IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

CUSTOMER CHURN PREDICTION IN THE FINANCIAL SECTOR USING SUPERVISED MACHINE LEARNING TECHNIQUES

[1]SK. Tabasum Fathima, [2]Dr. Y. Padma, [3]Y. Yougender, [4]S. Sreya, [5]M. Yaswini

[1,3,4,5]IV B. Tech Students. Department of Information Technology
[2]Assistant Professor, Department of Information Technology,
[1-5] Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India.

Abstract : Nowadays churning becomes the most popular issue in any sector because of the increase in service providers. Similarly, there are many options for customers to keep their money in whatever bank they want. This may lead to an increase in churn rate and a decrease in profit and the growth of particular banks. Identification of churn is most important to understand the reasons behind leaving the bank and can apply strategies to stop churning rate so that they can boost their business growth. This project proposes a method to predict customer churn in a Bank using machine learning techniques, a branch of artificial intelligence. The research promotes the exploration of the likelihood of churn by analyzing customer behavior. This study aims to find a machine-learning model that predicts customer churn in the taken churn_modelling dataset. The overall accuracy is taken as the metric to define the best classifier. Supervised algorithms like KNN [7], SVM [11], XGBoost [10], Logistic Regression [8], and Naïve Bayes [9]. From all algorithms, XGBoost [10] performed well with an accuracy of 86.25%, a precision of 88%, a recall of 95%, and an f1-score of 92% [1][3][4][12].

IndexTerms **- Customer churn, Machine learning techniques [7][8][9][10][11], Data-preprocessing, Python libraries-pandas, seaborn, numpy, matplotlib, sklearn [1][6].**

I.INTRODUCTION

Churning means a customer leaving or stopping the usage of a particular product or business. In the financial sector, there can be many reasons for customer churn like a lack of many products, interest rates, location of branches, and having better options from other banks. Customer retention is more important than getting a new customer. Customer retention success can depend not only on acquiring new customers but also depends on the satisfaction of existing customers. As the bank database will have information on n number of people, it will be difficult to predict customer churning as it takes more time manually. So, an effective machine learning model is required which will be trained easily and predict efficiently whether the customer will churn or not [1][3].

II. METHODOLOGY

In the process of developing a machine-learning model for customer churn prediction number of methodologies are used. Section 2.1 explains the dataset in detail, Section 2.2 explains about data-preprocessing, then the architecture of customer churn prediction 2.3.

2.1 Dataset

Data is very important in order to train a machine learning model. This study is based on European countries' data. The dataset is collected from Kaggle. Churn_modelling is the name of the dataset taken and it consists of 10,000 columns with 13 attributes. RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NoOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, and Excited are the attributes mentioned in the dataset [2].

Table-1: Dataset Description

Feature Name	Feature Description
Row number	Row numbers from 1 to 10000.
Customer Id	Unique Ids for bank customer identification.
Surname	Customer's last name.
Credit Score	Credit score of the customer.
Geography	The country from which the customer belongs.
Gender	Male or Female.
Age	Age of the customer.
Tenure	Number of years for which the customer has been with the bank.
Balance	Bank balance of the customer.
Num of Products	Number of bank products the customer is utilizing(savings account, mobile banking, internet banking etc.).
Has Cr Card	Binary flag for whether the customer holds a credit card with the bank or not.
Is Active Member	Binary flag for whether the customer is an active member with the bank or not.
Estimated Salary	Estimated salary of the customer in Dollars.
Exited	Binary flag 1 if the customer closed account with bank and 0 if the customer is retained.

In 10000 columns 7963 are non-churn customers and 2037 are churn customers. Exited is the target variable which represents 1 for the customers who are likely to churn and 0 for those who are not churning.

