples from the minority class is to combine new models from the minority class. This is a kind of information increase for even tabular data and can be extremely powerful. SMOTE first chooses a minority class instance aimlessly and tracks it down, and find its k nearest minority class neighbours. The manufactured occurrence is then made by picking one of the k nearest neighbours b randomly and associating it with an and b to become line segment the synthesized instances are created as an convex mix of the two picked examples an and b.

**5.3.4 DATA VISUALIZATION**

Information representation refers to the procedures used to impart information or data by encoding it as visual items (focuses points, lines or bars) contained in designs or graphics. Visualizing information is viable when done correctly. We characterize right when the information representations have served its need. A speedy test - when individuals can decipher your visualization by posing more questions on the data showed versus how it is shown, at that point one will realize that you are on the correct way. So to be exceptionally viable, it is essential to plan the correct representations for your information to permit yourself and colleagues to decipher and settle on choices dependent on what they notice. We make the appropriate representations by understanding the various kinds of perceptions and addressing the questions. Data visualization provide insight of the data graphically. There 5 types of visualization temporal, hierarchical, Network, Multidimensional, Geospatial. There are some plots which help one get the insights of data.

1. Bar Chart Visualization:

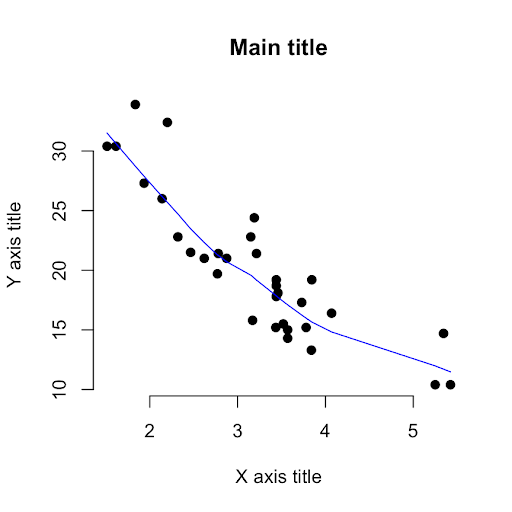
* Bar charts are such a favoured graph visualization due to how simple you can examine them for fast data. Bar diagrams arrange information into rectangular bars that make it a breeze to look at related informational datasets. Bar charts are used to compare two or more values in the same category, when you want to compare parts of a whole, when you don’t have too many groups, when you want to understand how multiple similar data sets relate with each other. When you’re using bar graph use a different colour for each category you’re comparing, and make sure you also use solid lines to keep the line chart clear and accurate.



**Fig 31: Random Example of Bar Plot**

2. Scatter Plot:

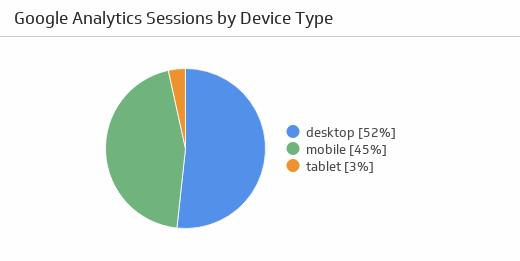
Scatter plots are the correct data visualization to utilize when there are various data points, and you need to feature similarity of the data set. This is valuable when searching for outliers or for understanding the distribution of your data. On the other if the data frames a band reaching out from lower left to upper right, there probably a positive relationship between the two variables. If band runs from upper left to down right, a negative correlation is probable. If it is difficult to identify a pattern, there is no correlation. Scatter plot is used to describe the relationship between the variables and make a compact data visualization. Trend line are best way to analyse the data present on the scatter plot.



**Fig 32: Random Scatter Plot Example**

3. Pie Chart:

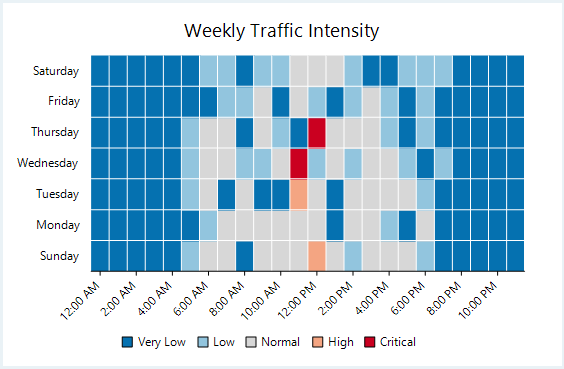
Pie diagrams are a fascinating chart visualization. At a significant level, they're not difficult to understand and comprehend on the grounds that the pieces of an entire relationship is made self-evident. In any case, top information visual specialists concur that one of their drawback is that the level of each part isn't clear without adding mathematical values to each cut of the pie. Pie charts are used to compare relative values.



