ML-Based Bank Churn Analysis for Improved Customer Retention

Mim Bin Hossain[1], Abidur Rahman[1], Sujana Islam Smrity[1], Md. Farhan Tonoy[1], Rahat Hasan Robin[1], Md. Mottakin Rahat[1], Rahma Mahbub[1], and Ragib Mahatab[1]

Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh

Abstract. The "ML-Based Bank Churn Analysis for Improved Customer Retention" project aims to address the critical issue of customer attrition and improve their strategies to hold the customers by providing more facilities in the banking sector through the application of machine learning techniques. In this study Bank customer churn or attrition prediction by using a comprehensive data analysis. The main objective is to forecast whether customers are going to leave the bank soon. This research aims to develop a predictive model that can accurately identify the risk of customer churn in banking sectors by analyzing large customer data. The dataset contains detailed information on customer personal data, relational data with the bank, transaction details, and credit profiles. This project includes data preprocessing, handling missing values, and ensuring data consistency which are the vital steps to achieve the goal. Compared predictive performance of some machine learning algorithms Support vector machine (SVM) and decision Tree here. Focused on evaluation metrics: Precision, Recall, and F1-score each model's effectiveness and importance. The findings from this study focused on the customer churn and insight of the banks to their customers retentions. Banks can improve their strategies by finding the issues for customers churn and also can improve their loyalty towards customers and this can boost their profit. This study identifies the importance of decision-making by analyzing the data to minimize customer attrition.

Keywords: Data analytics · Machine learning · Support Vector Machine (SVM) · Precision · Recall · F1-score · Decision Tree

1 Introduction

In the banking industry, Customer churn or attrition is an alarming issue that can affect the whole banking system. This can also affect the growth of the banks. This study will help to identify the reasons behind customer attrition and developing predictive models can help banks retain customers and reduce churn rates. Customer churn or attrition is the biggest issue in the banking industry. The profitability of banks depends on a strong customer base and customer churn breaks the relationship with the bank which can lead the bank to down

worth. Now it has become an alarming issue for the banks. If they could forecast customer attrition, it could be solved. Then the bank can be more profitable and economically strong. It is really hard to gain new customers than hold the old ones, so it is important to find the reasons for customer attrition. Customer churn affects the bank as a whole and spoils the whole customer base. Trust issues can be created among other customers. A bank can come to the road at a time. It should be predicted and solved before churn otherwise banks will be affected and the economy of our country as a whole. Commercial banks try to hold customers because customers staying and banking success are correlated. At this time bank churn prediction can play a vital role in customer holding. Where customer loyalty is equal to a bank's financial health, this bank churn prediction model will help the bank to survive in the competitive field of banking[1].

1.1 Importance of Bank Customer Churn Datasets

Banks can reduce risks using these datasets because there is relevant data for predicting customer churn like transaction details, last date of transaction, and number of transactions. These will help to detect possible risks for customer churn. Banks can plan a strategic solution using this dataset based on the client category as well and they can keep their valuable customers. Loss of a large number of customers can impact a bank's revenue, banks can optimize their revenue by putting some strategies to protect customers. In this case, the dataset can help. Banks can use cost-effective plans to keep the customers that already have. Using the dataset, they can identify the reasons for customer churn and solve it instantly. A complete view of data can help to implement a strategy to reduce customer churn. Banks can improve their service by applying insight into the reasons why customers are leaving their bank and taking initiative steps.

1.2 How Bank Customer Churn Datasets Help

A predictive model can be created using the dataset. There is historical data of the bank that describes the whole scenario of the customer and bank's relationship. An analytical overview of customers' relationship with the bank will help to predict what are the reasons for customer churn. This dataset will summarize the whole thing together. This predictive analytical model will also describe what customers want from the bank; and what facilities attract them to stay with the bank. Banks can easily understand what action should be taken to reduce customer churn. Banks can also find the reasons for dissatisfaction with the bank. The data is really useful and effective for the model. The banks can improve their services to satisfy customers. This dataset gives the bank feedback to learn from customers behavior and it leads the banks to better improvement. This can also help to create useful methods of modification[1].

