A Report on

BANK CHURN ANALYSIS USING MACHINE LEARNING

Submitted for partial fulfilment of award of

DEGREE
OF
BACHELOR OF COMPUTER APPLICATIONS

Submitted By

TANISHA
(210934106318)

Under the supervision of
Mr. Prateek Gupta

GHAZIABAD
(Estd : 1995)
NAAC Accredited 'A' Grade Institute
NBA Accredited MCA Programme

INSTITUTE OF TECHNOLOGY & SCIENCE
MOHAN NAGAR, GHAZIABAD

Batch: 2021-2024

ACKNOWLEDGEMENT

I would like to express my gratitude to the individuals who contributed to the completion of the "Bank Churn Analysis Using Machine Learning" project. My sincere appreciation goes to Mr. Prateek Gupta, whose guidance and insightful feedback were instrumental throughout the research process. I extend my thanks to the faculty of Institute of Technology & Science for providing supportive environment & valuable resources. I am indebted to my classmates for their engaging discussion. I want to express my heartfelt thanks to my family for their support during this project. Their belief in my abilities served as a constant source of motivation.

In conclusion, the success of this project is a testament to the collective efforts of these individuals.

TANISHA

ABSTRACT

Now a -days there are a lot of service providers available in every business. There is no shortage of customers in any options. Mainly, in the banking sector when people want to keep their money safe, they have a lot of options. As a result, customer churn and loyalty of customers have become a major problem for most banks. In this paper, a method that predicts customer churn in banking using Machine learning. This research promotes the exploration of the likelihood of churning by customer loyalty.

The Logistic regression, Random Forest, Decision tree and XG Boost Machine Learning algorithms are used in this study. This study is done on a dataset called churn modelling. The dataset was collected from Kaggle. The results are compared to find an appropriate model with higher accuracy. As a result, the XG Boost algorithm achieved higher accuracy than other algorithms. And accuracy was nearly 85%. The least accuracy was achieved by the Logistic Regression algorithm and it was 70% accuracy.

When the number of service providers is increasing very rapidly in every business. These days, there is no shortage of options for customers in the banking sector choosing where to put their money. As a result, customer churn and engagement have become one of the top issues for most banks. In this project, a method to predict the customer churn in a Bank, using machine learning techniques, which is a branch of artificial intelligence, is proposed. The research promotes the exploration of the likelihood of churn by analysing customer behaviour. Customer Churn has become a major problem in all industries including the banking industry and banks have always tried to track customer interaction so that they can detect the customers who are likely to leave the bank.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Churning means a customer who leaves one company and transfers to another company. It is not only a loss in income but also other negative effects on the operations and also mainly Customer Relation Management is very important for banking when the company considers it as they try to establish long-term relationships with customers and also it will lead to increase their customer base.

The service provider's challenges are found in the behaviour of the customer and their expectations. In the current generation, people are mostly educated compared to previous generations. So, the current generation of people is expecting more policies and their diverse demand for connectivity and innovation. This advanced knowledge is leading to changes in purchase behaviour. This is a big challenge for current service providers to think innovatively to reach their expectations.

Private sectors need to recognize customers Liu and Shih strengthen this argument in their paper by indicating that increasing pressures on companies to develop new and innovative ideas in marketing, to meet customer expectations and increase loyalty and retention.

For Customers, it is very easy to transfer their relations from one bank to another bank. Some customers might be keeping their relationship status null that means they will keep their account status inactive. By keeping this account inactive it might be the customer transferring their relationship with another bank. There are different types of customers are in the bank. Farmers are one of the major customers to the banks they will expect fewer monthly chargers as they were financially low. Businessperson, are also one of the major and important customers because a lot of transactions with huge amount is done by them only usually. These customers will expect better service consider customers and their needs to resolve these challenges delivering reliable service on 5 time and within budget to customer.

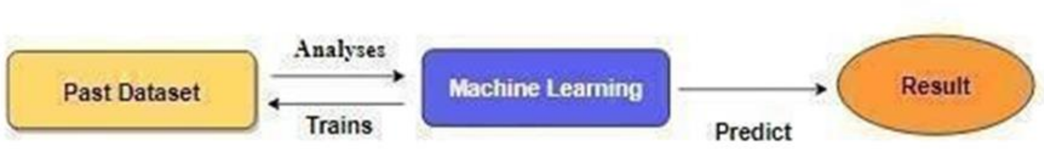
While maintaining a good working partnership with them is another significant challenge for them. If they failed to resolve these challenges this may cause churning. Recruiting a new customer is more expensive and harder than keeping already customers.

Customers quality, one of the most important categories was Middle-class customers, mostly in every bank these peoples are more than the type of customers. These people will expect fewer monthly charges, better service quality, and new policies.

So, maintaining different types of customers is not that easy. They need to holding on the other hand is usually more expensive because they have already gained the confidence and loyalty of present customers. So, the need for a system that can predict customer churn effectively in the early stages is very important for any banking. This paper aims at a framework that can predict the customer churning banking sectors using some machine learning algorithms.

