Bank Customer Segmentation using Different Machine and Deep Learning Methods

Submitted by

Niloy Mitra
ID: CSE 018 06713

A thesis submitted in conformity with the requirements for the degree of B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering

Port City International University
7-14, Nikunja Housing Society, South Khulshi,
Chattogram,Bangladesh

May 2023

Bank Customer Segmentation using Different Machine and Deep Learning Methods

Submitted by

Niloy Mitra
ID: CSE 018 06713

Under the supervision of

Mrs. Manoara Begum

Assistant Professor
Department of CSE
Port City International University

Department of Computer Science and Engineering

Port City International University
7-14, Nikunja Housing Society, South Khulshi,
Chattogram,Bangladesh

May 2023

14 May 2023

Publisher
Port City International University
7-14, Nikunja Housing Society, South Khulshi, Chattogram.
www.portcity.edu.bd

Dedication

"I want to dedicate this project to my parents for their full support of my study, without their support I couldn't be able to continue my study and supporting me all time And giving me hope.
I also dedicate this project to my honorable teacher ***Mrs. Manoara Begum*** *for supervising me in this project."*

Recommendation

This is to certify that **Niloy Mitra (CSE 018 06713)**, a student of Port City International University (PCIU) under the department of Computer Science and Engineering (CSE), had carried out the thesis entitled "**Bank Customer Segmentation Using Different Machine And Deep Learning Methods**" successfully under my supervision and guidance. To the best of my knowledge, the matter embodied in this dissertation has not been submitted to any other university/institute for the award of any degree.

(Signature of the Supervisor)
Mrs. Manoara Begum
Assistant Professor
Dept. of CSE, PCIU
Cell: 01518356201, 01881075015
Email: manoara.cse@portcity.edu.bd, manoara.cse34@gmail.com

Declaration of Authorship

We declare that this thesis entitled "**Bank Customer Segmentation Using Different Machine And Deep Learning Methods**" and the works presented in it are our own.

We confirm that:

- Any part of this thesis has not previously been submitted for a degree or any other qualification in this university or any other institution
- This thesis work is done entirely by us for our final undergraduate thesis.
- We have consulted the published works of others with appropriate references.

(Signature of the Supervisor)
Mrs. Manoara Begum
Assistant Professor
Dept. of CSE, PCIU
Cell: 01518356201, 01881075015
Email: manoara.cse@portcity.edu.bd, manoara.cse34@gmail.com

Acknowledgment

By the grace of Almighty God who is the most gracious and merciful, we have accomplished this work. We would like to express our deep appreciation and sincere gratitude to our honorable supervisor **Mrs. Manoara Begum** for her immense support throughout the entire work. Her valuable guidance and intuitive knowledge in this field helped and motivated us in writing this thesis.

We would also like to thank all the faculty members of the **Department of Computer Science and Engineering, Port City International University**, for helping us with all the necessary support.

Niloy Mitra

ID: CSE 018 06713

Batch: CSE-018

Port City International University

Abstract

Customer segmentation is the approach of dividing a large and diverse customer base into smaller groups of related customers that are similar in certain ways and relevant to the marketing of a bank's products and services. This is important step of Bank. This research mainly focuses on customer or not in Banking System. This data contains valuable information. Hence, it is very important to store, process, manage and analyze this data to extract knowledge from it. It helps to increase business profit. Banking industry plays very important role in economy of country. Customers are the main asset of the bank. The research aims to use a machine and deep learning algorithm to estimate predict on the entire Bank customer data whether the customer is in Bank customer or not. It can be help us to higher accuracy prediction by doing this research. It reveales that higher the level responsiveness and reliability higher would be the customer. We got the highest accuracy on my research. Using Bank Customer, 42639 data were classified as customer is in Bank customer or not in the proposed work. Among the numerous classification models, Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), ANN, Convolutional Neural Network and Long short-term memory (LSTM) have been employed for classification. RF outperformed all other algorithms with the highest accuracy of 95%.

Keywords: Customer Segmentation,Machine learning,ANN,LSTM,Prediction.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES **PAGE**

NOMENCLATURE

The list Describes several abbreviations and will be used later within the documentation.

- Naive Bayes
- LR Logistic Regression
- DT Decision Tree
- RF Random Forest
- SVM Support Vector Machine
- KNN K-Nearest Neighbors
- ANN Artificial Neural Network
- CNN Convolutional Neural Network
- LSTM Long Short-Term Memory

CHAPTER 1

INTRODUCTION

1.1 Introduction

The Banking industry generates a massive volume of data every day. It contains customer account information, transaction information, all financial data etc. Bank management must identify quality dimensions and improve service quality to satisfy customers.The Banking organization is highly competitive.The banks can not competing among each other.Customer is an important effective tool that banks can use to gain a strategic advantage and survive in competitive banking industry. It helps to uncover hidden information, hidden patterns and to discover knowledge from the large volume data.

Cronin and taylor (1992),said that the purchase intention of the customers in the banking industry depend on customer satisfaction.The Customer getting the best response from the service provider result in the increase of profitability,the positive word of mouth and brand loyalty which is the function of customer loyalty.

Parasuraman et al. (1988), proposed that the servqual analysis was suitable to analyse the service quality of the banking industry by collect data from customers.The servqual consist of 5 dimensions related to the service operation,that are tangibles services,reliability,responsiveness,assurance and empathy.

Oliver (1997) gave a static definition of the customer satisfaction by analyzing the total consumption process and experience of the customer.For Satisfaction to affect loyalty frequent or cumulative satisfaction is required so that individual satisfaction become blended.

In recent years,deep learning tools such as Artificial Neural Network (ANN) have become established.ANN works on the concept of neurons in our human brain. The association between the neurons outputs and neuron inputs can be viewed as the directed edges with weights.

Deep learning another tool is LSTM have been established.Increasingly popular and proven in many time series forecasting problems.Such problems focus on determining future values of time series data with High accuracy.
In this thesis, we are focusing on:

- Mining in unbalanced data.
- Encoding category features.
- Model calibrating to acquire the best performance.

1.2 Scope of the Study

This study Covers the types of services Provided by different Banks and also find the level of the customers or not.

1.3 Problem Definition

Banking transactions are going on from the ancient times in different ways. Operating only monetary transaction is not the work of a bank but customer satisfaction is also the most. By their service facilities they can attract customers to the bank. This shows that there should be good and positive relation between bank and the customers. Therefore, this study focuses to analyze the present situation of banking services. This Data also can reveal major insights into how customers relate to a brand and they will interact with future.

1.4 Motivation

- This thesis investigates the Customer classification problem for prediction, employing supervised approaches to determine the overall semantic of customer prediction by classifying them as Yes or No.
- A model could be assist possible clients with settling on an informed decision on their get services from the banks and banks to improve their services.
- Fewer deep learning model used on Bank Customer dataset.

1.5 Objectives

- To provide good customer service, customer satisfaction rises and when customers are satisfied and ensures repeat account.
- My research aims to do this by conducting Bank Customer prediction or not as Yes or No.
- To examine the relationship between Customer and Bank.
- It can be beneficial to the bank because bank can easily understand who are their customers and who are not their customers.
- To find the highly effective model with Bank Customer Prediction dataset.

1.6 Organization of the Study

Chapter 1 Introduction: A synopsis of this research-based initiative can be found in this document. In addition, this chapter adequately addresses the rationale for conducting such a research-based initiative. The most important aspect of this chapter is the study's purpose. The information can be included in the introduction chapter is: -- Scope of the study, Objective of the study and Organization of the Study.

Chapter 2 (Literature Review) : This chapter discusses what has been done in the past in this domain. The final section of this chapter discusses the required tools.

Chapter 3 Methodology : This chapter is about the theoretical discussion of this research. Besides it has the flow charts of our proposed work. This chapter contain 5 Sections.

Chapter 4 Implementation and Models : This chapter is about the implementation of models to build the model also we have discussed parameter tuning to get optimal solution.

Chapter 5 Result and Discussion : This chapter is related to the outcome of this project and also it has some pictures and graphs for a better result. A conclusion is thus a deduction based on the findings.

Chapter 6 Conclusion and Future Work: This is based on the conclusion topics of the thesis. We will discuss what we will do in the future.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this Chapter, some previous work review of the existing system for Bank Customer has been presented. Finally, the chapter ends with the required tools for our proposed work. In this the reviews of the various relevant literatures in relation to support the study to receive some ideas for developing a research design.

2.2 Previous Work

During the study original copy of review of some web are included:

Syeda Farjana Shetu, Israt Jahan , Mohammad Monirul Islam, Refath Ara Hossai , Nazmun Nessa Moon and Fernaz Narin Nur Predicting Satisfaction of Online Banking System in Bangladesh using Logistic Regression, Random Forest, Naïve Bayes, support vector machine, Neural network, Decision tree, K nearest neighbor algorithms. They Can achieved accuracy was (LR – 96%,RF - 96%,DT – 93.33%,SVM – 93.3% and NN – 86%).Highest accuracy was 96% that is achieved by three algorithms Logistic regression random forest and decision tree. They cannot collect data directly by survey form. They tried to build a model that can predict user satisfaction and find the best service quality bank. To predict banking System KNN (K=5) and Naïve bayes using random state and gained this accuracy KNN-96% and naïve bayes-89.3%.