**Fig 33: Pie Plot Example**

4. Heatmap:

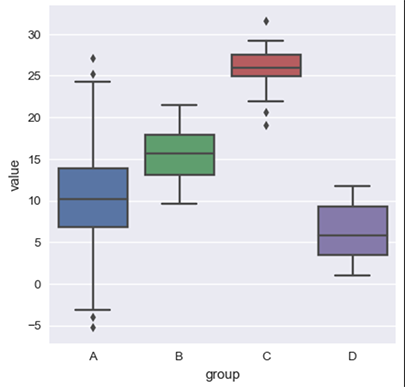
A heat map or choropleth map is a data visualization that shows the connection between two measures and gives rating data. The rating data is shown using changing tones or saturation and can display evaluations, for example, high to low or awful to amazing, and needs improvement to functioning admirably. It can likewise be a topical guide where the region inside perceived limits is concealed with respect to the information being addressed. Heat map are used to illustrate the details and display the relationship between two measures.



**Fig 34: Heatmap Plot Example**

5. Boxplot:

A box plot, or box and whisker diagram is a visual representation of showing an appropriation of information, for the most part across groups based on the five-number synopsis: the minimum, first quartile, the median (second quartile), third quartile, and the maximum. The easiest of box plots will show the full scope of variety from least to greatest, the reasonable scope of variety, and a common value. A box plot will likewise show the outliers. Box plots are used to compare the distribution of data and identify median, minimum and maximum.



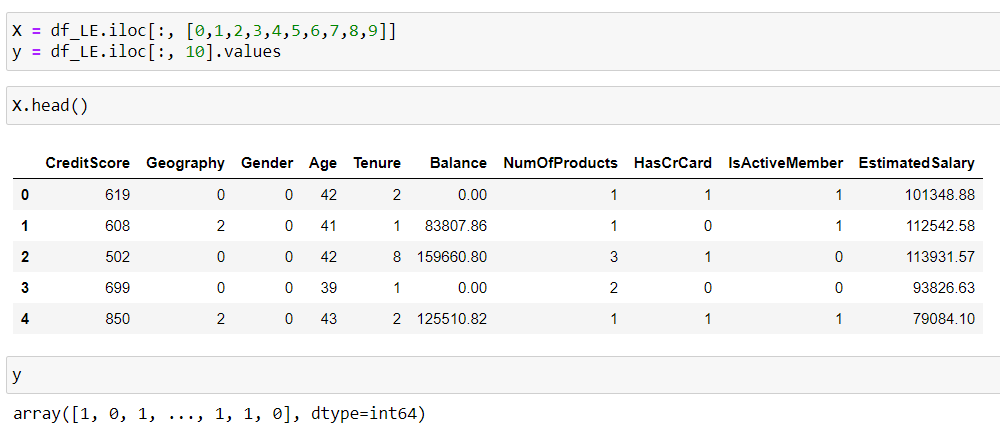
**Fig 34: Box Plot Example**

**5.4 Model Construction:**

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.

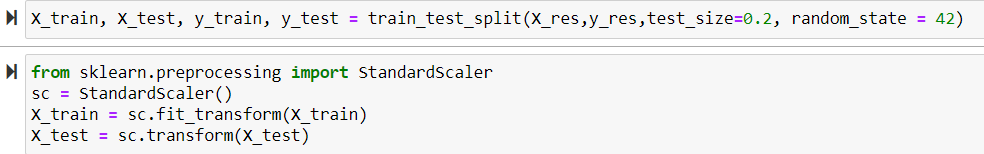
The Machine learning algorithms used in model construction are:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. K-NN classifier
5. AdaBoost Classifier
6. Gradient Boosting Classifier
7. eXtreme Gradient Boosting classifier

The model construction can be divided into three major parts:

**Fig 35: First five Records In Data**

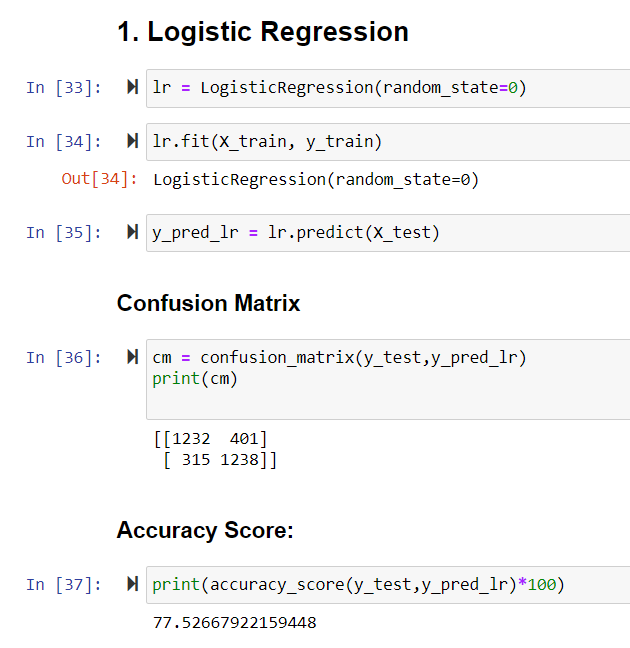
1. Splitting the Predictors(X) and the Target variable(y).
2. Splitting the X and y further. That is the test-train split. Now, we have 4 variables. X\_train, X\_test ; y\_train, y\_test. The train data is 80% of the complete data whereas the test data is 20% of the complete data sampled randomly.