2 Background

One of the problems affecting both financial health and competitive positioning in the banking industry is customer churn or the rate at which customers disappear from a business. Customer Churn Analysis: For banks, predicting and preventing customer churn is crucial for long-term growth & stability[2].

The concept of the "Bank Customers Churn Prediction" project is an advanced data analytics and machine learning solution offered by Datary (referred to as Project). It analyzes customer attrition by digging into historical data and tries to determine patterns that led customers away, so banks can take appropriate steps beforehand to prevent them.

The datasets used in the project were taken from publicly available Kaggle repositories: Customer demographics data, customers' bank order habits, and banking relationships based on transaction behaviors as well as credit profiling performed. The presence of these various data sets enables a holistic view of components that cause customer churn.

Our process: We will look at several datasets, learn more about handling missing values, do an exploratory data analysis, and try to predict which customers are likely to churn.

The results aim to provide banks with usable insights for developing targeted retention strategies. By understanding why customers come, banks can improve customer satisfaction, improve service quality, and increase profitability[2].

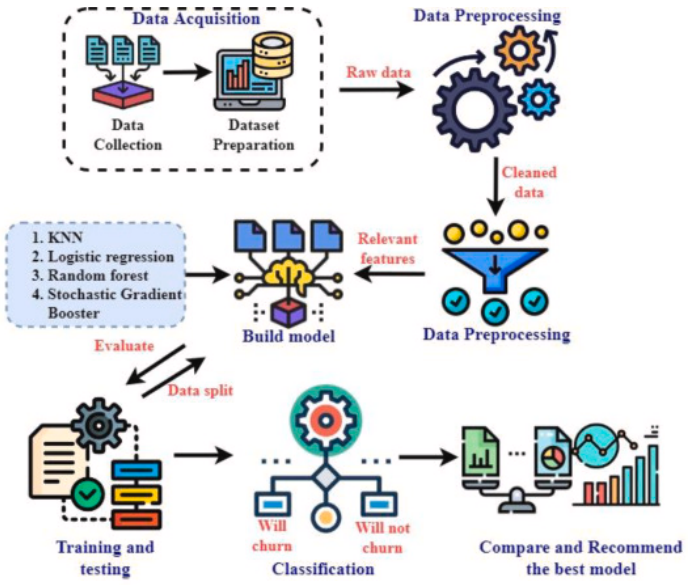

Fig. 1. System Layout

3 Objectives

1. **Anticipate Customer Automation:** The ultimate goal is to accurately anticipate bank customers who are as likely as possible to use the advanced machine learning model. This includes enhancing predictive segmentation models trained on historical customer statistics to identify key patterns and drivers of customer decline.
2. **Analyze Contributing Factors:** The goal of the program is to identify and select key factors contributing to customer success. By mastering those areas, banks can implement focused strategies to reduce customer loss.
3. **Model Comparison and Evaluation:** To study different system models, combined with Support Vector Machine (SVM), decision tree, and logistic regression, used to determine the most effective method for predicting customer churn and as a comparison. The overall performance of this model is fully evaluated on the metrics of accuracy, consensus, and F1-rating.

4 Datasets

These Datasets Are Bank Churn Analysis:

No	Dataset Name	Link
1	Bank Customer Churn	Link1
2	Credit Card Churn Prediction	Link2
3	Bank Customer Churn Data	Link3

Table 1. List of Datasets Used for Bank Churn Analysis

5 Types of Data

Statistical data includes customer's age, gender, income, and occupation, used to identify populations likely to leave the bank. Transactional data involves account activity, transaction details, withdrawals, and deposits to detect patterns of customer churn. Customer service interactions such as inquiries, complaints, and contacts impact customer attrition. Product and service usage details help evaluate the relationship between churn and product usage. Sentiment analysis data from reviews and feedback observe customer satisfaction.This section provides a comprehensive account of the study's methodology, the experimental setup, and the evaluation of the model's performance. It encompasses every phase and method used to conduct the research and evaluate its effectiveness.