Sadaf Ilyas , Sultan Zia, Zaib un Nisa, Umair Muneer Butt, Sukumar Letchmunan, Predicting the Future Transaction from Large and Imbalanced Banking Dataset using logistic regression (LR), Random forests (RF), Decision tree (DT), Multilayer perceptron (MLP), Gradient boosting method (GBM), Category boost (CatBoost), Extreme gradient boosting (XGBoost), Adaptive boosting (Adaboost) and Light gradient boosting (LigtGBM) method on the dataset.They can achieved accuracy was (LR-91%,DT-83%,GBM-90%,XGB-90%.CATBOOST-92%,RF-90%,LIGHGBM(Classifier) – 90%). They find out that selecting the metrics is the first and foremost step to know what exactly we want to get from the classifier when working on Imbalanced data. They have used fine-tune hyperparameters for our dataset and implemented them in combination with the LighGBM. This tuning improves performance, and we have achieved 89% accuracy. They can LighGBM Threshold(t): 0.1025 and LighGBM Threshold(t): 0.1200 and they can gain the accuracy has 65% and 71%.

Priyanka S. Pati, Nagaraj V. Dharwadkar, Analysis of Banking Data Using deep Learning (ANN) with divided into train and test score.They are using a two Dataset D1 and D2. There are total 24 inputs and one output. It contains records of 1000 customers. It describes Root Mean Square Error (RMSE) and accuracy on training and testing datasets. They can gain achieved 72% and 90% accuracy for D1 and D2.

Muhammet Sinan Başarslan, İrem Düzdar Argun, Prediction of Potential Bank Customers: Application on Data Mining Using NB,KNN,LR AND RF.They can achieved accuracy with 5-fold (NB-88%,KNN-88%,LR-89% and RF-90%) and again 10-fold (NB-87%,KNN-86%,LR-90% and RF-92%).They can highest performance was achieved with the RF algorithm.The DT and Ada had a Same accuracy 89%. To predict the Was the customer became a bank customer?

G.Arutjothi,Dr.c.Senthamarai,prediction of loan status using Machine Learning Classifier.They Could achieved accuracy based on percentage Split(80-20%,70-30%,60-40% and 50-50%) and the accuracy was 74.5%,73.4%,74.9% and 75.08%.The RMSE Gained 2.04,1.91,2.29 and 2.41.They could iteration level of 30 KNN model gives significant accuracy.They can predict the loan status in banks.

Renato Alexandre de lima Lemos,Thiago christiano Silva,Benjamin Miranda Tabak,Propension to customer churn in a Financial institution Using Machine Learning.They could achieved Accuracy (DT-86%,KNN-84%,LR-84%,SVM-87% and RF-90%).They can predict the customer churn in Banking System when the customer is Existing customers than it is to acquire new customers.They can Predict Observation period and Prediction period about 6 months.

Emad Abd Elaziz Dawood, Essam Elfakhrany, Fahima A. Maghrab, Improve profiling bank customer's behavior using machine learning. They showed that improved k-means are the best accuracy technique equal to 37.61%.They Can only used in KNN Model and short Description about Neural Network.

Amir E. Khandani, Adlar J. Kim, Andrew W. Lo, Consumer credit-risk models via machine-learning algorithms.They can Predicted and actual 90-days-or-more delinquency rates (6-month) and (12 month).They achieved accuracy linear regression R2 ’s of forecasted/realized delinquencies of 85%.They can be shown at the grap impact of income ratio drop on future 3 and 6 months default rate.

Regina Esi Turkson, Edward Yeallakuor Baagyere, Gideon Evans Wenya, A Machine Learning Approach for Predicting Bank Credit Worthiness Using ML.The Algorithms can be used RF,RF(With only 5 features),K-neighbours,SVM,LR and NB.They could gain accuracy RF-81%,SVM-78%,K-neighbors-76% and NB-38% and neural networks-80%.The experimental results showed no significance difference in their predictive accuracy and other metrics. They Can develop a hybrid machine learning system that will incorporate the most important features.

Chaitrali S. Kulkarni, Amruta U. Bhavsar, Savita R. Pingale, Prof. Satish S. Kumbhar, BANK CHAT BOT – An Intelligent Assistant System Using NLP and Machine Learning. They algorithms can be used DT,RF and SVM.They can gain accuracy DT-90%,RF-92% And SVM-94%. They Can not working in responsive images,links and not showing account related information using Bank Gateway. This paper explains the dataset that we have prepared from FAQs of banks websites, architecture and methodology used for developing such chat bot.

Ajmal Amiry, An Analysis of Customer Satisfaction Level in Banking Sector of Afghanistan they can said that customer satisfaction level in increasing customer loyalty in the banking industry specifically the Azizi bank. The study will highlight the numerous factors affecting the satisfaction level of the customers. They could do so by improving their managerial strategies and having robust hold over the functions so that the customers are satisfied with their services.

Nasser Mohammadi, Maryam Zangeneh, Customer Credit Risk Assessment using Artificial Neural Networks they can be applied MLPNN, LR, C4.5. The accuracy can be achieved (MLPNN-78.40%,LR-75.90%,C4.5-70.70%). The MSE Accuracy got 0.1593, 0.3127, 0.3459. the cost that Type II error rate imposes to the model is too high, therefore, Receiver Operating Characteristic curve is used to find appropriate cut-off point for a model that in addition to high Accuracy, has lower Type II error rate.

2.3 Required Tools

The following tools will be used in the implementation of the designed system:

- Colaboratory Notebook
- NumPy
- Pandas
- OS
- Sequential
- Keras
- Scikit-Learn
- Dense
- Tensorflow
- Seaborn
- Matplotlib
- Python
- Jupyter Notebook
- Google Drive

2.3.1 Colaboratory Notebook

The Collaboratory, or "Colab," is a Google Research creation. Because of the web-based Python editor Colab, anyone can write and run Python programs. It has applications in education, data analysis, and machine learning. Colab is a hosted Jupyter notebook service that does not require installation and provides free access to computing resources such as GPUs. We can use and share Jupyter notebooks with others through Colab without downloading, installing, or running anything. Colab's resources are not always available or limitless, and usage guidelines are subject to change at any time. This is required for Colab to provide free materials. Go to Resource Limits for more information.

2.3.2 NumPy

NumPy, a Python library that works with arrays, has a variety of applications. It also includes tools for working with a multidimensional array object with high performance. This is the most important Python package for scientific computing. It has several characteristics, including the following: An N-dimensional array's strong object (Broadcasting) C/C++ and Fortran code integration tools with advanced features Linear algebra, the Fourier transform, and random number generation are all capabilities. In addition to its obvious scientific applications, NumPy can be used as a multidimensional

container of generic data. NumPy can connect to a variety of databases quickly and cleanly because it allows the creation of any data type.

2.3.3 Pandas

Pandas, an open-source Python toolkit, uses solid data structures to provide high-performance data manipulation and analysis. The term "Pandas" comes from the term "Panels Data," which is used in econometrics to refer to multidimensional data. Python was mostly used for data preparation and munging before Pandas. It was only marginally helpful for data analysis. This problem was solved by pandas. We can use Pandas to import, prepare, alter, model, and analyze data from any source. Finance, economics, statistics, analytics, and other academic and professional fields use Python with Pandas.

2.3.4 OS

Python's OS module contains tools for interacting with the operating system. OS is one of Python's most common utility modules. This module allows us to use operating system-specific features while traveling. A large number of file system interface functions are available in the *os* and *os. path* modules**.**

2.3.5 Sequential

The core idea of Sequential API is simply arranging the Keras layers in a sequential order and so, it is called *Sequential API*.
Can be simply calling Sequential() API as specified below:

```
from keras.models import Sequential
model = Sequential()
```

2.3.6 Keras

Keras is an open-source high-level Neural Network library, which is written in Python is capable enough to run on Theano, TensorFlow, or CNTK. . It is written in Python and is used to implement neural networks easily.

2.3.7 Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. It is a free software machine learning library for the Python programming language. The primary characteristics of Scikit-Learns are as follows:

• Efficient and simple data mining and analysis methods. Other classifications, regression, and clustering algorithms are also mentioned, such as support vector machines, random forests, gradient boosting, and k-means.
• It can be utilized by anyone and repurposed in several circumstances.
• Because it is built on top of NumPy, SciPy, and matplotlib, it is quick.

2.3.8 Dense

Dense layer is the regular deeply connected neural network layer. It is most common and frequently used layer. Dense layer does the below operation on the input and return the output.

output = activation(dot(input, kernel) + bias)

2.3.9 Tensorflow

TensorFlow is an open source machine learning framework for all developers. It is used for implementing machine learning and deep learning applications.

2.3.10 Seaborn

Seaborn is an amazing visualization library for statistical graphics plotting in Python.It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn works easily with dataframes and the Pandas library.

2.3.11 Matplotlib

Matplotlib is an excellent Python visualization library for 2D array charts. Matplotlib is a crossplatform data visualization toolkit built on NumPy arrays and designed to work with the entire SciPy stack. John Hunter first mentioned it in 2002. One of the most important advantages of visualization is that it provides us with visual access to massive amounts of data in easily 20 understandable images. Line, bar, scatter, histogram, and other plot types are available in Matplotlib.

2.3.12 Python

Python is a high-level programming language that is dynamically semantic, interpreted, and object-oriented. Its high-level data structures, combined with dynamic typing and dynamic binding, make it ideal for Rapid Application Development as well as as a scripting or glue language for connecting existing components. Python's concise, simple syntax prioritizes readability, reducing software maintenance costs. Python supports modules and packages, which encourages program modularity and code reuse.