1.1 MACHINE LEARNING

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. This technology finds applications across various domains, including image and speech recognition, natural language processing, recommendation systems, and predictive analytics. By leveraging large datasets, machine learning algorithms can discover complex patterns and insights, contributing to advancements in automation, efficiency, and decision-making processes in fields ranging from healthcare and finance to education and beyond. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves the use of specialized algorithms. It feeds the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data.



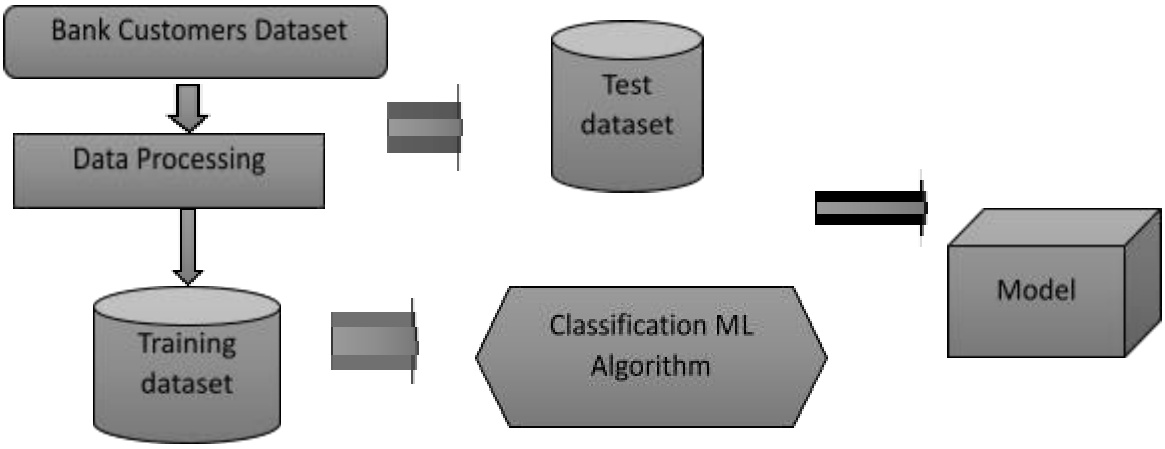
Process Of Machine Learning

1.2 PROPOSED SYSTEM

The proposed method is to build a Bank Customer Churn prediction using Machine learning Technique. We are going to develop an AI based model, we need data to train our model. We can use Bank Customer dataset in order to train the model.

To use this dataset, we need to understand what the intents that we are going to train. An intent is the intention of the user interacting with a predictive model or the intention behind each Data that the Model receives from a particular user. According to the domain that you are developing an AI solution, these intents may vary from one solution to another.

The strategy is to define different intents and make training samples for those intents and train your AI model with those training sample data as model training data and intents as model training categories. The model is build using the process of vectorisation where the vectors made to understand the data. To use different algorithm, we can get a better AI model and best accuracy. After building a model we evaluate the model using different metrics like confusion metrics, precision, recall, sensitivity and F1 scorer.



1.3 OBJECTIVE

The goal is to develop a machine learning model for Bank Churn Prediction, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithms.

CHAPTER 2

LITERATURE SURVEY

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is a secondary source and discusses published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a reorganization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them

A comparison of machine learning techniques for customer churn prediction Praveen Asthana 2018 We present a comparative study on the most popular machine learning methods applied to the challenging problem of customer churning prediction in the

telecommunications industry. In the first phase of our experiments, all models were applied and evaluated using cross-validation on a

popular, public domain dataset. In the second phase, the performance improvement offered by boosting was studied. In order to determine the most efficient parameter combinations we performed a series of Monte Carlo simulations for each method and for a wide range of parameters. Our results demonstrate clear superiority of the boosted versions of the models against the plain (non-boosted) versions. The best overall classifier was the SVMPOLY using AdaBoost with accuracy of almost 97% and F-measure over 84%. Customer Churn Analysis in Banking Sector G. Jignesh Chowdary1, Suganya. G 2, Premalatha. M32019 The role of ICT

in the banking sector is a crucial part of the development of nations. The development of the banking sector mostly depends on its valuable customers. So, customer churn analysis is needed to determine customers whether they are at risk of leaving or worth retaining. From an organizational point of view, gaining new

customers are usually more difficult or more expensive than retaining existing customers. So, customer churn prediction has been popular in the banking industry. By reducing customer churn or attrition, the commercial banks gain not only more profits but also enhancing core competitiveness among the competitors. Although many researchers proposed many single prediction models and some hybrid models, accuracy is still weak and computation time of some algorithms is still increased. In this research, the churn prediction model of classifying bank customers is built by using the hybrid

model of k-means and Support Vector Machine data mining

methods on bank customer churn dataset to overcome the instability and limitations of single prediction model and predict churn trend of high value users.

Developing a prediction model for customer churn from electronic banking services using data mining Abbas Karamat, Hajar Ganei and Seyed Mohammad Mir Mohammadi. 2016 Given the importance of customers as the most valuable assets of

organizations, customer retention seems to be an essential, basic requirement for any organization. Banks are no exception to this rule. The competitive atmosphere within which electronic banking services are provided by different banks increases the necessity of customer retention. Methods: Being based on existing information technologies which allow one to collect data from organizations' databases, data mining introduces a powerful tool for the extraction of knowledge from huge amounts of data. In this research, the decision tree technique was applied to build a model incorporating this knowledge. Results: The results represent the characteristics of churned customers.