2.2 Data pre-processing

Data preprocessing is a useful technique that helps in converting raw data to clean data. It undergoes many steps like attribute selection, noise removal, and encoding, etc. Attribute selection means considering useful data features and removing unused features by using the noise removal method drop. In the churn_modelling dataset RowNumber, CustomerId, SurName, and CreditScore are not used as input so, we can remove them. LableEncoder is used to convert non-categorical variables into categorical variables so that they become machine-readable. [2][5].

```
In [14]: from sklearn.preprocessing import LabelEncoder
         label = LabelEncoder()
         X['Gender'] = label.fit_transform(X['Gender'])
         print(X['Gender'].head(7))

         0    0
         1    0
         2    0
         3    0
         4    0
         5    1
         6    1
         Name: Gender, dtype: int32

In [15]: X['Geography']=label.fit_transform(X['Geography'])
         print(X['Geography'].head())
         X['Geography'].value_counts()

         0    0
         1    2
         2    0
         3    0
         4    2
         Name: Geography, dtype: int32

Out[15]: 0    5014
         1    2509
         2    2477
         Name: Geography, dtype: int64
```

Fig 1: Converting non-numerical attributes like Gender and Geography to numerical attributes using LabelEncoder

2.3 Architecture of Customer Churn Prediction

The architecture of customer churn prediction depicts various steps from taking the dataset to predicting results. The following figure shows the system architecture of churn prediction.

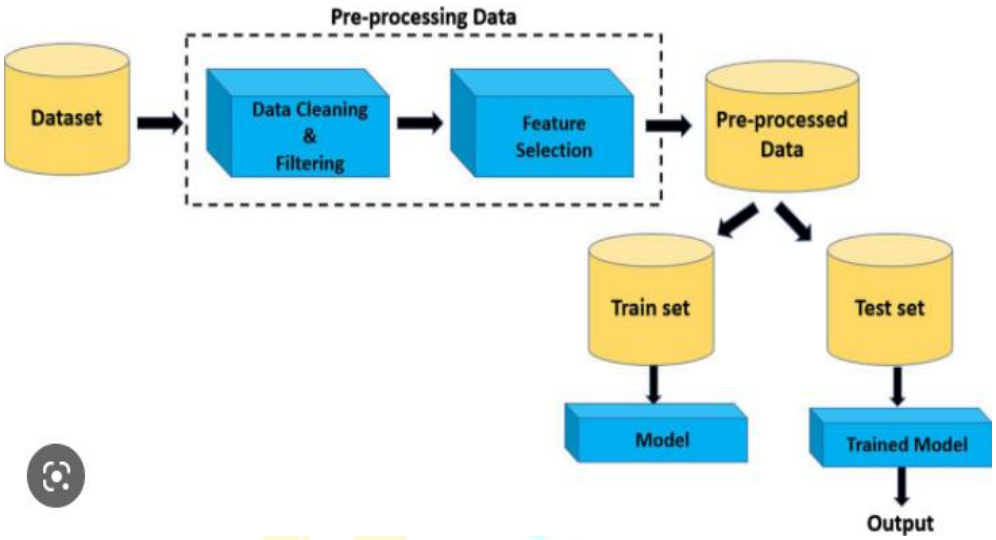

Fig 2: Architecture of Customer Churn Prediction

The dataset will be collected and undergo pre-processing steps and then the data will split into two parts- training data and testing data. 80% of data will undergo training data and 20% of data will treat as testing data. The model will train using training data and then the trained model will predict the output by taking the test data as input. The output shows 0 for non-churn customers and 1 for churn customers [2][4].

III. TECHNOLOGIES USED

3.1 Jupyter Notebook

Jupyter Notebook is an interactive web application used to run code by uploading datasets. It is an open-source application that enables users to share documentation including code, multimedia resources, etc. Jupyter Notebook is easy to use.

3.2 Python Libraries

In this study, we mainly use these libraries- numpy, pandas, matplotlib, seaborn, and sklearn. The Numpy library is used to work with array operations. Pandas library is used to manipulate the data. It is built on top of numpy so it can also provide support for multidimensional arrays. Matplotlib library is used for visualizations- to plot graphs and figures. Seaborn is also used for plotting graphs with themes. Sklearn is one of the famous and most important libraries in Python because it provides tools for classification, regression ad clustering [6].