2.3.13 Jupyter Notebook

JupyterLab is the most recent web-based interactive development environment for notebooks, code, and data. Using the interface's flexibility, users can create and arrange workflows for machine learning, computational journalism, scientific computing, and data science. A modular architecture encourages extensions to increase and improve functionality.

2.3.14 Google Drive

Google Drive, a free cloud storage service, allows users to upload files and access them from anywhere on the planet. Documents, photos, and other data are synced between all of the customers' devices, including PCs, cellphones, and tablets. Google Docs, Gmail, Android, Chrome, YouTube, Google Analytics, and Google+ are among the other services and platforms that Google Drive is compatible with. Microsoft OneDrive, Apple iCloud, Box, Dropbox, and Sugar Sync compete with Google Drive.

2.4 Hardware

- Processor: Intel i5-8265U @ 1.60GHz 1.80 GHz
- Ram: 8 GB DDR4 2600MHz
- OS: Windows 10(22000.978)

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter covers the research method employed in this study. It is a Clear concept of my Research. It is defined as techniques that are used for conducting research such as in data collection, data analysis, and evaluation of the accuracy of the research results. We will be giving the best model that we found for our system.

3.2 Dataset & Data collection

The data has been used in this study is taken from Kaggle. We have collected the Data from Kaggle is Bank Customer Data. There are 17 properties and 42639 customer records in the bank data set. Our Dataset also had numerical and Categorical Data. This dataset has been labeled with two classes, which are Yes Or No.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	Customer
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
42634	21	student	single	secondary	no	2488	no	no	telephone	12	jan	661	2	92	1	success	yes
42635	87	retired	married	primary	no	2190	no	no	telephone	12	jan	512	2	-1	0	unknown	yes
42636	34	blue-collar	married	primary	no	6718	no	no	cellular	13	jan	278	4	97	1	other	no
42637	22	student	single	secondary	no	254	no	no	cellular	13	jan	143	2	-1	0	unknown	yes
42638	32	management	single	tertiary	no	1962	no	no	cellular	13	jan	130	1	-1	0	unknown	no

42639 rows × 17 columns

Figure 3.2: Sample of the dataset

3.2.1 Attributes description

The Dataset Contains 17 Attributes. We have chosen Bank Customer or not. The Duration attribute is highly affected the Output.

Table 3.1: Explanation of dataset for Customer

No	Attributes	Explanation	Example
1	age	Customer Age	58
2	job	type of job	management
3	marital	marital status	married
4	education	Customer Education	secondary
5	default	has credit in default	Yes
6	housing	has housing loan	Yes
7	loan	has personal loan	NO
8	contact	contact communication type	cellular
9	Customer	Was the customer became a bank customer?	Yes/no
10	day	last contact day of the week	5
11	month	last contact month of year	may
12	campaign	includes last contact	2
13	pdays	after the client was last contacted from a previous campaign	-1
14	previous	number of contacts performed before this campaign and for this client	0
15	poutcome	outcome of the previous marketing campaign	success
16	duration	last contact duration, in seconds	1077
17	Balance	Customer Account Balance	1506

3.3 Working Procedure

The main goal of my research is to Bank Customer predict from bank and classify what type of Customer it is. So, I can say that it is a binary classification problem. This work is accomplished through the use of various machine learning and deep learning models. I preprocessed the Dataset to remove noisy data before predicting the classifiers. Then I encoding features from preprocessed data. After this, I used a technique called Oversampling to deal with the issues of imbalanced data which I have in our dataset. Machine learning and deep learning kept 80% of training and 20% for testing purposes. Finally fed the training data into classifiers and deep learning models. Then used different evaluation methods to predict the outcome.

3.3.1 Methodology

The research methodology is shown below in fig:

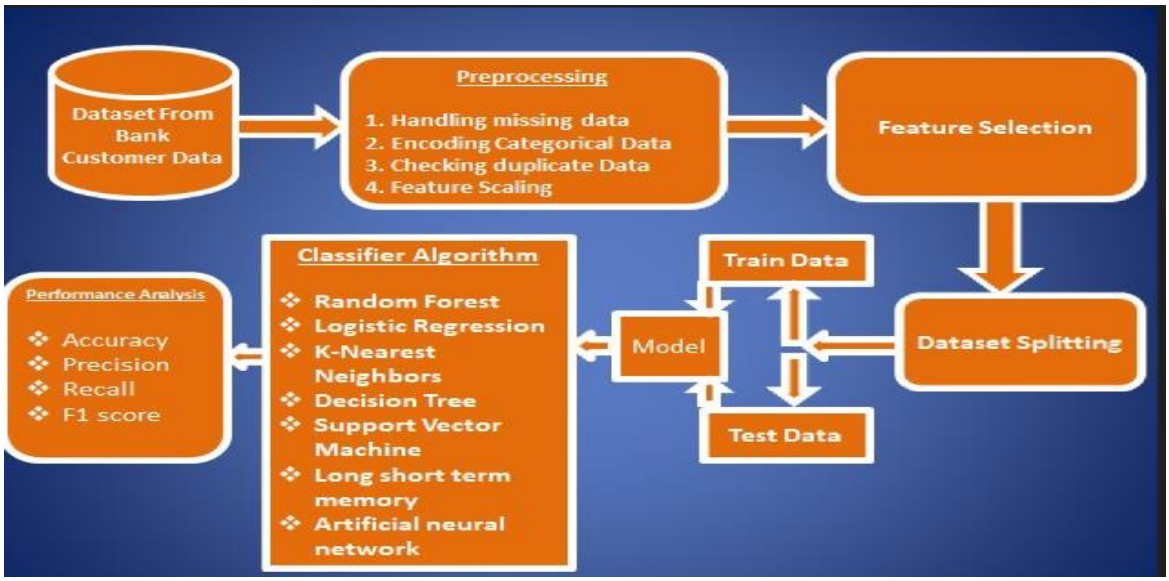

Fig 3.3: Research methodology

3.4 Data Pre-processing

In addition to the original data, I have some unnecessary data in this dataset that I do not require. In the beginning, I have to remove that unnecessary thing. The processes will be discussed below.

According to the statistical analysis, there was no missing data in the bank dataset.

```
age            0
job            0
marital        0
education      0
default        0
balance        0
housing        0
loan           0
contact        0
day            0
month          0
duration       0
campaign       0
pdays          0
previous       0
poutcome       0
Customer       0
dtype: int64
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42639 entries, 0 to 42638
Data columns (total 17 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   age         42639 non-null  int64
 1   job         42639 non-null  object
 2   marital     42639 non-null  object
 3   education   42639 non-null  object
 4   default     42639 non-null  object
 5   balance     42639 non-null  int64
 6   housing     42639 non-null  object
 7   loan        42639 non-null  object
 8   contact     42639 non-null  object
 9   day         42639 non-null  int64
 10  month       42639 non-null  object
 11  duration    42639 non-null  int64
 12  campaign    42639 non-null  int64
 13  pdays       42639 non-null  int64
 14  previous    42639 non-null  int64
 15  poutcome    42639 non-null  object
 16  Customer    42639 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.5+ MB
```

Figure 3.4: Check Missing Values

But input Dataset contains are categorical and numerical value.All categorical Data is converted into numerical values.when we can convert this Categorical to numerical values then we can divided into two parts.One is Training and other is Testing part.

3.4.1 Feature Encoding

Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. This process is called *feature encoding*.

Data frame analytics automatically performs feature encoding.

3.4.1.1 Before Feature Encoding

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	Customer
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Figure 3.4.1.1: Before Feature Encoding

3.4.1.2 After Feature Encoding

	age	balance	day	duration	campaign	pdays	previous	Customer	job_blue-collar	job_entrepreneur	...	month_may	month_nov	month_oct	month_sep	loan_yes	contact_telephone	contact_unknown	poutcome_other
0	58	2143	5	261	1	-1	0	no	0	0	...	1	0	0	0	0	0	1	0
1	44	29	5	151	1	-1	0	no	0	0	...	1	0	0	0	0	0	1	0
2	33	2	5	76	1	-1	0	no	0	1	...	1	0	0	0	1	0	1	0
3	47	1506	5	92	1	-1	0	no	1	0	...	1	0	0	0	0	0	1	0
4	33	1	5	198	1	-1	0	no	0	0	...	1	0	0	0	0	0	1	0

5 rows × 43 columns

Figure 3.4.1.2: After Feature Encoding

3.4.2 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization or standardization. Feature scaling is generally performed during the data pre-processing stage, before training models using machine learning algorithms. The goal is to transform the data so that each feature is in the same range (e.g. between -1 and 1). This ensures that no single feature dominates the others, and makes training and tuning quicker and more effective.

	age	balance	day	duration	campaign	pdays	previous	job_blue-collar	job_entrepreneur	job_housemaid	...	month_may	month_nov	month_oct	month_sep	loan_yes	contact_telephone	contact_unknown	p
0	0.519481	0.092259	0.133333	0.053070	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	1.0	
1	0.337662	0.073067	0.133333	0.030704	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	1.0	
2	0.194805	0.072822	0.133333	0.015453	0.0	0.0	0.0	0.0	1.0	0.0	...	1.0	0.0	0.0	0.0	1.0	0.0	1.0	
3	0.376623	0.086476	0.133333	0.018707	0.0	0.0	0.0	1.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	1.0	
4	0.194805	0.072812	0.133333	0.040260	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	1.0	

5 rows × 42 columns

Figure 3.4.2: Feature Scaling

3.4.3 Normalization

Normalization of a dataset means rescaling the range[0,1] meaning that the minimum and maximum value of a feature/variable are going to be 0 and 1, respectively.