Conclusions: Bank managers can identify churners in future using the results of the decision tree. They should be provided some strategies for customers whose features are getting more likely to churner's features.

A Critical Examination of Different Models for Customer Churn Prediction using Data Mining Seema, Gaurav Gupta 2019 Due to competition between online retailers, the need for providing improved customer service has grown rapidly. In addition to reduction in sales due to loss of customers, more investments are needed to be done to attract new customers. Companies now are working continuously to improve their perceived quality by way of giving timely and quality service to their customers. Customer churn has become one of the primary challenges that many firms are facing nowadays.

Several churn prediction models and techniques are proposed previously in literature to predict customer churn in areas such as finance, telecom, banking etc. Researchers are also working on customer churn prediction in ecommerce using data mining and machine learning techniques. In this paper, a comprehensive review of various models to predict customer churn in ecommerce data mining and machine learning techniques has been presented.

A critical review of recent research papers in the field of customer churn prediction in e-commerce using data mining has been done. Thereafter, important inferences and research gaps after studying the literature are presented. Finally, the research significance and concluding remarks are described in the end. bank customer retention prediction and customer ranking based on deep neural networks Dr A.P. Jagadeesan, Ph.D., 2020 Retention of customers is a major concern in any industry. Customer churn is an important metric that gives the hard truth about the retention percentage of customers. A detailed study about the existing models for predicting the customer churn is made and a new model based on Artificial Neural Network is proposed to find the customer churn in banking domain.

The proposed model is compared with the existing machine learning models. Logistic regression, Decision Tree, XG Boost and Random Forest mechanisms are the baseline models that are used for comparison, the performance metrics that were compared are accuracy, precision, recall and F1 score. It has been observed that the artificial neural network model performs better than the logistic regression model, random forest model and decision tree model. But when the results are compared with the XG Boost model considerable difference is not noted. The proposed model differs from the existing models in a way that it can rank the customers in the order in which they would leave the organization.

CHAPTER 3

METHODOLOGY

This section explains the various works that have been done in order to predict the customer churn. It includes totally 5 modules are using in this project they are as follows:

3.1 MODULE IMPLEMENTATION

3.1.1 Collecting Bank Churning Data:

The bank churn dataset used in this analysis was collected from Kaggle. This dataset includes information of 10000 bank customers. It contains 13 features; they are mentioned below:

(1) Customer Id: Id of a customer, it is a unique identification code that is usually got from the respective bank.

(2) Surname: Name of the customer or the customer's first name.

(3) Credit Score: It is the measure of independent capacity to pay back the borrowed amount. It is a numerical representation of their creditability.

(4) Geography: In this dataset, there 3 geographical areas were mentioned France, Spain, and Germany.

(5) Gender: Gender of a customer. In this dataset, it is mentioned as a string (Female and Male).

(6) Age: Age of the customer.

(7) Tenure: Length of time that will be taken by the lender to repay their loan along with the interest.

(8) Balance: Amount of money present in the account either it can be a savings account or salary account.

(9) Number of Products: Various services offered by a bank to the customers. The number of services that a customer is using is mentioned in the dataset and the objective of this is int.

(10) Has Cr Card: Does a customer contain the credit card or not? If yes, it is mentioned as 1, and if no it is mentioned as 0.

(11) Is Active Member: Does a customer is active or not? It is decided by their transactions. In this dataset, it is mentioned as 1 if the customer was active. If the customer was not active it is mentioned as 0.

(12) Estimated Salary: Salary of a customer (in rupees).

(13) Exited: Whether the customer is excited or not. It is also mentioned in binary values. If exited it is mentioned as 1 if not it is mentioned a 0.

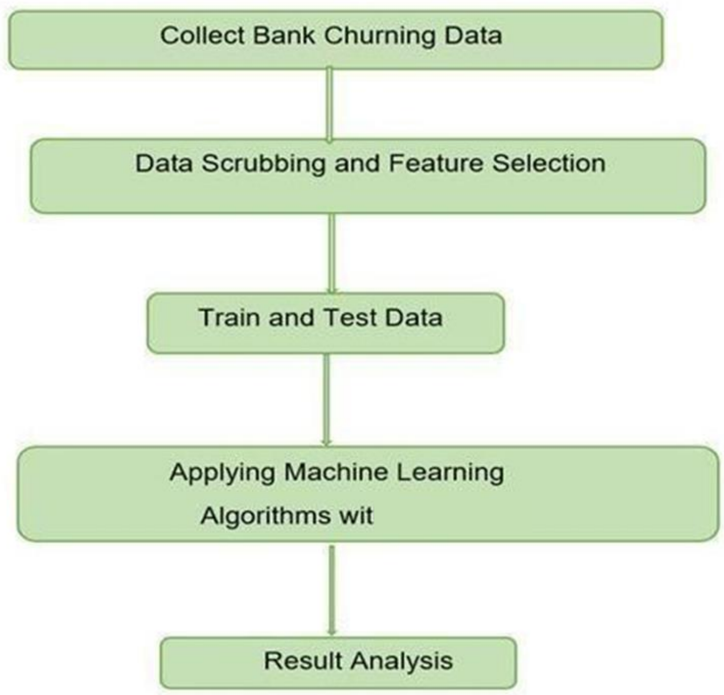
3.1.2 Data Scrubbing and Feature Selection:

Data Scrubbing: Removing, correcting, or identifying the errors in the dataset. Which may impact the predictive model.