6 Review of Datasets

Dataset 1: Bank Customer Churn

Introduction: In today's highly competitive banking environment, retaining customers is crucial and one of the core features that decide upon long-term success for banks. Your current customers are the easiest to service, and gaining new patronage is costly. Understanding the causes of customer churn allows a bank to develop new features and targeted actions aimed at satisfying customers. There are key datasets that tell us what factors contribute to customer churn, and focusing on those allows a financial analyst or another stakeholder to see the big picture[4].

Characteristics: RowNumber, CustomerId, and Surname are columns that serve as identifiers that are not used in predicting customer churn models. Credit score is an important variable because it is linked to lower churn rates. Geography is also an important variable because location can influence customers to churn and impact the decision to leave the bank. Gender is valuable for targeted marketing strategies. Longer tenure has a higher possibility of higher loyalty and lower churn rates. The balance variable represents the information about the customer's bank balance, which indicates financial stability because higher account balances correlate with lower churn rates. NumOfProducts, HasCrCard, and IsActiveMember provide insights into a customer's engagement with the bank[4].

Limitation: RowNumber, CustomerId, and Surname cannot be used in statistical calculations because they have no predictive value. The credit score is important because it is linked to lower churn rates, but it does not capture other financial behaviors that may influence churn. Geography may simplify complex regional variations and cultural factors. Gender can influence customers because specific gender populations may prefer different choices[4].

In conclusion, understanding customer churn by analyzing a large set of factors is crucial.

Dataset 2: Credit Card Churn Prediction

Introduction: In the competitive banking industry today, retaining customers is an important necessity. Knowing why customers leave helps make those initiatives targeted: it could be that loyalty programs need improvement. This analysis aims to predict and take action steps to reduce customer churn, which can help save long-term growth and profit. The ten columns in this dataset show important characteristics of the customers such as attrition, age, gender, dependents, level of education, marital status, income, card type, and ownership duration [5].

Characteristics: The dataset contains columns of the following characteristics: Client Number is a unique identifier for each client and has no use in model building. Attrition-Flag is the target categorical variable for churn prediction (yes, no). Customer-Age shows the frequency distribution of customers. Gender shows the gender distribution among customers. Dependent-Count indicates the

number of dependents a customer has. Education-Level provides information about the academic field of customers. Marital-Status indicates the marital status of the customers. Income-Category categorizes the customers according to their income. Card-Category shows the type of credit card the customer holds. Lastly, Months-on-Book indicates the number of months a customer has been with the company[5].

Limitation: Different attributes have limitations. Client Number provides only identification and no analytics. Attrition-Flag suffers from data imbalance issues. Customer-Age may have data holes due to not being real-time. Gender distribution may introduce bias if not properly balanced. Each column presents different complications for analysis and modeling[5].

Finally, understanding when a bank loses customers is crucial for improving services and retention. This dataset helps predict customer churn by analyzing various factors like age, income, and card preferences.

Dataset 3: Bank Customer Churn Data

Introduction: The purpose of this dataset is to forecast customer churn in the banking sector. The dataset contains data on 28,382 clients[6], covering banking signals, transaction history, and personal characteristics. Using this dataset, machine learning models can be built to predict whether a client will leave or stay at the bank, allowing proactive customer retention[6].

Characteristics: The dataset includes columns like Customer-ID, which is a unique identifier. Vintage informs about customer loyalty and retention time. Locations provide data on city and consumer behavior. Net Worth relates to financial stability and service tiers. Customer-New-Category indicates engagement levels. There are three branch codes relevant for trend analysis and forecasting financial sustainability[6].

Limitation: Each column has unique characteristics and limitations. Customer-ID is just an identifier with no analytical meaning. Vintage records duration but does not change with further analysis. Demographic context provided by Age and "Gender" may introduce bias. Columns like Current-Month-Credit or Current-Month-Debit are useful for behavior analysis but lack real-time perspective[6].