3.5 Splitting dataset into Train and Test set

After data preprocessing, We will Divide The Data into Two Steps.The One step is Training and the Other One is Test Step. By Dividing The two Steps,It can improve Performance of our model and give better predictability. We can easily evaluate our model and Find the accuracy. For splitting the dataset, we can use the train_test_split function of scikit-learn.

Split(Percentage)	Train	Test
(80-20)%	(61884, 42)	(15472, 42)

Table 3.5: Data Splitting

3.6 Application of Classification Models

This section gives a brief overview of the algorithms used in my thesis study. Here, I evaluated two deep learning models and six machine learning models in our system.

Models are established with various classification algorithms using the bank dataset to estimate bank customers. The classification algorithms used in Machine Learning By Random Forest Classifier, Logistic Regression, SVM,K-Neighbors Classifier, Naïve Bayes And Decision Tree Classifier. The Models has a Four Performance measures are Used.

The Deep Learning Model Can be Used in ANN,CNN And LSTM. ANN Models has a Four Performance measures are used in the accuracy,precision,recall and F1-score with a Confusion Matrix. The accuracy Can be divided Into Macro Avg and Weighted Avg. The LSTM Model convert into Dataset matrix with a timestep and reshaping the Data for the Future Prediction.

3.6.1 Machine Learning Model

SL NO	Model
01	Random Forest Classifier
02	Logistic Regression
03	SVM
04	K-Neighbors Classifier
05	Naïve Bayes
06	Decision Tree Classifier

Table 3.6.1: ML Model

3.6.2 Deep Learning Model

SL NO	Model
01	ANN
02	CNN
03	LSTM

Table 3.6.2: DL Model

CHAPTER 4

IMPLEMENTATION OF MODEL

4.1 Introduction

In this chapter different machine learning and deep learning, models were implemented to predict the Customer Satisfaction or Unsatisfaction in this research. So We will identify the best model here in our research and we will describe those models here.

4.2 Machine Learning Model

Machine Learning: It can learn from Data, identify patterns and make decisions like a human intervention.

4.2.1 Random Forest

During the training phase of the random forests or random decision forests ensemble learning approach, which is used for classification, regression, and other tasks, a large number of decision trees are built. For classification problems, the random forest output is the class that the majority of the trees chose. For regression tasks, the mean or average prediction of each tree is returned. In RF, the classifier data variables are drawn at random from a large number of trees. The RF is built using the four streamlined stages that follow. There are N instances (cases) in the training data, and M attributes in the classifier. The number of examples (or "cases") in the training data is N, while the number of attributes (or "attributes") in the classifier is M.

4.2.2 Logistic Regression

The most basic version of logistic regression, though there are many more complex variations, employs a logistic function to describe a binary dependent variable. Regression analysis employs the logistic regression method (also known as logistic regression) to estimate the parameters of a logistic model (a form of binary regression). At the turn of the twentieth century, the biological sciences began to employ logistic regression. Later, it was used for a variety of social science applications. Logistic regression is a technique for estimating the probability of a discrete outcome from an input variable. Types of Logistic Regression-

• Binary Logistic Regression-The categorical response has only two 2 possible outcomes. Example: Spam or Not.

• Multinomial Logistic Regression-Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan).

• Ordinal Logistic Regression-Three or more categories with ordering. Example: Movie rating from 1 to 5.

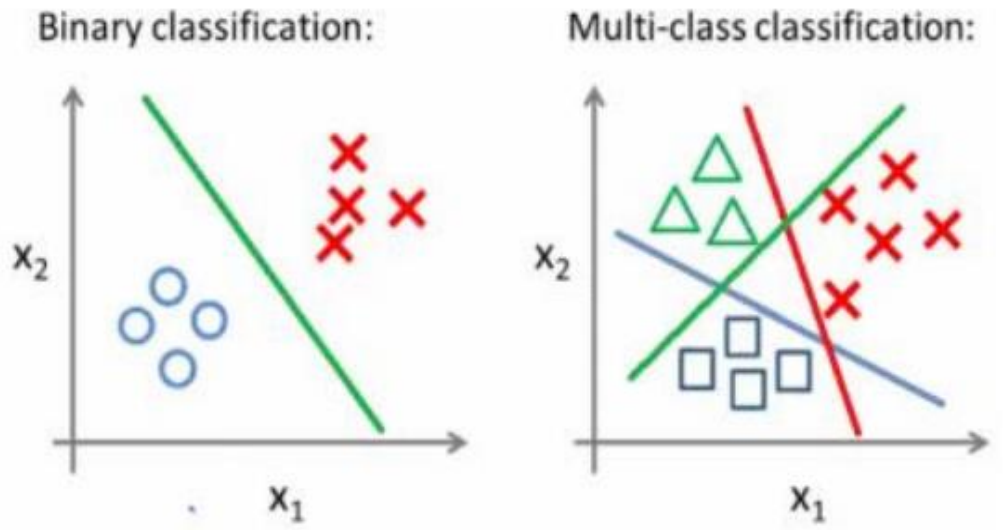

Figure 4.2.2: Logistic Regression

To model the probability of a given class or occurrence, a supervised machine learning method known as logistic regression can be used. When the outcome is binary or dichotomous and the data can be separated linearly, this strategy is used. As a result, logistic regression is frequently used to solve problems involving binary classification.

4.2.3 Support Vector Classifier (SVC)

A Linear SVC (Support Vector Classifier) is to fit the data you provide, returning a "best fit" hyperplane that divides or categorizes your data. After you've obtained the hyperplane, you can feed some features into your classifier to see what the "predicted" class is. The Linear SVC model has more parameters than the SVC model, such as penalty normalization (L1 or L2) and loss function.

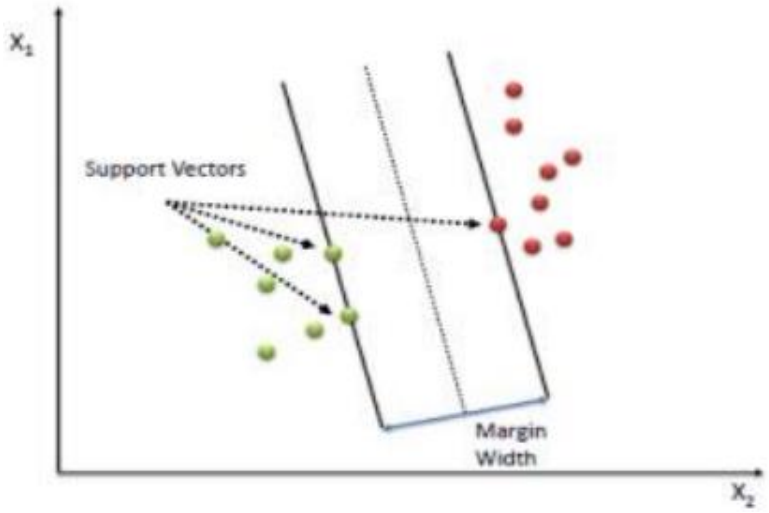

Figure 4.2.3: SVC Classifier

Kernel functions are a method of converting input data into the format required for processing. The Kernel Function modifies the training set of data in this way so that a non-linear decision surface can be converted into a linear equation in higher dimension spaces. The implementation makes use of the SVC library. The fit time may become prohibitive after tens of thousands of data points because it scales at least quadratically with the number of samples. When dealing with large datasets, consider using Linear SVC or SGD Classifier. A one-to-one system is used to manage multiclass support.

4.2.4 K-Nearest Neighbor(KNN)

The K-Nearest Neighbors algorithm and technique can be used for both regression and classification tasks. K-Nearest Neighbors examines the labels of a predetermined number of data points surrounding a target data point to predict which class the data point belongs to. K-Nearest Neighbors (KNN) is a conceptually simple yet extremely powerful algorithm, and it is one of the most widely used machine learning algorithms. Let's take a closer look at the KNN algorithm and see how it works. Understanding how KNN works will allow you to appreciate the best and worst use cases for KNN.

4.2.5 Naïve Bayes

It is used for solving classification problem and make first predictions. It predict the basis of probability of an object. It depends on the conditional probability and it can be used to formula for Bayes theorem. The combination of the prediction for all parameters is the final prediction, that returns a probability of the dependent variable to be classified in each group. The final classification is assigned to the group with the higher probability.

4.2.6 Decision Tree

The Decision Tree algorithm is a member of the supervised learning algorithm family. The goal of using a decision tree is to build a training model that can be used to predict the class or value of the target variable by learning simple choice rules generated from prior data (training data). Decision trees are a type of supervised machine learning that divides data indefinitely based on a parameter. Using the binary tree from earlier, one can comprehend a decision tree. Most decision tree algorithms work top-down, selecting a variable that best divides the set of objects at each stage. Decision trees use a variety of techniques to determine whether to divide a node into two or more sub-nodes. Sub-node formation increases the homogeneity of newly formed sub-nodes. The decision tree divides the nodes based on all of the available factors, then selects the split that produces the most homogeneous sub-nodes. It can help with both decision making and decision representation. A condition is represented by each internal node in a decision tree, and a choice is represented by each leaf node in a decision tree.