Feature Selection: To reduce the number of input features. Selecting wanted features and removing the unwanted features.

By applying, Data scrubbing and feature selection will improve the data's consistency and accuracy. The selection phase is very important it will help us to reduce the training and skip the high-dimensional curse.

Transformation: Data transformation is the data was available may not be in the right format or may need transformations to make it more useful. By removing null values, unwanted duplicates, incorrect indexing, and incompatible formats. In this dataset Credit Score, Age, Tenure, Balance, the Estimated salary was big value. By using MinMaxScaler these values normalized or transformed into the range [0,1]. Columns like Row number, Surname, Customer Id are unwanted columns.


Collect Bank Churning Data
Data Scrubbing and Feature Selection
Train and Test Data
Applying Machine Learning Algorithms wit
Result Analysis

3.1.3 Train and Test Data:

Training and testing data by applying different Machine learning algorithms like SVM, KNN, Decision Tree, Random Forest, Logistic regression and XG Boost. The dataset is going to be split into two as train and test datasets by using the train_test_split() method. It is a quick and simple procedure to perform, the result of which allows us to compare the performance of Machine Learning algorithms for predictive modelling problems. The training dataset was used to fit the machine learning model and a test dataset was used to evaluate and fit the Machine learning model. To control the size of the train and test dataset with parameters train_size and test_size respectively. Either the train_size or test_size should be included in train_test_split() method. Over all the dataset was 1 then the dataset is split into 2 subsets as train and test. Then train_size should be more than the test_size.

3.1.4 Applying ML Algorithms:

Decision Tree Algorithm:

Decision tree (DT) is one of the supervised learning algorithms used for both regression and classification. Data in the dataset is constant split according to a certain parameter. This tree can be classified by entities namely decision nodes and leaves. The leaves are the outcomes. A simple way to implement the decision tree algorithm is by using SCIKIT learn library and calling related functions to implement the Decision Tree model. Decision Tree classifier is used to fit the data provide returning the best fit.

Random Forest Algorithm:

Random Forest is also one of the most popular supervised learning techniques. It is used for both classification and regression. It builds multiple decision trees and merges them to get more accurate. Compare to all algorithms in this model Random Forest gives more accuracy. By using the SCIKIT learn library Random Forest can be implemented easily. By calling Random Forest classifier that will ensemble learning method for regression and calling some related functions to implement Random Forest.

Logistic Regression Algorithm:

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of

logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent

variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

XG Boost:

XG Boost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XG Boost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.

3.2 LIBRARIES

Pandas:

It is a Popular Python based data analysis toolkit which can be imported using import pandas as PD. It represents a diverse range of utilities, ranging from parsing multiple file-formats to converting an entire data table into a NumPy matrix array.

NumPy:

It is a python library used for working with arrays. It also has functions for working with an array. It also has functions for working in the domain of linear algebra, Fourier transform & matrices.

Seaborn:

It is a library for making statistical graphics in python. It builds on top of matplotlib & integrates closely with Pandas data structures. It helps us to explore & understood our data. Its plotting functions operate on Data frames & arrays containing whole datasets & internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

Matplotlib:

It is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

Scikit-learn:

It is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering, and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy, and Matplotlib.

3.3 PROJECT REQUIREMENTS

1. Functional requirements
2. Non-Functional
3. Environment requirements
    a. Hardware requirement
    b. Software requirement

Functional Requirements:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details follow the special libraries like SK-learn, pandas, numpy, matplotlib and seaborn.

Non-Functional Requirements:

Process of functional steps:

1. Problem defining
2. Preparing data

3. Evaluating algorithms
4. Improving results
5. Prediction the result

Software Requirements:

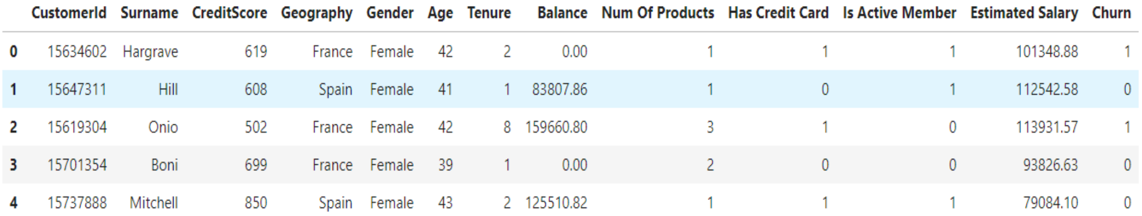
Operating System	:	Windows
Tool	:	Anaconda with Jupyter Notebook