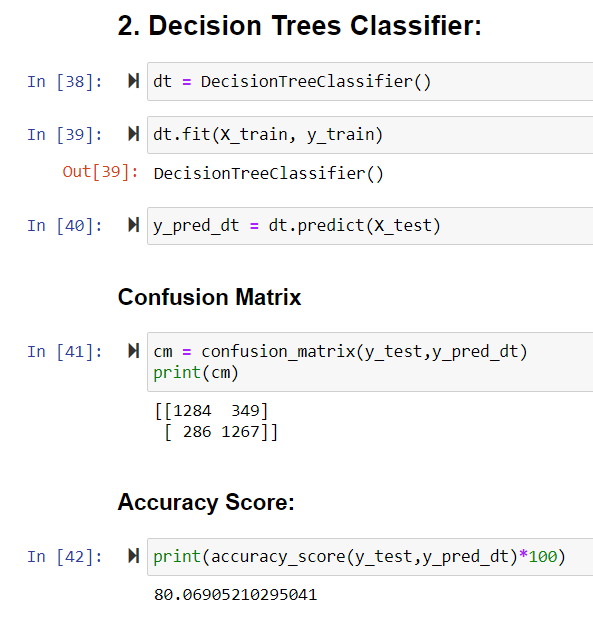
**Fig 36:** **Splitting the Data Into Train and Test**

1. Fitting the Required machine learning model using the algorithms.

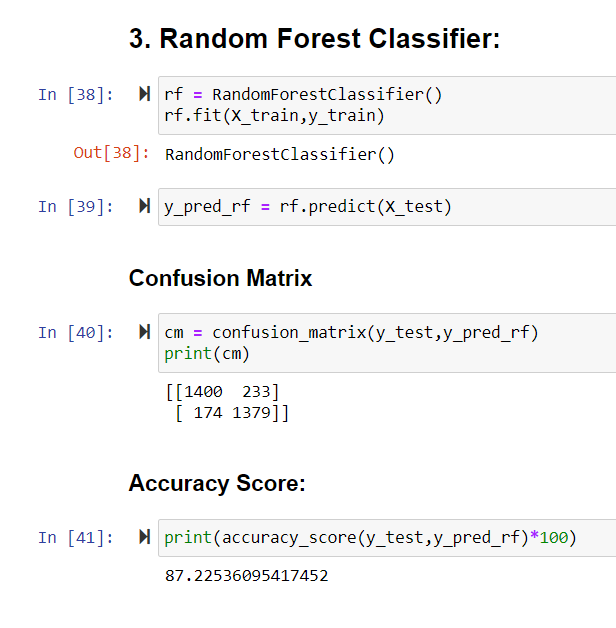
**5.4.1 Model Training, Testing and Evaluation:**