4.3 Deep Learning Model

Deep Learning: It is a subset of machine learning. To achieve state-of-the-art accuracy, sometimes exceeding human-level performance.

4.3.1 ANN

It works the way human brain processes information. It includes a large number of connected processing units.

To understand the concept of the architecture of an artificial neural network, we have to understand what a neural network consists of. In order to define a neural network that consists of a large number of artificial neurons, which are termed units arranged in a sequence of layers. Lets us look at various types of layers available in an artificial neural network.

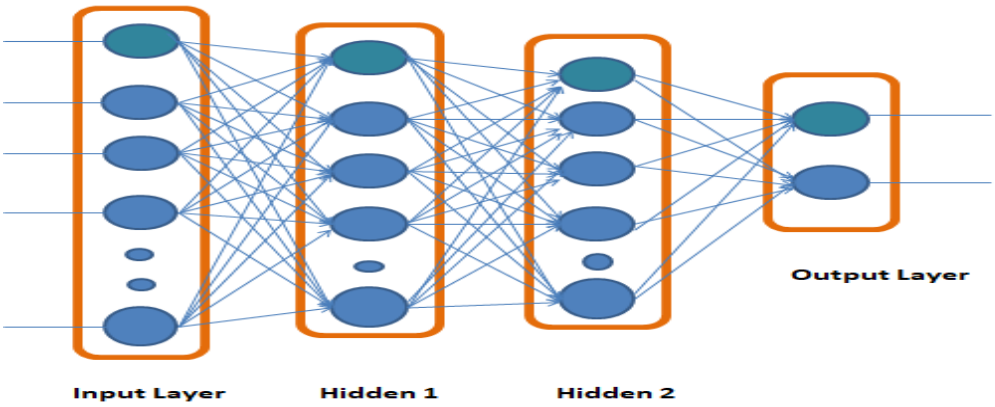

Figure 4.3.1: Layer combination of the ANN architecture

It consists of three layers:

- **Input Layer:** It accepts inputs in several different format.
  Our Dataset has a 42 inputs.
- **Hidden Layer:** To perform all the calculations to find hidden features and patterns.
- **Output Layer:** We can get the output weighted sum of inputs and includes a bias.

4.3.2 Convolutional neural network (CNN)

Deep learning has emerged as a very useful approach in recent years due to its ability to handle massive amounts of data. Hidden layers have surpassed traditional methods in popularity, particularly in pattern recognition. The convolutional neural network is a popular deep neural network. A Convolutional Neural Network (CNN) is a Deep Learning system that can take in an input image, assign importance to different elements and objects in the image (learnable weights and biases), and distinguish between them.

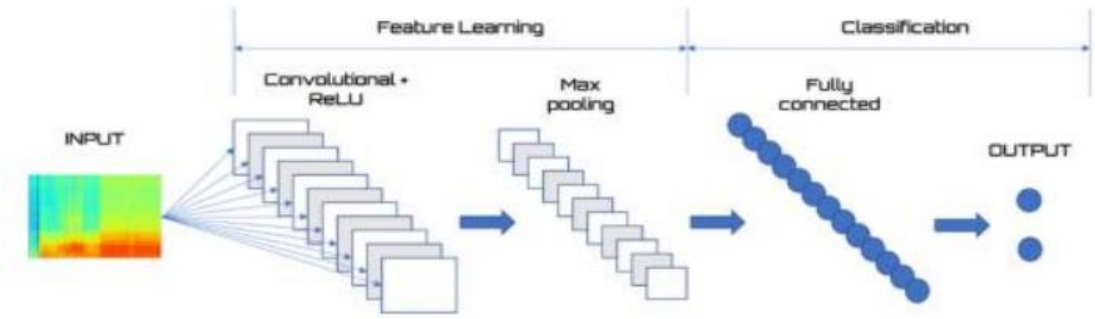

Figure 4.3.2: Layer combination of the CNN architecture

❖ **Convolutional layer**

The first and most important layer is the convolution layer. The convolutional layer's K filters look for phrases in the input that are similar to the filter. A one-dimensional convolution is an operation that involves a weight vector, m R m, and an input vector, s R s, where m is the convolution's filter. Convolution is the degree of overlap when one function is transferred to another. In this instance, the filter scans the entered sentence in search of word patterns that match the filter and the phrase. Furthermore, filters are taught during training, allowing us to train them on more data and build up their strength to match the demands of our network. We fix certain parameters (hyper), which are independent of training, such as the number of filters and the stride (the offset of how far to shift the filter across the text).

❖ **Max pooling layer**

Pooling layers often come after convolutional layers. The input that is supplied to the pooling layers is divided into subsamples. The most common pooling technique involves doing a max operation on the output of each filter. Pooling in NLP often produces just one value for each filter and is applied to the total output. For a 22 window, the following example shows maximum pooling. Pooling reduces storage and over-fitting.

❖ **Fully connected layer**

A totally connected input layer "flattens" the output of the preceding layers into a single vector that may be utilized as an input for the following layer. To forecast the correct label, the first fully connected layer applies weights to the feature analysis inputs. A fully connected output layer provides the final probability for each label.

4.3.3 LSTM

LSTM is a fundamental RNN extension. It reduces the problem of disappearing gradients and can track dependencies across large gaps.

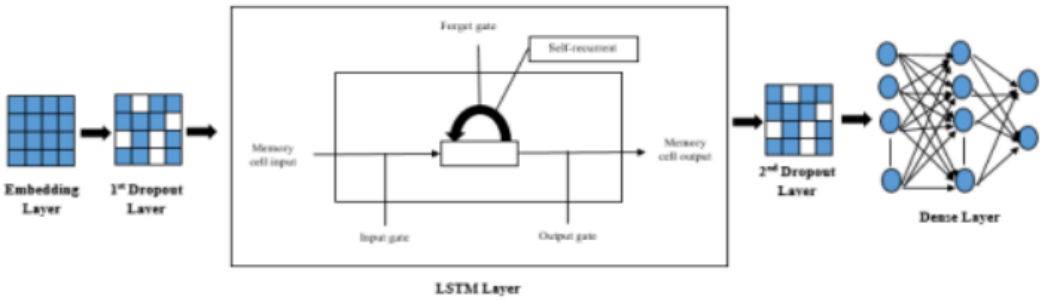

Figure 4.3.3: Layer combination of the LSTM architecture

An LSTM's recurrent hidden layer contains distinct units known as memory blocks. Memory blocks include memory cells with self-connections that store the network's temporal state, in addition to specific multiplicative units called gates that regulate the flow of information. In the original architecture, each memory block had one of three different gate types, namely:

- **Input gate:** The input gate regulates the flow of input activations into the memory cell.
- **Output gate:** this gate regulates how cell activations exit the network and enter other nodes.
- **Forget gate:** scales the internal state of the cell before adding it as input through the self recurrent connection of the cell, adaptively forgetting or resetting the memory of the cell.

4.4 Confusion matrix

It is a N*N matrix and evaluate the performance of a classification model. Tabular summary of the number of correct and incorrect predictions. They are following 4 types:

True Positives (TP): Actual value positive and predicted value positive.
True negatives (TN): Actual value Negative and predicted value Negative.
False positives (FP): Actual Negative and predicted positive.
False negatives (FN): Actual positive and predicted Negative.

TABLE 4.4: Evaluation Metrics and Definition

Metric	Equation	Definition
Accuracy	$\frac{TP + TN}{TP + FP + FN + TN}$	predicts the correct output
Precision	$\frac{TP}{TP + FP}$	Proportion of the predicted positive cases that are correct.
Recall	$\frac{TP}{TP + FN}$	total positive classes how our model predicted correctly.
F1-measure	$\frac{2 * Recall * Precision}{Recall + Precision}$	Weighted harmonic mean of precision and recall.

CHAPTER 5

RESULTS & DISCUSSION

5.1 Introduction

In this section, We will be containing a description about the main findings of our research. We provides a comparison of all the models and their individual graph and confusion matrix of this section.

5.2 Experimental Results

In work, we applied 6 different Machine learning algorithms and 3 deep learning algorithm. Machine learning model such as Random Forest, Logistic regression, SVM, naïve bayes, k-neighbors and Decision Tree. Deep learning model such as ANN,CNN and LSTM.

The selected algorithm performs very well. Table 1 and 2 represent the accuracy of all algorithm with a precision, recall and f1-score.

Performance Comparison of ML algorithm:

Parameter	Random Forest	Logistic Regression	SVM	Naïve Bayes	K-neighbors	Decision Tree
Accuracy	95.2%	85.6%	85.8%	63.8%	90.7%	79.8%
Precision	95%	86%	86%	64%	91%	80%
Recall	95%	86%	86%	64%	91%	80%
F1-Score	95%	86%	86%	64%	91%	80%

Table 1: Comparison of ML algorithm

Table 1 it's found out that all the ML has different accuracy due to their way of working methodology. Since RF has the highest accuracy (95.2%) among all the tested models, so in this case, I can say RF is the best classifier for our dataset.