CHAPTER 4

RESULT & DISCUSSION

4.1 Data Pre-processing:

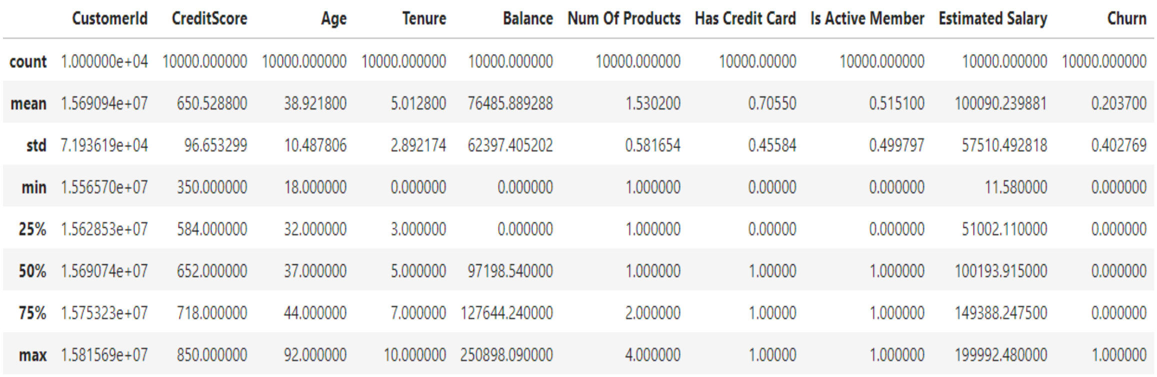
Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.



	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	Num Of Products	Has Credit Card	Is Active Member	Estimated Salary	Churn
0	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

A number of different data cleaning tasks using Python's Pandas library and specifically, it focuses on probably the biggest data cleaning task, missing values and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modelling.

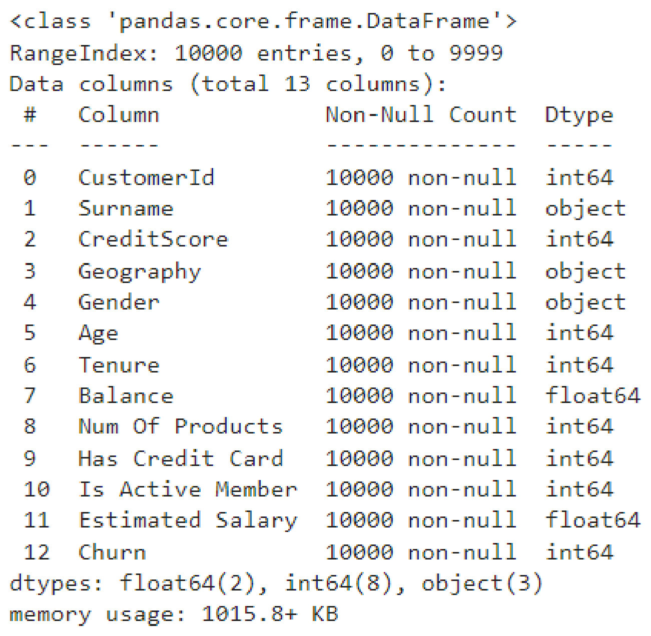


	CustomerId	CreditScore	Age	Tenure	Balance	Num Of Products	Has Credit Card	Is Active Member	Estimated Salary	Churn
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing.

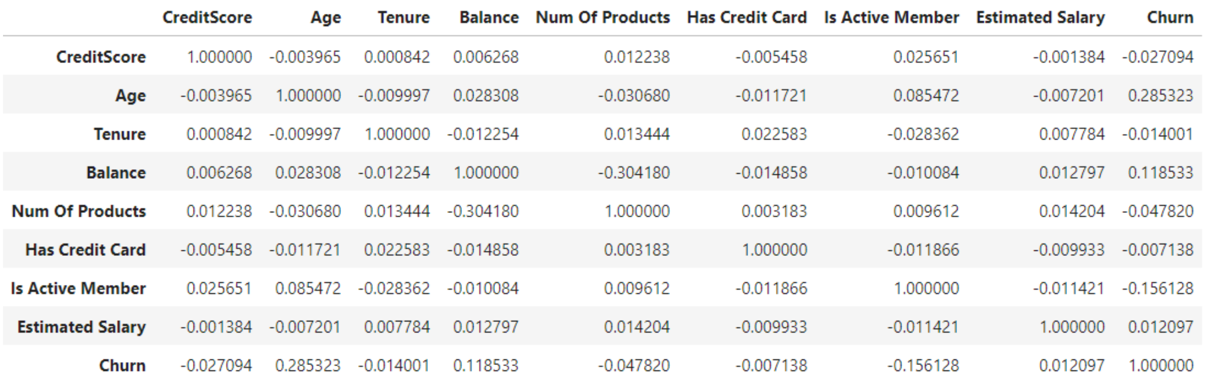
It’s important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data.