In short, this dataset is good for predicting client attrition. It provides various data points, transaction history, and account balances to create a prediction model that identifies customers at risk for churn.

7 Review of Algorithms or Techniques

Technique 1: Decision Tree

The dataset highlights the flexibility and transparency of selection timber in device mastering. Decision trees are celebrated for his or her simplicity, making them accessible to novices and specialists because of their visible tree-like

diagrams This readability offers an intuitive knowledge of the decision-making manner, of which want in areas together with financial institution churn forecasting. Despite this strength, selection bushes are regularly overdressed, mainly in complex systems that suit schooling facts nicely but war with different records. To be powerful in predictive modeling obligations, it's far essential to test how they tend to overfit.

Technique 2: Support Vector Machine (SVM)

The support Vector Machine or SVM is a classification technique that belongs to supervised learning in machine learning theory. It works by setting up an optimal hyperplane that divides the various classes, hence it is capable of dealing with linear as well as non-linear relationships. Within the banking churn prediction use case, SVM can perform the more complicated task of churning out customers that historically tend to violate bank policy cues; drawing on cover historical insight for early predictions and technically using feature engineering methods to ensure model accuracy. High accuracy Ability to build flexible models robust in high dimension space Good generalization performance with customers and unseen datasets was very effective. However, It has limitations such as it performs well only with good data quality and is computationally expensive, especially for large datasets. Even under these limitations, SVM remains a strong tool for predicting and reducing customer churn in the banking sector[8].

8 Result and Discussions of Datasets

Dataset 1: Bank Customer Churn

Dataset Summary

Description	Value
Number of Rows	10,000
Number of Columns	18
Target Column	IsActiveMember (categorical, integer)
Data Split	70% Training, 30% Testing
Models Used	SVM, Decision Tree
Null Values	No Null Values
Duplicate Values	None Found
Data Type for Each Column	Presented

Table 2. Summary of The Dataset Used For Bank Customer Churn Prediction

Performance Comparison of Precision, Recall & F1 Score:

Inactive Customers

Model	Precision	Recall	F1 Score
SVM	49%	16%	24%
Decision Tree	53%	55%	54%

Table 3. Performance Comparison for Inactive Customers

Active Customers

Model	Precision	Recall	F1 Score
SVM	52%	85%	64%
Decision Tree	56%	55%	56%

Table 4. Performance Comparison for Active Customers

Since the research problem prioritizes classification performance for active customers, the SVM model is considered more suitable. However, if a balanced approach is desired, especially concerning inactive customer detection, the Decision Tree model offers a compelling alternative.

Dataset 2: Credit Card Churn Prediction

Dataset Summary

Description	Value
Dataset Size	232,921 samples
Number of Rows	10,127
Number of Columns	23
Data Split	70% Training, 30% Testing
Duplicate Values	None
Models Used	SVM, Decision Tree
Target Column	Card Type
Models Used	SVM, Decision Tree

Table 5. Summary of The Dataset Used for Credit Card Churn Prediction

Performance Comparison of Precision, Recall & F1 Score:

SVM Technique

Card Type	Precision	Recall	F1 Score
Blue	93%	100%	96%
Gold	0%	0%	0%
Silver	0%	0%	0%
Platinum	0%	0%	0%

Table 6. Performance Comparison

Decision Tree Technique

Card Type	Precision	Recall	F1 Score
Blue	99%	98%	99%
Gold	39%	57%	47%
Silver	0%	0%	0%
Platinum	77%	73%	75%

Table 7. Performance Comparison

This structure will help communicate the results and the reasoning behind the conclusion that the Decision Tree model is more suitable for this dataset.