Performance Comparison of Deep Learning Algorithms given below:

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ANN	91.46%	92%	91%	91%
CNN	89%	90%	89%	89%
LSTM	90.7%	91%	1.00%	95%

Table 2: Comparison of Deep Learning Algorithm

Table 2 it's found out that all the Deep Learning has different accuracy due to their way of working methodology. Since ANN has the highest accuracy (91.46%) among all the tested models, so in this case, I can say ANN is the best classifier for our dataset.

5.3 Confusion Matrix

5.3.1 Random Forest

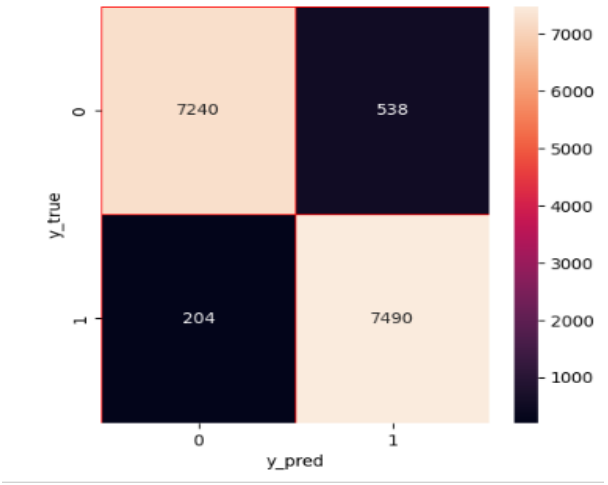

Fig 5.3.1: Confusion Matrix of Random Forest

Figure 5.3.1 Describes the confusion matrix of the RF model. From the confusion matrix, I calculated the Accuracy, Precision, Recall, and F1 scores of the RF model. Here I used the weighted average for calculating Precision, Recall, and F1 score for our multiclass classification problem. In this confusion matrix of the RF model, we can see that the true positive and true negative are 7240 and 7490, respectively, while the false positive and false negative numbers are 538 and 204.Our model performs well in terms of class prediction.

5.3.2 Logistic Regression

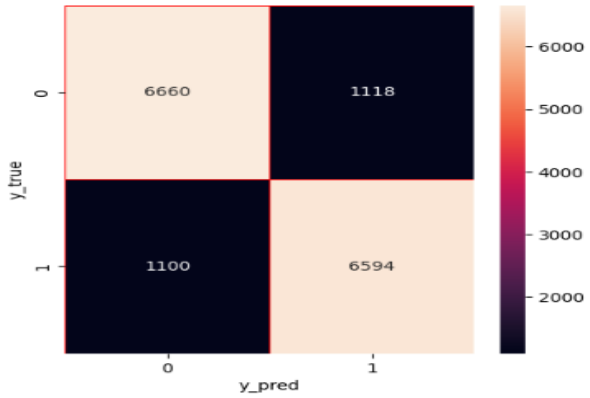

Fig 5.3.2: Confusion Matrix of Logistic Regression

Figure 5.3.2 Describes the confusion matrix of the Logistic Regression model. From the confusion matrix, I calculated the Accuracy, Precision, Recall, and F1 scores of the LR model. Here I used the weighted average for calculating Precision, Recall, and F1 score for our multiclass classification problem. In this confusion matrix of the LR model, we can see that the true positive and true negative are 6660 and 6594, respectively, while the false positive and false negative numbers are 1118 and 1100. Our model performs well in terms of class prediction.

5.3.3 SVM

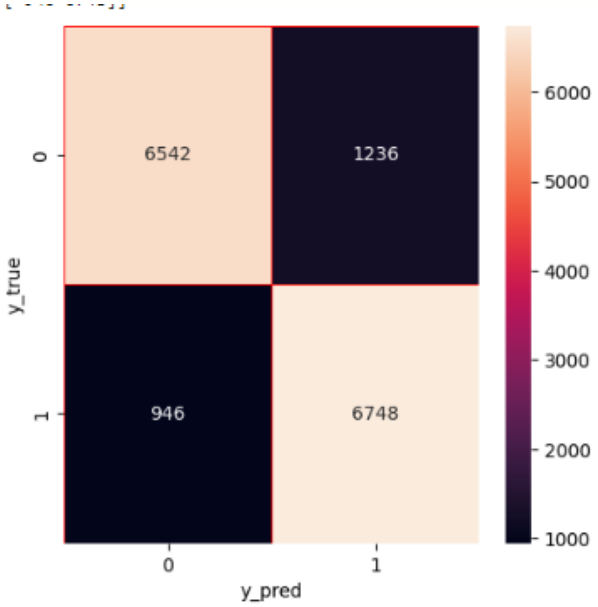

Fig 5.3.3: Confusion Matrix of SVM

Figure 5.3.3 Describes the confusion matrix of the SVM model. From the confusion matrix, I calculated the Accuracy, Precision, Recall, and F1 scores of the SVM model. Here I used the weighted average for calculating Precision, Recall, and F1 score for our multiclass classification problem. In this confusion matrix of the SVM model, we can see that the true positive and true negative are 6542 and 6748, respectively, while the false positive and false negative numbers are 1236 and 946.Our model performs well in terms of class prediction.

5.3.4 KNeighbors

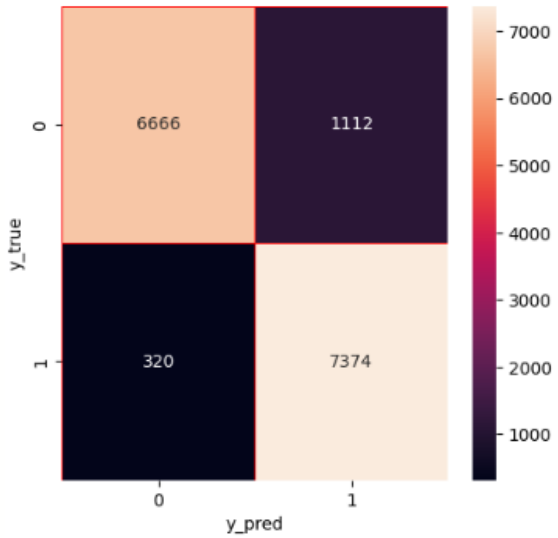

Fig 5.3.4: Confusion Matrix of KNeighbors

Figure 5.3.4 Describes the confusion matrix of the KNeighbors model. From the confusion matrix, I calculated the Accuracy, Precision, Recall, and F1 scores of the KNeighbors model. Here I used the weighted average for calculating Precision, Recall, and F1 score for our multiclass classification problem. In this confusion matrix of the KNeighbors model, we can see that the true positive and true negative are 6666 and 7374, respectively, while the false positive and false negative numbers are 1112 and 320.Our model performs well in terms of class prediction.

5.3.5 Naïve Bayes

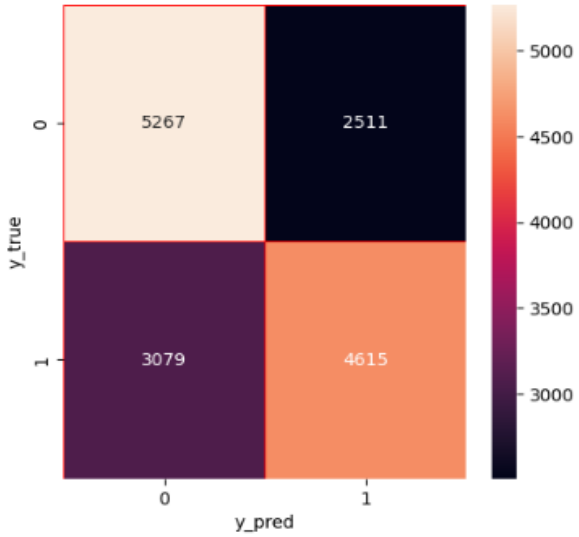

Fig 5.3.5: Confusion Matrix of Naïve Bayes

Figure 5.3.5 Describes the confusion matrix of the Naïve Bayes model. From the confusion matrix, I calculated the Accuracy, Precision, Recall, and F1 scores of the Naïve Bayes model. Here I used the weighted average for calculating Precision, Recall, and F1 score for our multiclass classification problem. In this confusion matrix of the Naïve Bayes model, we can see that the true positive and true negative are 5267 and 4615, respectively, while the false positive and false negative numbers are 2511 and 3079.Our model performs well in terms of class prediction.

5.3.6 Decision Tree

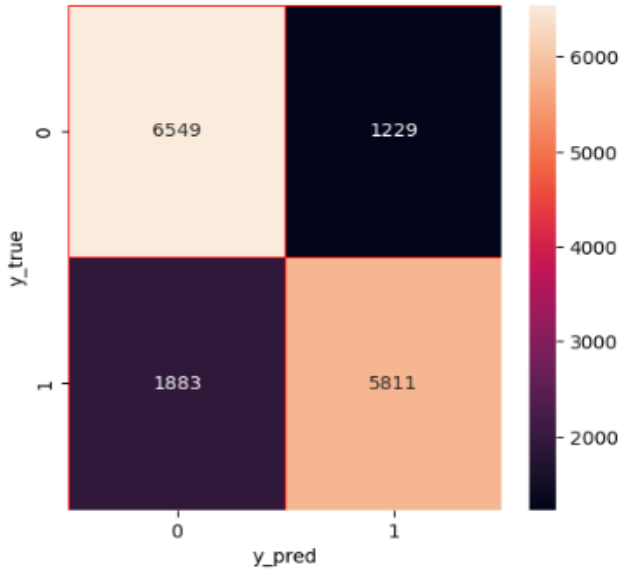

Fig 5.3.6: Confusion Matrix of Decision Tree

Figure 5.3.6 Describes the confusion matrix of the Decision Tree model. From the confusion matrix, I calculated the Accuracy, Precision, Recall, and F1 scores of the Decision Tree model. Here I used the weighted average for calculating Precision, Recall, and F1 score for our multiclass classification problem. In this confusion matrix of the Decision Tree model, we can see that the true positive and true negative are 6549 and 5811, respectively, while the false positive and false negative numbers are 1229 and 1883.Our model performs well in terms of class prediction.