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn.metrics as metrics
from sklearn.metrics import accuracy_score,precision_score,recall_score,f1_score, classification_report,confusion_matrix
%matplotlib inline
```

Fig 3: Libraries used in this project

IV. EXPERIMENT AND RESULTS

A number of steps were performed for the experiment. In this study, 4.1 explains the experimental setup and 4.2 explains the results and outcomes of the experiment.

4.1. Experiment

To perform an experiment one should need the experimental setup. Various methods will be performed on the system. So, the system must consist of essential hardware and software requirements. Hardware requirements: 8 GB of RAM, an Intel(R) Core i7 processor clocked at 1.80 GHz, and Windows 10 as the operating system must be fulfilled. Software requirements like Python, jupyter notebook and must support required libraries like numpy, seaborn, matplotlib, sklearn, pandas, etc [4][6].

4.2. Results

In this study, five supervised machine learning algorithms are considered and compared based on overall accuracy, precision, and recall. KNN [7], SVM [11], XGBoost [10], Logistic Regression [8], and Naïve Bayes [9] are the data mining techniques used to predict customer churn in the banking sector.

Table-2: Accuracy Table

	model	accuracy
0	KNN	0.8350
1	SVM	0.7835
2	XGBoost	0.8625
3	LogisticRgeression	0.7950
4	NaiveBayes	0.8220

All the above five algorithms are trained and tested against the testing data and predict how accurately the models are predicting the churn rate of a customer. XGBoost [10] algorithm performed well compared to other algorithms with an accuracy score of 86.25%, a precision of 88%, a recall of 95%, and an f1-score of 92% followed by the KNN [7] algorithm with an accuracy score of 83%, a precision of 84%, a recall of 97%, and an f1-score of 90% followed by Naïve Bayes [9] algorithm with an accuracy score of 82%, a precision of 83%, a recall of 97%, and an f1-score of 90% followed by Logistic Regression [8] with an accuracy score of 79%, a precision of 81%, a recall of 97%, and an f1-score of 88% followed by SVM [11] with an accuracy of 78%, a precision of 78%, a recall of 96%, and an f1-score of 88% [12].

The following figure shows the bar graph of five supervised algorithms representing their accuracies on the y-axis and algorithms on the x-axis.

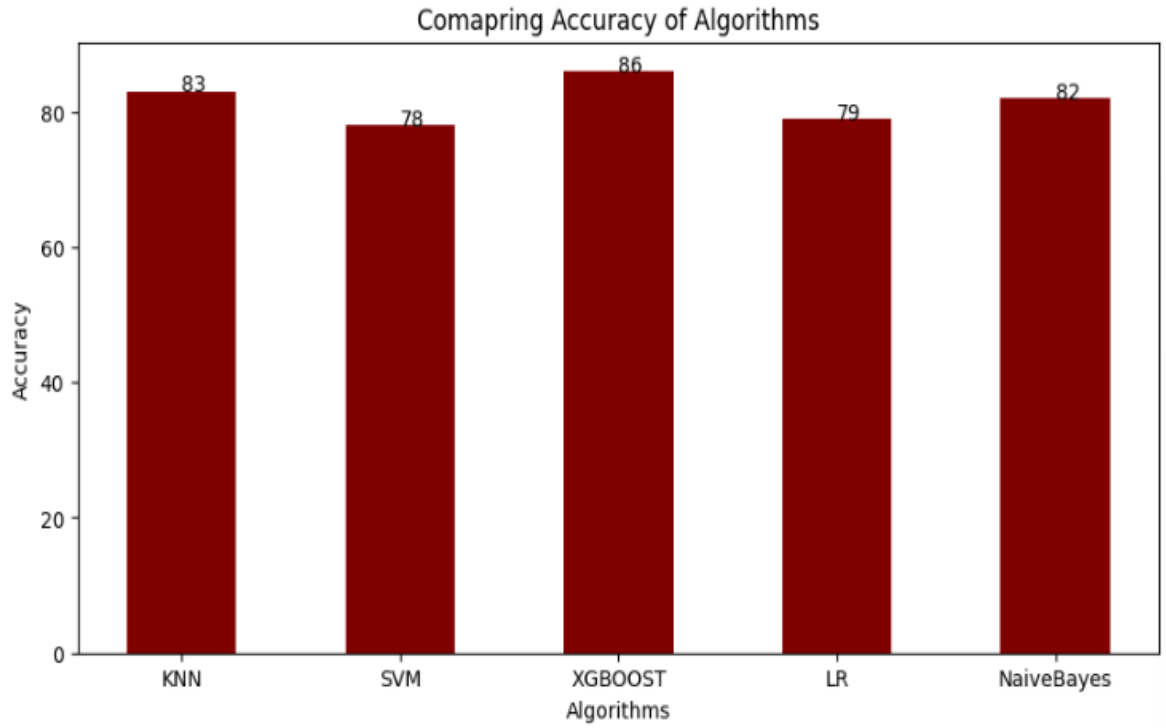

Fig 4: Accuracy comparison graph

As shown in Fig 4 and Table 2 the accuracy of the XGBoost [10] algorithm is high and it predicts customer churn effectively. On the y-axis, the accuracy percentage is taken at a 20 percent difference. From this, we can say that XGBoost [10] is an effective model for predicting customer churn in banking [12].

V. SCOPE OF FUTURE USE