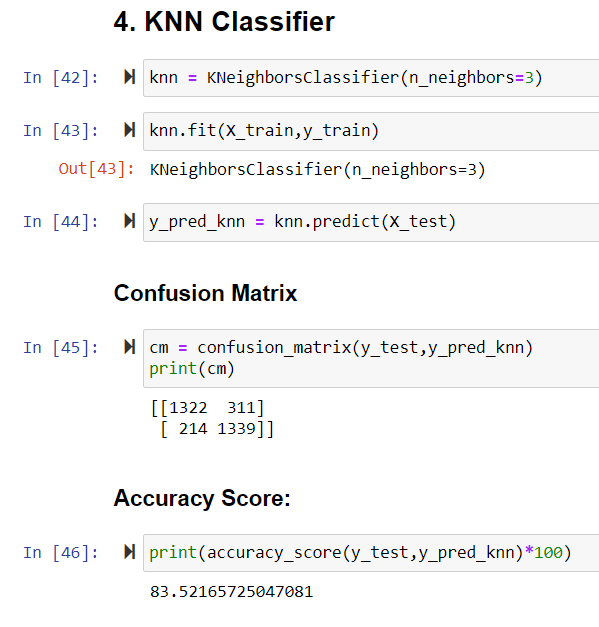
**Fig 37: Logistic Regression**



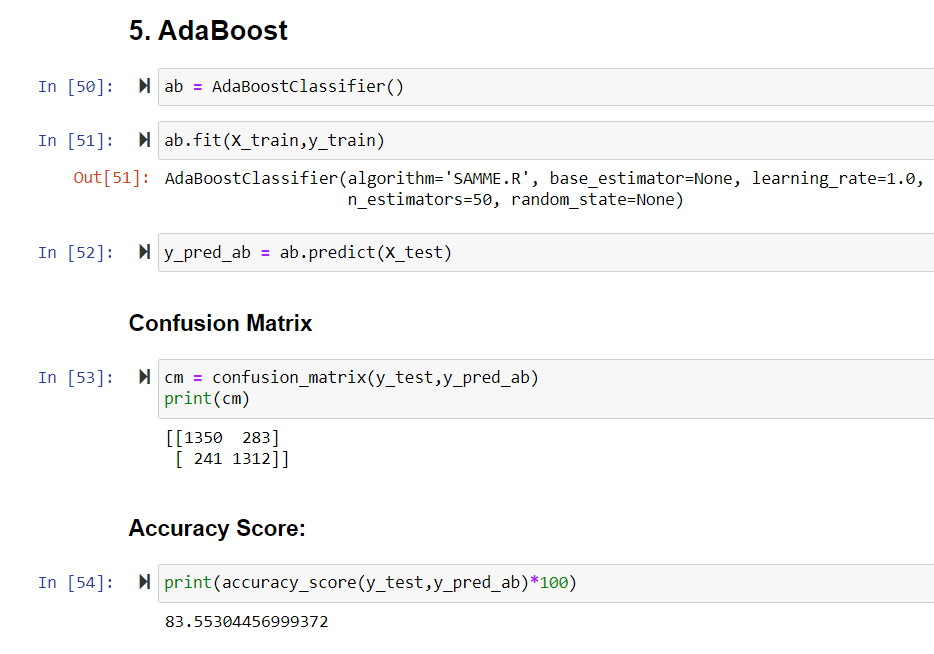
**Fig 38: Decision Tree Classifier**

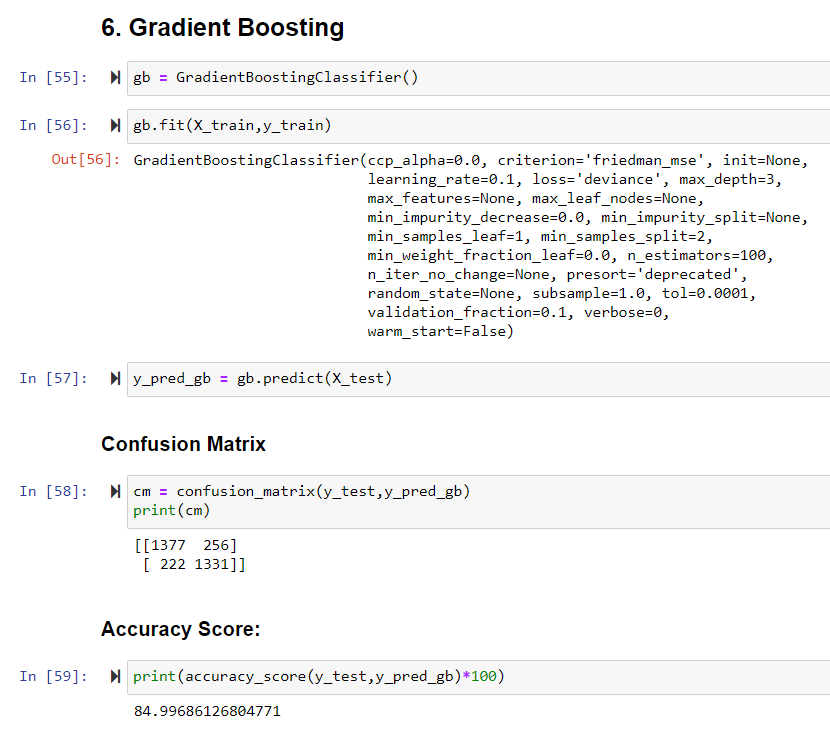


**Fig 39: Random Forest Classifier**

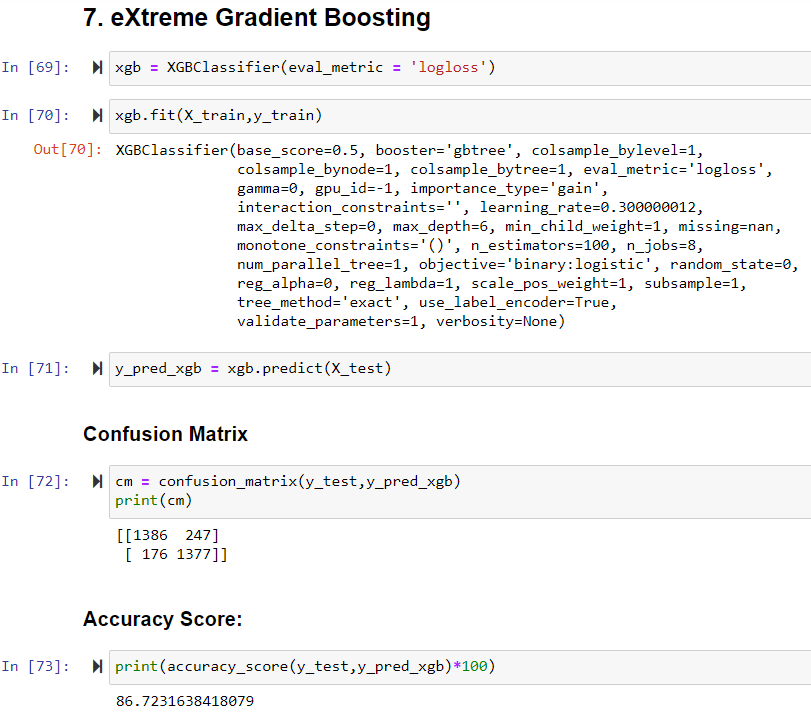


**Fig 40: KNN Classifier**

**Fig 41: AdaBoost Classifier**

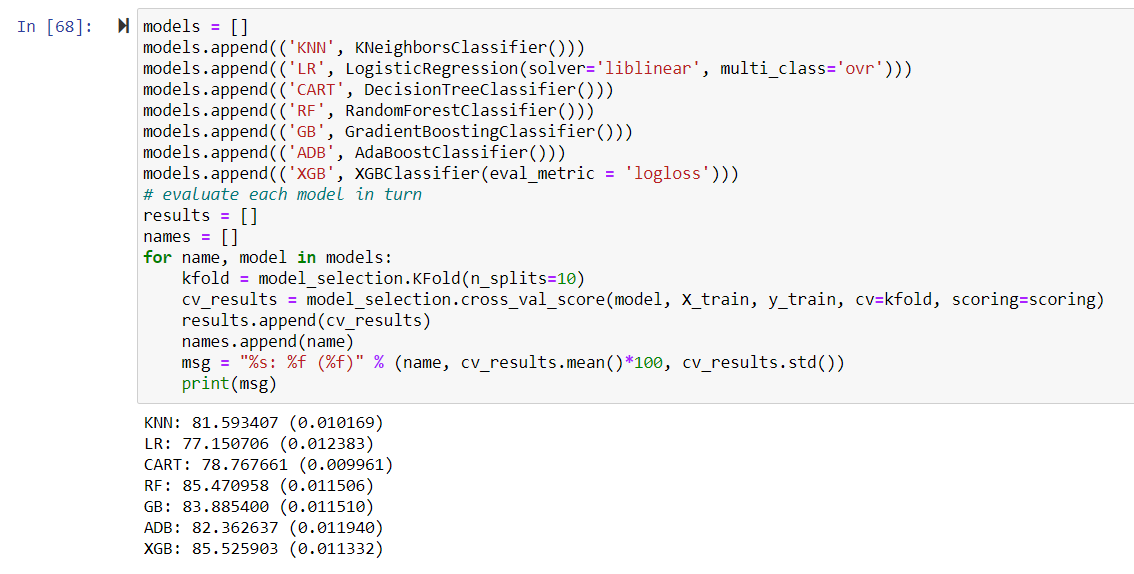
****

**Fig 42: Gradient Boosting Classifier**

**Fig 43: Extreme Gradient Boosting Classifier**

The evaluation Metrics used were Train/Test split validation as well as K fold Cross validation techniques.

Since these two are different validation techniques, there will be a slight deference in the accuracy scores of the same model.



**Fig 44: Accuracy Associated With Each Classifier**

**6.TESTING**

**6.1 INTRODUCTION**

In the proposed system, the dataset