Before, joint into code, it is important to understand the sources of missing data. Here are some typical reasons why some data is missing.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   CustomerId         10000 non-null  int64
 1   Surname            10000 non-null  object
 2   CreditScore        10000 non-null  int64
 3   Geography          10000 non-null  object
 4   Gender             10000 non-null  object
 5   Age                10000 non-null  int64
 6   Tenure             10000 non-null  int64
 7   Balance            10000 non-null  float64
 8   Num Of Products    10000 non-null  int64
 9   Has Credit Card    10000 non-null  int64
 10  Is Active Member   10000 non-null  int64
 11  Estimated Salary   10000 non-null  float64
 12  Churn              10000 non-null  int64
dtypes: float64(2), int64(8), object(3)
memory usage: 1015.8+ KB
```

- ➢ User forgot to fill in a field.
- ➢ Data was lost while transferring manually from a legacy database.
- ➢ There was a programming error.
- ➢ Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.



	CreditScore	Age	Tenure	Balance	Num Of Products	Has Credit Card	Is Active Member	Estimated Salary	Churn
CreditScore	1.000000	-0.003965	0.000842	0.006268	0.012238	-0.005458	0.025651	-0.001384	-0.027094
Age	-0.003965	1.000000	-0.009997	0.028308	-0.030680	-0.011721	0.085472	-0.007201	0.285323
Tenure	0.000842	-0.009997	1.000000	-0.012254	0.013444	0.022583	-0.028362	0.007784	-0.014001
Balance	0.006268	0.028308	-0.012254	1.000000	-0.304180	-0.014858	-0.010084	0.012797	0.118533
Num Of Products	0.012238	-0.030680	0.013444	-0.304180	1.000000	0.003183	0.009612	0.014204	-0.047820
Has Credit Card	-0.005458	-0.011721	0.022583	-0.014858	0.003183	1.000000	-0.011866	-0.009933	-0.007138
Is Active Member	0.025651	0.085472	-0.028362	-0.010084	0.009612	-0.011866	1.000000	-0.011421	-0.156128
Estimated Salary	-0.001384	-0.007201	0.007784	0.012797	0.014204	-0.009933	-0.011421	1.000000	0.012097
Churn	-0.027094	0.285323	-0.014001	0.118533	-0.047820	-0.007138	-0.156128	0.012097	1.000000

Variable identification with Uni-variate, Bi-variate and Multivariate analysis:

Import libraries for access and functional purpose and read the given dataset

General Properties of Analysing the given dataset:

Display the given dataset in the form of data frame show columns shape.

To describe the data frame.

Checking data type and information about dataset.

Checking for duplicate data.

Checking Missing values of data frame.

Checking unique values of data frame.

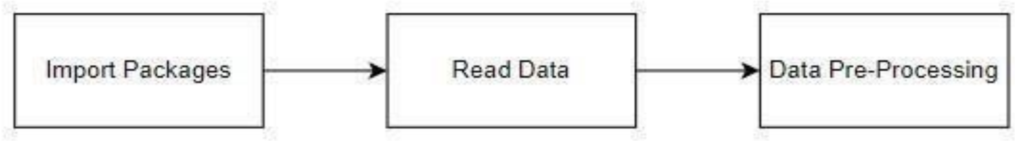
Checking count values of data frame.

Rename and drop the given data frame.

To specify the type of values.

To create extra columns.