Dataset 3: Bank Customer Churn Data

Dataset Summary

Description	Value
Total Samples	596,022
Number of Rows	28,382
Number of Columns	21
Random State	42
Data Split	70% Training, 30% Testing
Duplicate Values	None
Models Used	SVM, Decision Tree
Target Column	Churned Customers Existing Customers

Table 8. Summary of The Dataset Used for Bank Customer Churn Data

Performance Comparison of Precision, Recall & F1 Score:

SVM Technique

Metric	Churned Customers	Existing Customers
Precision	87%	74%
Recall	97%	37%
F1-Score	92%	49%

Table 9. Performance Comparison

Decision Tree Technique

Metric	Churned Customers	Existing Customers
Precision	84%	44%
Recall	87%	45%
F1-Score	87%	44%

Table 10. Performance Comparison

Therefore, the SVM model is more suitable for this research problem, providing better classification performance overall.

9 Conclusion

Consequently, this research study greatly benefits the field by applying the data-driven approach and prescriptive models to estimate and control customer turnover in the Banking industry. Thus, making use of characteristics such as demographics transactions and balances brought into use models that could establish customers most likely to churn. Thus, it is possible to use machine learning approaches, for example, the method of support vector machines and decision trees, for the classification of such customers.

These results assert the importance of basing fixed data with evolving data in enhancing the cord model's reliability. The present study was useful in establishing the link between the level of economic information, which gives knowledge that would allow for more control and better-targeted efforts to maintain customers. The study also brings to light the relevance of statistical tools, in enhancing CRM in the context of banking.

In other words, this research not only indicates some of the reasons for customer attrition but also outlines a specific strategy that banks can use to alleviate turnover and strengthen the client base. With the help of analytics, the

banks save money and time and can improve factors associated with turnover problems, customer satisfaction, and therefore, performance.

References

1. M. Wiryaseputra, "Bank Customer Churn Prediction Using Machine Learning," Analytics Vidhya, p. 120, Oct. 12, 2022.
2. A. Cole, "Predicting Customer Churn With Classification Modeling," Towards Data Science, vol. 2, no. 3, p. 18, May 9, 2020.
3. P. Verma, "Churn Prediction for Savings Bank Customers: A Machine Learning Approach," Journal of Statistics Applications & Probability, vol. 9, no. 3, pp. 535-547, Nov. 1, 2020.
4. S. Shaibu, "Predicting Churn Rate in a Bank Using Artificial Neural Network with Keras," Medium, Apr. 5, 2022.
5. H. Tran, N. Le, and V.-H. Nguyen, "Customer Churn Prediction in the Banking Sector Using Machine Learning-Based Classification Models," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 18, pp. 87-105, Feb. 21, 2023.
6. Y. Kavyarshitha and V. Sandhya, "Churn Prediction in Banking Using ML with ANN," Institute of Science and Technology, Mar. 2022.
7. N. Saeed, "Bank Customer Churn Prediction Model," Medium, May 16, 2023.
8. R. Lopez, "Reduce Customer Churn in a Bank Using Machine Learning," Neural Designer, Aug. 31, 2023.
9. D. Al-Najjar, N. Al-Rousan, and H. Al-Najjar, "Machine Learning to Develop Credit Card Customer Churn Prediction," Theoretical and Applied Electronic Commerce Research, vol. 17, pp. 1529–1542, Nov. 16, 2022.
10. A. Widya, "Machine Learning Model Random Forest," Dqlab, vol. 2, no. 3, p. 18, Jul. 18, 2023.
11. M. T. Chowdhury, H. Rahman, M. I. Sumon and A. Talha, "Classification of satellite images with VGG19 and Convolutional Neural Network (CNN)," 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, p. 397-402, 2024.
12. Chowdhury, M.T. et al. (2024). Detecting Crop Pests and Diseases Through Deep Learning Techniques for Improved Yields. In: Dutta, S., Bhattacharya, A., Shahnaz, C., Chakrabarti, S. (eds) Cyber Intelligence and Information Retrieval. CIIR 2023. Lecture Notes in Networks and Systems, vol 1025, no. 18, p 343-478, 2021.