is prepared upon collection of data from various resources. The dataset contains data of hundreds of customer details which include their customer-id, credit score, location, gender, age, tenure with the bank, bank balance, their credit card status, whether he is an active member or not. These data is separated as the train and test data, where the train data is 80% of the complete data and the test data is of 20% of complete data.

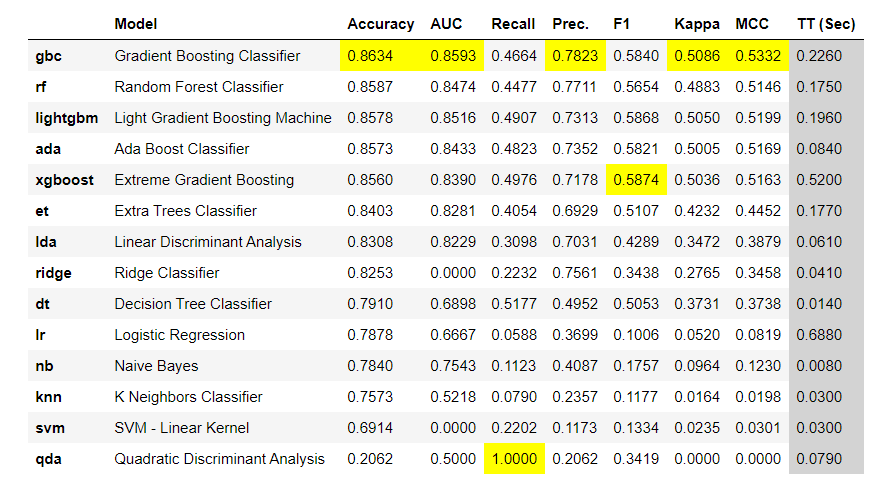
With the models used, the accuracies we obtained are:

|  |  |
| --- | --- |
| Models Used | Accuracy |
| Logistic Regression | 77.30% |
| Decision Tree Classifier | 78.93% |
| Random Forest Classifier | 87.22% |
| K-NN Classifier | 83.52% |
| AdaBoost Classifier | 83.55% |
| Gradient Boosting Classifier | 84.99% |
| eXtreme Gradient Boosting Classifier | 86.72% |