This study can improve further by creating a module in respective bank applications so that when a customer installs a particular bank app and uses it for any purpose, the module should be able to detect whether that customer will remain in a particular bank or not at the beginning itself. Of this, an alarm might ring in a bank and they can take care of the reasons behind churning and they can stop that customer from churning [1][3].

VI. CONCLUSION

In this paper, we have studied the comparison of five supervised machine learning techniques- KNN [7], SVM [11], XGBoost [10], Logistic Regression [8], and Naïve Bayes [9]. The study of comparison was done on the Churn_modelling [2] bank dataset which consists of 10000 instances in which 7963 are non-churn customers and 2037 are churn customers and it contains 13 attributes. Through this study, we finally conclude that the XGBoost [10] algorithm achieved the best results in accuracy as well as precision and recall. It obtains an accuracy of 86.325%, a precision of 88%, a recall of 95%, and an f1-score of 92%. And the second-best

algorithm is the KNN [7] followed by Naïve Bayes [9]. By seeing the Accuracy we can promisingly say that it is possible to predict customer churn in banking using selected attributes [12].

REFERENCES

[1].https://mode.com/blog/predicting-and-preventing-churn/#:~:text=up%20for%20success ,What%20is%20churn%20prediction%3F,their%20behavior%20with%20your%20product.
[2]. https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling
[3]. https://www.zoho.com/subscriptions/guides/what-is-churn-rate.html#:~:text=Different%20types%20of%20churn,underlying%20causes%20and%20prevention%20strategies.
[4].https://www.researchgate.net/publication/348094541_Machine_Learning_Based_Customer_Churn_Prediction_In_Banking
[5]. https://contactsunny.medium.com/label-encoder-vs-one-hot-encoder-in-machine-learning-3fc273365621
[6]. https://opendatascience.com/top-7-most-essential-python-libraries-for-beginners/
[7]. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
[8]. https://www.javatpoint.com/logistic-regression-in-machine-learning
[9]. https://www.javatpoint.com/machine-learning-naive-bayes-classifier
[10]. https://www.geeksforgeeks.org/xgboost/
[11]. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
[12]. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd
[13].https://www.irjmets.com/uploadedfiles/paper/issue_9_september_2022/29877/final/fin_irjmets1663071650.pdf
[14].https://www.researchgate.net/publication/340855263_Churning_of_Bank_Customers_Using_Supervised_Learning
[15].https://www.researchgate.net/publication/366148946_Customer_Churn_Prediction_Using_Machine_Learning_Commercial_Bank_of_Ethiopia
[16].https://www.researchgate.net/publication/357539438_Customer_churn_analysis_in_banking_sector_Evidence_from_explainable_machine_learning_models