Module diagram:

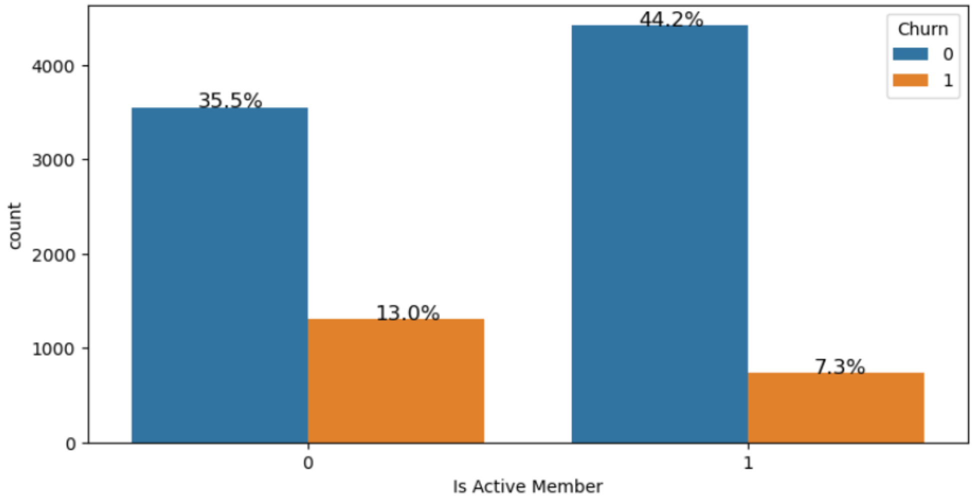


4.2 Data Validation/Cleaning/Preparing Process:

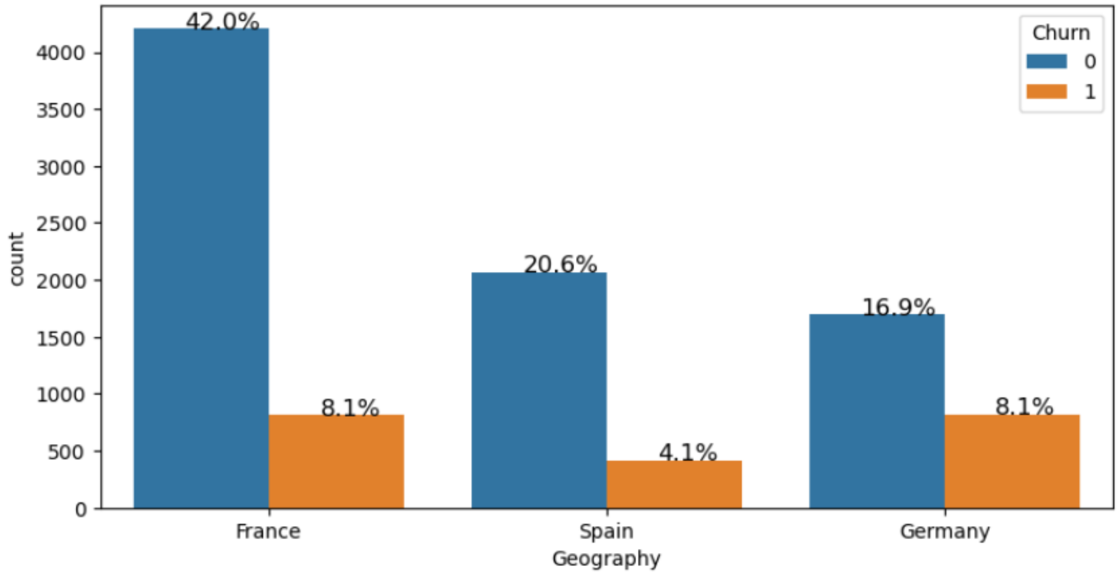
Importing the library packages with loading given dataset. To analyse the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning/preparing by rename the given dataset and drop the column etc. to analyse the Uni-variate, bivariate and multivariate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

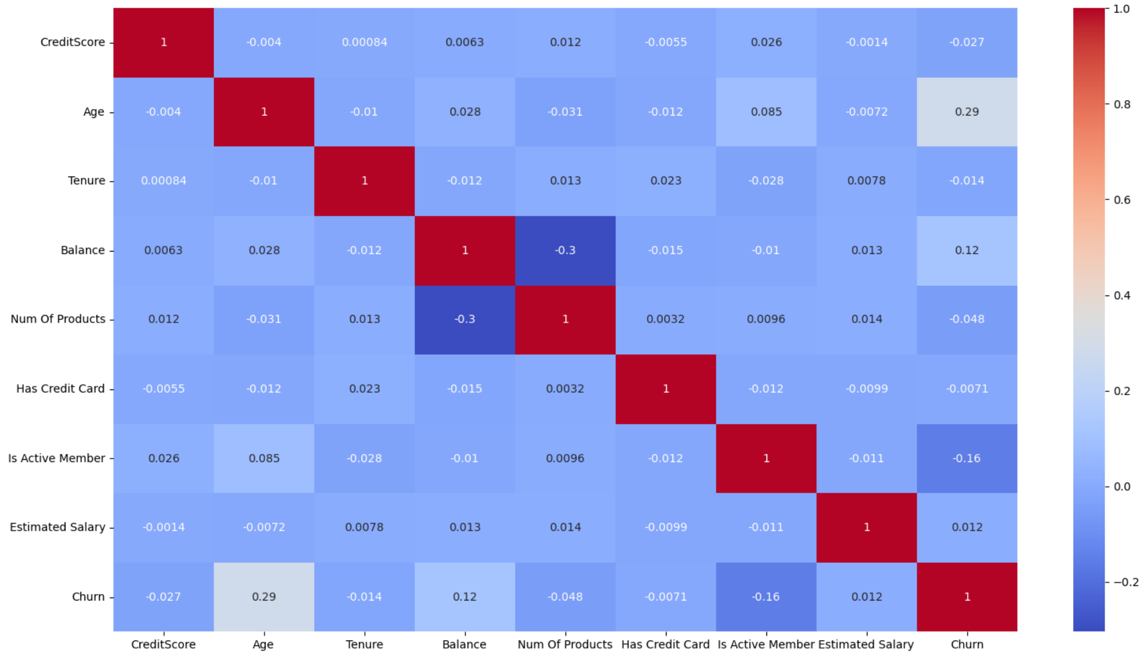
Exploration data analysis of visualisation:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding.



This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.





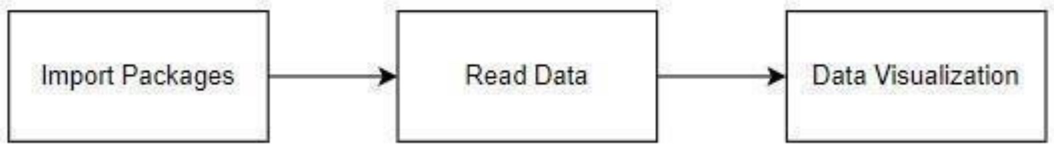
	CreditScore	Age	Tenure	Balance	Num Of Products	Has Credit Card	Is Active Member	Estimated Salary	Churn
CreditScore	1	-0.004	0.00084	0.0063	0.012	-0.0055	0.026	-0.0014	-0.027
Age	-0.004	1	-0.01	0.028	-0.031	-0.012	0.085	-0.0072	0.29
Tenure	0.00084	-0.01	1	-0.012	0.013	0.023	-0.028	0.0078	-0.014
Balance	0.0063	0.028	-0.012	1	-0.3	-0.015	-0.01	0.013	0.12
Num Of Products	0.012	-0.031	0.013	-0.3	1	0.0032	0.0096	0.014	-0.048
Has Credit Card	-0.0055	-0.012	0.023	-0.015	0.0032	1	-0.012	-0.0099	-0.0071
Is Active Member	0.026	0.085	-0.028	-0.01	0.0096	-0.012	1	-0.011	-0.16
Estimated Salary	-0.0014	-0.0072	0.0078	0.013	0.014	-0.0099	-0.011	1	0.012
Churn	-0.027	0.29	-0.014	0.12	-0.048	-0.0071	-0.16	0.012	1

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

How to chart time series data with line plots and categorical quantities with bar charts.

How to summarize data distributions with histograms and box plots.

Module Diagram:



Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in a given dataset.

False Positives (FP): A person who will pay is predicted as a defaulter. When the actual class is no and the predicted class is yes. E.g. if the actual class says this passenger did not survive but the predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When the actual class is yes but the predicted class is no. E.g. if the actual class value indicates that this passenger survived and the predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay is predicted as a defaulter. These are the correctly predicted positive values which means that the value of the actual class is yes and the value of predicted class is also yes.

E.g. if the actual class value indicates that this passenger survived and the predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. If the actual class says this passenger did not survive and the predicted class tells you the same thing.

Comparing Algorithm with prediction in the form of best accuracy result:

It is important to compare the performance of multiple different machine learning algorithms consistently and to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When having a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

In the example below 4 different algorithms are compared:

Logistic Regression

Random Forest

Decision Tree Classifier

XG Boost

The K-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. Before comparing algorithms, build a Machine Learning Model using Scikit-Learn libraries. In this library package, we have to do preprocessing, linear model with logistic regression method, cross validating by K Fold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predict the result by comparing accuracy.

Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It needs the output of the algorithm to be classified variable data. Higher accuracy predicting the result is a logistic regression model by comparing the best accuracy.

True Positive Rate (TPR) = TP / (TP + FN)

False Positive Rate (FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positives and false negatives are almost the same.

Precision: The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that ar labelled as survived, how many actually survived?

High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted.

(The proportion of actual defaulters that the model will correctly predict)

Recall = TP / (TP + FN)

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false

negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost.

If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula:

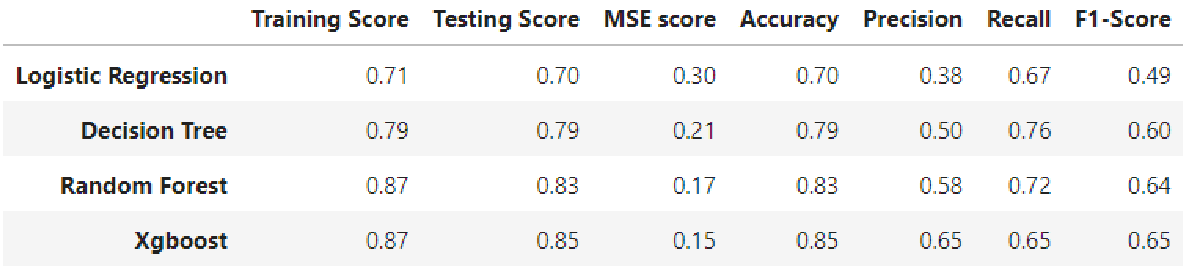
F- Measure = 2TP / (2TP + FP + FN) F1

Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

On comparing the four models used in this, we get:

The results are compared to find an appropriate model with higher accuracy. As a result, the XG Boost algorithm achieved higher accuracy than other algorithms. And accuracy was nearly 85%. The least accuracy was achieved by the Logistic Regression algorithm and it was 70% accuracy.



	Training Score	Testing Score	MSE score	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.71	0.70	0.30	0.70	0.38	0.67	0.49
Decision Tree	0.79	0.79	0.21	0.79	0.50	0.76	0.60
Random Forest	0.87	0.83	0.17	0.83	0.58	0.72	0.64
Xgboost	0.87	0.85	0.15	0.85	0.65	0.65	0.65

CHAPTER 5

FUTURE SCOPE OF THE PROJECT

Here the scope of the project is that integration of Bank support with computer-based records could reduce, enhance Bank safety, decrease the customer churn, and improve Bank Customer Support. This suggestion is promising as data modelling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of Bank support.