5.3.7 ANN

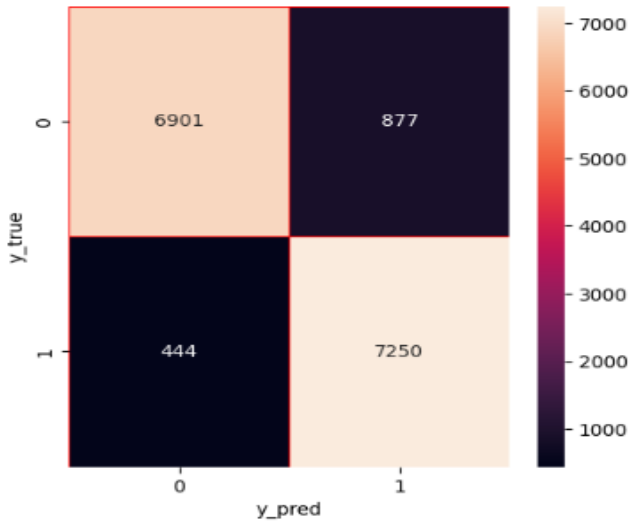

Fig 5.3.7: Confusion Matrix of ANN

Figure 5.3.7 Describes the confusion matrix of the ANN model. From the confusion matrix, I calculated the Accuracy, Precision, Recall, and F1 scores of the ANN model. Here I used the weighted average for calculating Precision, Recall, and F1 score for our multiclass classification problem. In this confusion matrix of the ANN model, we can see that the true positive and true negative are 6901 and 7250, respectively, while the false positive and false negative numbers are 877 and 444. Our model performs well in terms of class prediction.

5.3.8 CNN

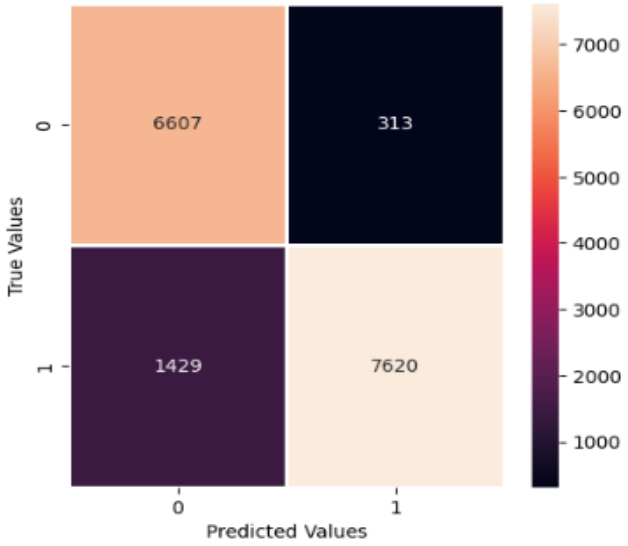

Fig 5.3.8: Confusion Matrix of CNN

Figure 5.3.8 Describes the confusion matrix of the CNN model. From the confusion matrix, I calculated the Accuracy, Precision, Recall, and F1 scores of the CNN model. Here I used the weighted average for calculating Precision, Recall, and F1 score for our multiclass classification problem. In this confusion matrix of the CNN model, we can see that the true positive and true negative are 6607 and 7620, respectively, while the false positive and false negative numbers are 313 and 1429. Our model performs well in terms of class prediction.

5.4 Classification Report

5.4.1 Random Forest

```
Accuracy score of the  RandomForestClassifier(n_estimators=350)  =  0.9520423991726991
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      7444
           1       0.97      0.93      0.95      8028

    accuracy                           0.95     15472
   macro avg       0.95      0.95      0.95     15472
weighted avg       0.95      0.95      0.95     15472
```

Fig 5.4.1: Classification report of Random Forest

Figure 5.4.1 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 95%.

5.4.2 Logistic Regression

```
Accuracy score of the  LogisticRegression(max_iter=1000, random_state=1)  =  0.8566442605997932
              precision    recall  f1-score   support

           0       0.86      0.86      0.86      7760
           1       0.86      0.86      0.86      7712

    accuracy                           0.86     15472
   macro avg       0.86      0.86      0.86     15472
weighted avg       0.86      0.86      0.86     15472
```

Fig 5.4.2: Classification report of Logistic Regression

Figure 5.4.2 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 86%.

5.4.3 SVM

```
Accuracy score of the  SVC(kernel='linear')  =  0.858971044467425
              precision    recall  f1-score   support

           0       0.84      0.87      0.86      7488
           1       0.88      0.85      0.86      7984

    accuracy                           0.86     15472
   macro avg       0.86      0.86      0.86     15472
weighted avg       0.86      0.86      0.86     15472
```

Fig 5.4.3: Classification report of SVM

Figure 5.4.3 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 86%.

5.4.4 KNeighbors

```
Accuracy score of the  KNeighborsClassifier()  =  0.9074457083764219
              precision    recall  f1-score   support

           0       0.86      0.95      0.90      6986
           1       0.96      0.87      0.91      8486

    accuracy                           0.91     15472
   macro avg       0.91      0.91      0.91     15472
weighted avg       0.91      0.91      0.91     15472
```

Fig 5.4.4: Classification report of KNeighbors

Figure 5.4.4 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 91%.

5.4.5 Naïve Bayes

```
Accuracy score of the  BernoulliNB()  =  0.6387021716649431
              precision    recall  f1-score   support

           0       0.68      0.63      0.65      8346
           1       0.60      0.65      0.62      7126

    accuracy                           0.64     15472
   macro avg       0.64      0.64      0.64     15472
weighted avg       0.64      0.64      0.64     15472
```

Fig 5.4.5: Classification report of Naïve Bayes

Figure 5.4.5 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 64%.

5.4.6 Decision Tree

```
Accuracy score of the  DecisionTreeClassifier(max_depth=3, random_state=0)  =  0.7988624612202688
              precision    recall  f1-score   support

           0       0.84      0.78      0.81      8432
           1       0.76      0.83      0.79      7040

    accuracy                           0.80     15472
   macro avg       0.80      0.80      0.80     15472
weighted avg       0.80      0.80      0.80     15472
```

Fig 5.4.6: Classification report of Decision Tree

Figure 5.4.6 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 80%.

5.4.7 ANN

	precision	recall	f1-score	support
0	0.94	0.89	0.91	7778
1	0.89	0.94	0.92	7694
accuracy			0.91	15472
macro avg	0.92	0.91	0.91	15472
weighted avg	0.92	0.91	0.91	15472

Fig 5.4.7: Classification report of ANN

Figure 5.4.7 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 91%.

5.4.8 CNN

	precision	recall	f1-score	support
0	0.95	0.82	0.88	8036
1	0.84	0.96	0.90	7933
accuracy			0.89	15969
macro avg	0.90	0.89	0.89	15969
weighted avg	0.90	0.89	0.89	15969

Fig 5.4.8: Classification report of CNN

Figure 5.4.8 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 89%.

5.4.9 LSTM

	precision	recall	f1-score	support
0	0.91	1.00	0.95	7736
1	0.00	0.00	0.00	792
accuracy			0.91	8528
macro avg	0.45	0.50	0.48	8528
weighted avg	0.82	0.91	0.86	8528

Fig 5.4.9: Classification report of LSTM

Figure 5.4.9 Describes the classification report, I got the overall overview of our model by getting the value of precision, recall, accuracy, and f1-score and also got an accuracy of 91%.

5.5 Evaluation

For final algorithm we have chosen Random Forest algorithm because it generates better accuracy for precision,recall and f1 score rather than others. So, We can choose it.
We can evaluate it by confusion matrix with accuracy,precision,recall and f1-score.

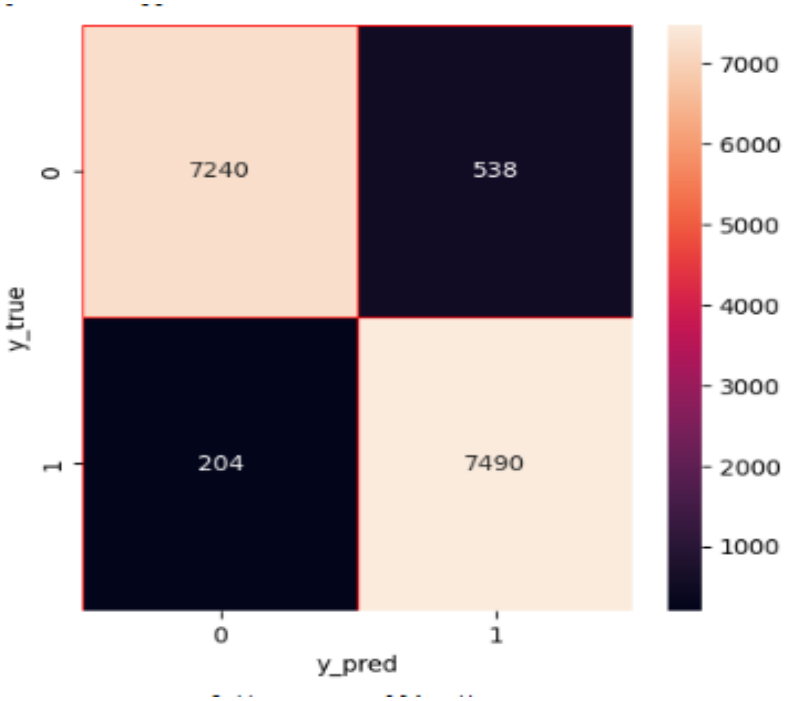

Fig 5.5:Evaluation

We can see that the model is better for Customer Prediction because there are 7490 number of Customer is True positive and True negative is 204. By predicting Bank customer is 0 or 1 with a actuals and predictions label.