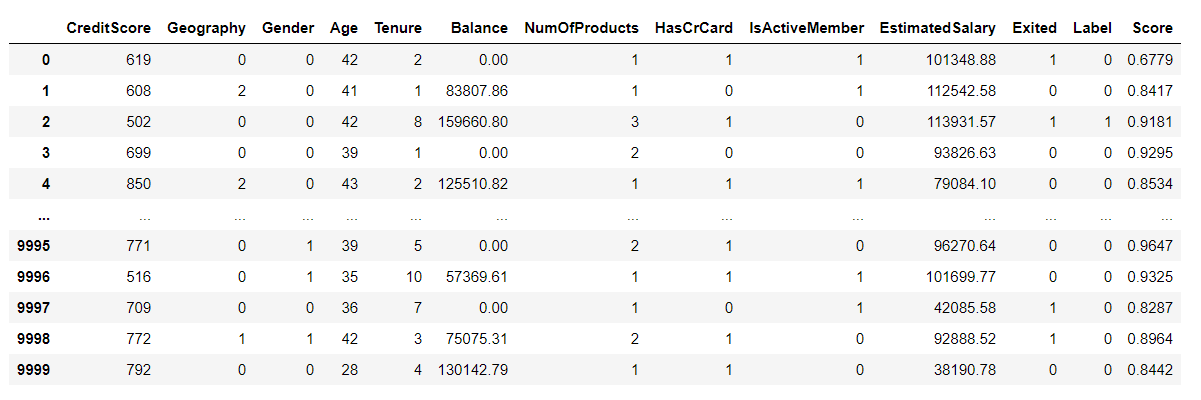
Out of these all the models used, Random Forest classifier gave us the highest accuracy of 87.2%.

But the validation technique used here was Train/Test split validation. We also checked with another validation technique.

LOOCV(Leave one out Cross Validation)

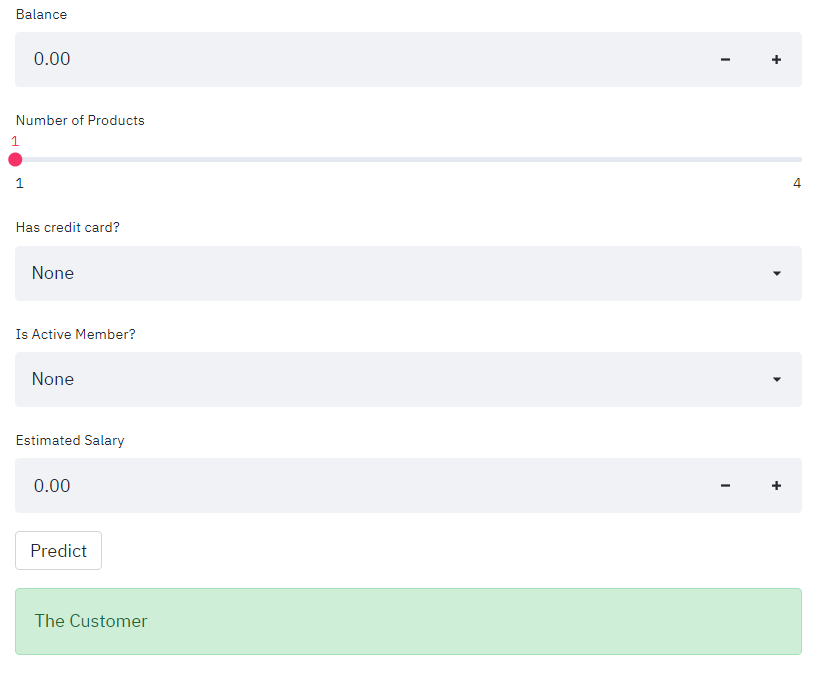
Using LOOCV, we obtained the following accuracy scores:

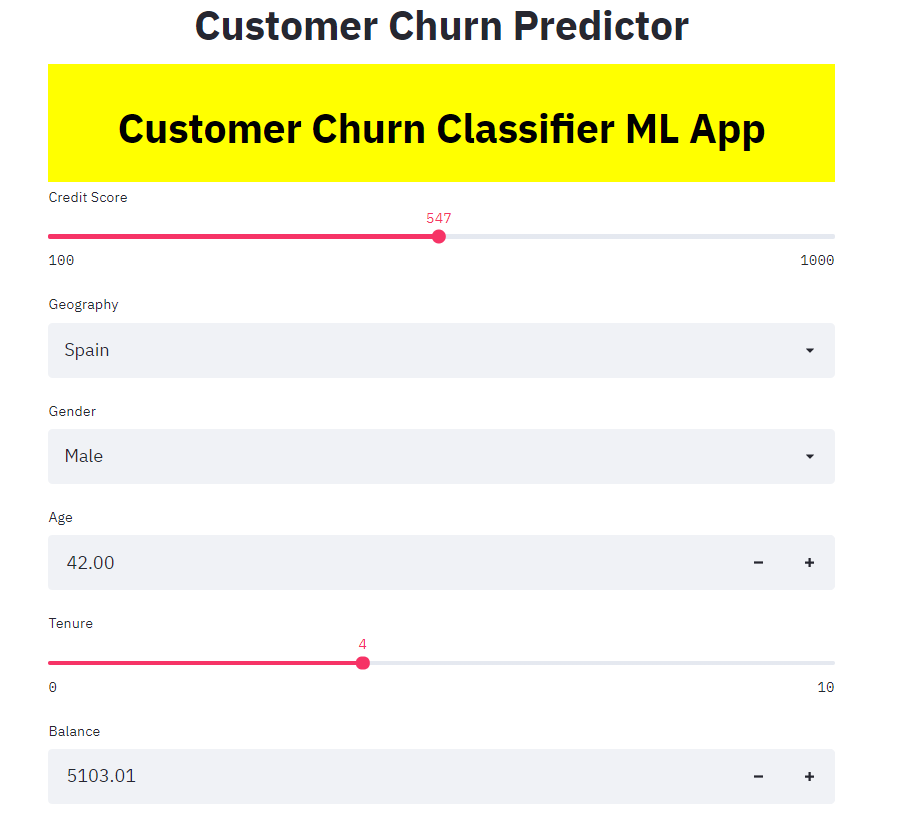
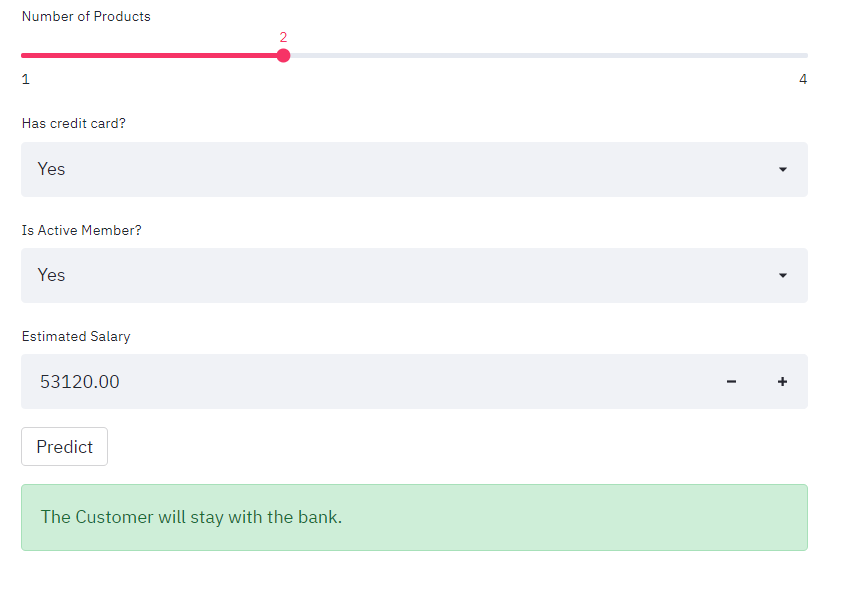
**Fig 45: Accuracy, Recall and Precision of Various ML Models**

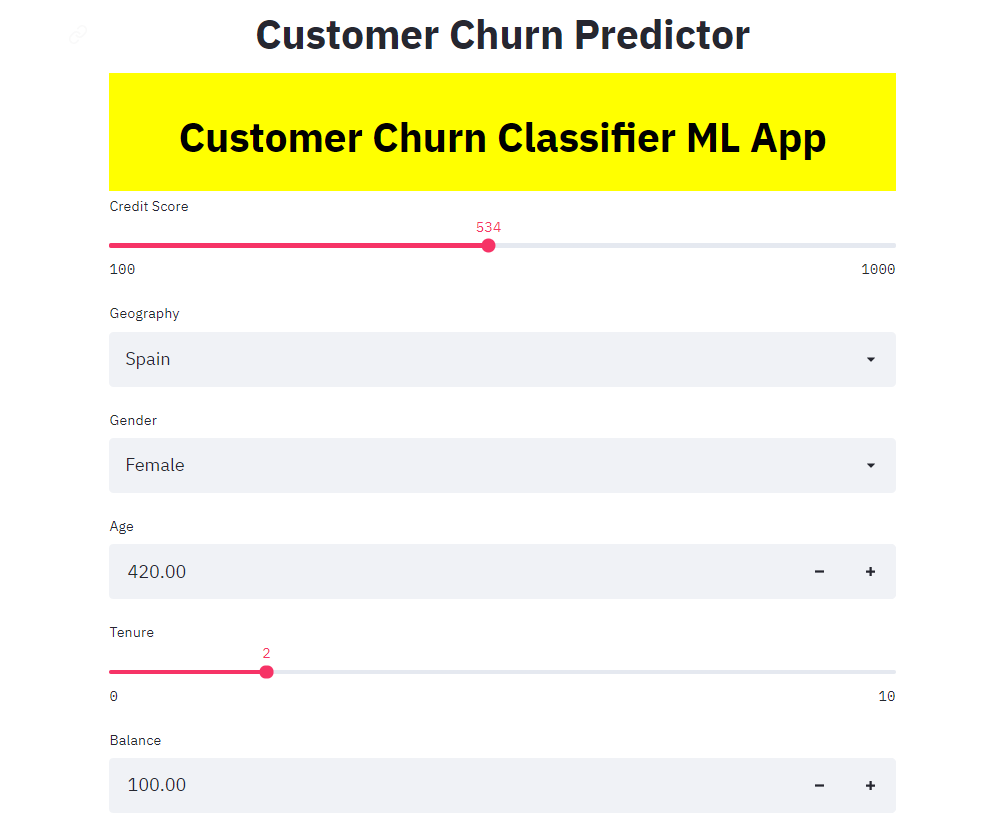
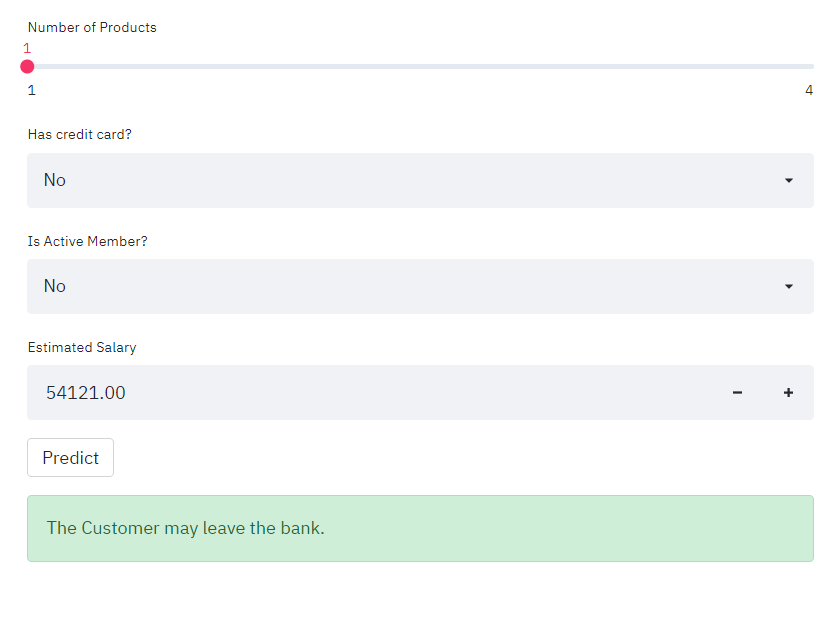
We also displayed the predictions in a tabular format:

**Fig 46: Predicted values after developing the model.**

Also, as part of our proposed system, we also designed a real-time web application in which the bank can feed in the customer details which in turn will help the bank to understand which customer is at a risk of being churned.



**Fig 47: Details Entered For A Customer, Who Will Stay With The Bank**



**Fig 48: Details Entered For A Customer, Who May Be Churned From The Bank**

**7.CONCLUSION**

Following an elongated period of time with bygone low interest rates, the banking sector around the world has been undergoing significant structural transformations in its business model to achieve profitability goals. On one side, revenue has been improved by the escalting and substantial rises in commissions in an attempt to resist low financial margins. On the other hand, the previously mentioned transformations have been driven primarily by a thought to shift towards digital, which has culiminated in the cutting down of operations and shutting down of numerous branches. Such decisions have had a strong affect on customer satisfaction, and predestined increase in churn rates.

This project intended to form a basis to address churn in a bank that is presently not utilizing its robust database and analytic applications to solve this critical problem. The initial step consisted in gathering a dataset of customers for a certain period of time, who would probably turn out to be churners during that period of time. The aim is to keep track of the behavior of these customers during the period of time given in the dataset, which would possibly be a representation of risk of being churned in the near future. As such, the choice of variables depended completely on the use of 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 in comparison to some other point in time. Pre-processing task was carried out with the intent of outlier elimination and data transformation, to which 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 classifier, random forest classifier, K-Neighbors classifier, AdaBoost classifier, Gradient boosting classifier and XGBoost classifier. Each of these models were trained on the same dataset consisting 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 random forest, extra trees, decision tree and logistic regression classifier scored an accuracy of 85%, 84%, 79% and 78% respectively.

At 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 on 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.

1. **FUTURE ENHANCEMENTS**

The current project aims solely on predicting churn by taking into account data from one single six-month time window. Hence, the results obtained may be biased when compared to datasets involving numerous time periods. In the end, the proposed procedure would be reinforced by a study gathering datasets for different time periods to further validate and verify the results presented in this work.

The seven machine learning techniques were used in this project on the obtained dataset. Further other algorithms can be explored and employed as well. Different machine learning algorithms can be explored, and data can be analyzed in a better way.

**9. REFERENCES**

[1] <https://www.programiz.com/python-programming>

[2] <https://realpython.com/jupyter-notebook-introduction/>

[3] <https://en.wikipedia.org/wiki/Pandas_(software)>

[4] <https://en.wikipedia.org/wiki/NumPy>

[5] <https://en.wikipedia.org/wiki/Matplotlib>

[6] <https://www.diva-portal.org/smash/get/diva2:1435454/FULLTEXT01.pdf>

[7] <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>

[8] <https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/>

[9] <https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825>

[10] <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>

[11] <https://www.mygreatlearning.com/blog/data-preprocessing/>

[12] <https://chartio.com/learn/charts/what-is-a-scatter-plot/>

[13] <https://www.klipfolio.com/resources/articles/what-is-data-visualization#DataViz1>

[14] <https://medium.com/analytics-vidhya/sampling-statistical-approach-in-machine-learning-4903c40ebf86>

[15] <https://www.mygreatlearning.com/blog/label-encoding-in-python/>

[16] <https://www.stitchdata.com/resources/data-transformation/>

[17] <https://medium.com/@danberdov/types-of-missing-data-902120fa4248>

[18] <https://www.questionpro.com/blog/data-collection/>

[19] <https://www.researchgate.net/publication/32040209_Data_Inspection_and_Data_Quality>

[20]<https://www.coursera.org/lecture/basic-data-processing-visualization-python/data-filtering-and-cleaning-zFhvv>