5.6 Comparison on previous paper

In this section my research was tried to be compared with target paper works. The comparative analysis was based on accuracy. The comparison can be seen in the table below:

Paper Title	RF	LR	NB	KNN	DT	ANN
Prediction of Potential Bank Customers [2]	93%	90%	88%	86%	83.12%	88.18%
Proposed Model	95.20%	85.66%	63.8%	91%	79.88%	91.46%

Table 5.6: Comparative Analysis

Table 5.6 based on different researches listed in the table have conducted different pre-processing steps and feature encoding processes. As part of my research, I had to improve all of the encoding and scaling processes and preprocessing steps and select the ones with the highest accuracy. Use of different preprocessing process helped sorting out unnecessary data. And finally taking the best features encoding and features scaling from the datasets and learning through proper classifiers it was possible to attain greater accuracy. From the table it can be decided that the approaches used in approaches my proposed model shows more effectiveness and could achieve a better result than target paper works.

5.7 Model's Prediction

There are Two Classes in our Dataset. One is Yes and other one is No. Here we replace our dataset with yes and no with 0 and 1 and predict in our model.

5.7.1 Predict 1

```
input1 = X_test.iloc[[12638]]
input1
```

	age	balance	day	duration	campaign	pdays	previous	job_blue-collar	job
53952	0.438697	0.080622	0.10113	0.121815	0.031711	0.0	0.0	0.0	

1 rows × 42 columns

```
input1.shape
```

```
(1, 42)
```

```
prediction1 = ann.predict(input1)
prediction1
```

```
array([1])
```

```
prediction1 = ['1' if prediction1>0.5 else '0' for y in prediction1]
prediction1
```

```
['1']
```

Fig 5.7.1: Prediction for the Deep learning algorithm(Outcome 1)

5.7.2 Predict 0

```
input2 = X_test.iloc[[9654]]
input2
```

	age	balance	day	duration	campaign	pdays	previous	job_blue-collar	job_entrepreneur	job_hou
11275	0.350649	0.096145	0.566667	0.029687	0.016129	0.0	0.0	0.0	0.0	

1 rows × 42 columns

```
input2.shape
```

```
(1, 42)
```

```
prediction1 = ann.predict(input2)
prediction1
```

```
array([0])
```

```
prediction1 = ['1' if prediction1>0.5 else '0' for y in prediction1]
prediction1
```

```
['0']
```

Fig 5.7.2: Prediction for the Deep learning algorithm(Outcome 0)

CHAPTER 6

CONCLUSION & FUTURE WORK

This chapter develops conclusions and evaluates the outcomes based on the observations. It also identifies certain limits as well as the future works of the research.

6.1 Conclusion

Each algorithm works well in our work. We have found the best algorithm and best accuracy in this research. We have a model from which we can determine the bank customer. My research aims to do this by conducting Bank Customer prediction or not as Yes or No. After balancing the data with almost equal ratio of Yes or No prediction, Nine classification models have been used to our Dataset. Out of the Nine classifiers, i.e. Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial neural network (ANN), Convolutional neural network (CNN) and Long short-term memory (LSTM), predictive accuracy of RF is found to be the best. The accuracy results have been cross validated and the highest value of accuracy achieved was 95% for RF among the eight models. For Evaluation, We are used different techniques and choose the best model performed. For predicting we selected this algorithm which is highest accuracy. Only Machine learning was applied in the target paper that we selected but we used machine and deep learning in our paper. We have selected the right model in our paper so that customers get the best prediction in the bank.

6.2 Limitation

According to our naïve bayes model, we can see that the accuracy of naïve bayes is much lower than other models.

6.3 Future Work

This work can be extended in the following manner in future

- To use more Machine Learning and better hybrid models.
- To use more Deep Learning Models.
- Improve deep learning model accuracy
- Adding more data
- Try to use more feature encoding and scaling.

References

[1]Syeda Farjana Shetu, Israt Jahan, Mohammad Monirul Islam, Refath Ara Hossain , Nazmun Nessa Moon and Fernaz Narin Nur, "Predicting Satisfaction of Online Banking System in Bangladesh by Machine Learning", 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST).

[2]Muhammet Sinan Başarslan, İrem Düzdar Argun, "Prediction of Potential Bank Customers: Application on Data Mining", © Springer Nature Switzerland AG 2020, ICAIAME 2019, LNDECT 43, pp. 96–106, 2020.

[3]Renato Alexandre de Lima Lemos,Thiago Christiano Silva,Benjamin Miranda Tabak, "Propension to customer churn in a financial institution: a machine learning approach", Neural Computing and Applications (2022) 34:11751–11768.

[4]Sadaf Ilyas, Sultan Zia, Umair Muneer Butt, Sukumar Letchmunan, "Predicting the Future Transaction from Large and Imbalanced Banking Dataset", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020.

[5]Sahar F. Sabbeh, "Machine-Learning Techniques for Customer Retention: A Comparative Study", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018

[6]Chaitrali S. Kulkarni, Amruta U. Bhavsar, Savita R. Pingale, Prof. Satish S. Kumbhar, BANK CHAT BOT – An Intelligent Assistant System Using NLP and Machine Learning, International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 05 | May -2017.

[7]Regina Esi Turkson, Edward Yeallakuor Baagyere, Gideon Evans Wenya, "A Machine Learning Approach for Predicting Bank Credit Worthiness", ISBN: 978-1-4673-9187-0 2016 IEEE.

[8]Mst. Shuly Aktar, "The Impacts of Service Quality on Client Satisfaction: An Empirical Study on Private Commercial Banks in Bangladesh", Canadian Journal of Business and Information Studies, 3(5), 80-90, 2021.

[9]Ajmal Amiry, "An Analysis of Customer Satisfaction Level in Banking Sector of Afghanistan", International Journal of Science and Research (IJSR), Volume 11 Issue 5, May 2022.

[10]Chengming Chang, "Research on Domestic Bank Customer Satisfaction Based on Logistic Regression Analysis", International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2015).

[11]Dr. Prakash Shrestha, “Service Quality and Customer Satisfaction: Evidence of Nepalese Banks”.

[12]Nasser Mohammadi, Maryam Zangeneh, “Customer Credit Risk Assessment using Artificial Neural Networks”, I.J. Information Technology and Computer Science, 2016, 3, 58-66.

[13]Kaj Storbacka1, “Segmentation Based on Customer Profitability Retrospective Analysis of Retail Bank Customer Bases”, Journal of Marketing Management, 1997, 13, 479-492.

[14]Benjamin Osayawe Ehigie, “Correlates of customer loyalty to their bank: a case study in Nigeria”, International Journal of Bank Marketing Vol. 24 No. 7, 2006 pp. 494-508 q Emerald Group Publishing Limited 0265-2323.

[15]Sasha Fathima Suhel, Vinod Kumar Shukla, Sonali Vyas, Ved Prakash Mishra, “Conversation to Automation in Banking Through Chatbot Using Artificial Machine Intelligence Language”, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. June 4-5, 2020.

[16]Sonia Rezina,Nur Ahmad,Mohitul Mustafi,”Customer Perception on Bank Service Quality: A Comparative Study Between Conventional Commercial Banks and Islamic Commercial Banks in Bangladesh”,Global Disclosure of Economics and Business, Vol. 5(2), 2016.

[17]Jamal Ali Jaballa,Dr Bojan Gorgevic,”The Relationship between Bank Employee Training and Customer's Satisfaction: A Field Study on Commercial Banks Operating in Libya”,International Journal of Scientific and Research Publications, Volume 12, Issue 2, February 2022 475 ISSN 2250-3153

[18]O.A. Novokreshchenova,N.A. Novokreshchenova,S.E. Terehin3,”Improving Bank’s Customer Service on the Basis of Quality Management Tools”,European Research Studies, Volume XIX, Special Issue 3, Part B, 2016,pp. 19 – 38

[19]Arvid O.I. Hoffmann,Cornelia Birnbrich,”The impact of fraud prevention on bank-customer relationships An empirical investigation in retail banking”,International Journal of Bank Marketing Vol. 30 No. 5, 2012 pp. 390-407 Emerald Group Publishing Limited 0265-2323.

[20]Manidayanand,Dr. K. Neelamegam,”CUSTOMER PERCEPTION TOWARDS SERVICES PROVIDED BY PUBLIC SECTOR AND PRIVATE SECTOR BANKS: A COMPARATIVE STUDY”,Ilkogretim Online - Elementary Education Online, Year; Vol 20 (Issue 5): pp. 2005-2013.