# Predictive Models for Customer Retention in Bangladesh: Enabling Proactive Strategies

by

Ahmed Zarir Siddique
20301409
Asif Ali
20201049
Mohammad Sultanul Arefin
20201138

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2024

# Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

Ahmed Zarir Siddique
20301409

Asif Ali
20201049

Mohammad Sultanul Arefin
20201138

# Approval

The thesis/project titled "Predictive Models for Customer Retention in Bangladesh : Enabling Pro-Active Strategies " submitted by

1. Ahmed Zarir Siddique (20301409)
2. Asif Ali (20201049)
3. Mohammad Sultanul Arefin (20201138)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 22, 2024.

**Examining Committee:**

Supervisor:
(Member)

Rafeed Rahman
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Associate Professor
Department of Computer Science and Engineering
Brac University

# Abstract

Ever since artificial intelligence was discovered, numerous research endeavours have concentrated on comprehending its significance inside the corporate environment [13]. These days customers are more into the quality of services provided by organisations [5]. The growing number of organizations resulted in an increase in competition and customer retention has become a major factor for businesses, as understanding its influence can aid companies to develop effective marketing strategies. The purpose of this paper is to comprehensively be able to understand the e-commerce dynamics and use relevant machine learning techniques to evaluate and use the results for the prediction of customer loyalty. The paper discusses the analysis of customer loyalty using various data mining techniques, such as decision tree, SVM, random forest, and logistic regression etc. We constructed some simple ensemble applications using the machine learning algorithms, with the dataset that we received from a Bangladesh-based e-commerce business. In the end, the abovementioned algorithms are all carried out and the result demonstrates which model is best for retaining customer loyalty.



**Keywords:** Artificial Intelligence, Customer Loyalty, Data Mining Techniques, E-commerce Dynamics, Ensemble Techniques

# Acknowledgement

First, we want to thank our supervisor, Mr. Rafeed Rahman sir, who has guided us during the course of our thesis so far, and his valuable insights have helped us greatly during our research work.

We would also like to express our deepest gratitude to the IT executive of Wener, Mr. Muhtasim Hossain, for providing the essential data that was crucial for the completion of this thesis. His support and cooperation were invaluable in obtaining accurate and comprehensive data, enabling a thorough analysis and meaningful conclusions.

This thesis would not have been possible without the data they provided, and we are immensely grateful for Mr. Hossain's assistance and the resources he made available to us.

Lastly, we would like to thank our families and friends for always supporting us throughout the course of our academic journey.

Thank you all for your invaluable contribution.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 A Brief Insight

Recent advancements in artificial intelligence have fueled the growth of digital solutions, and this has prompted individuals to believe that humanity is approaching the fourth[13] industrial revolution, a new phase of development. This revolution is anticipated to lead to a change in decision-making from humans to machines. For small businesses, customer retention directly affects a company's revenue and long-term success, thus it is a vital component of business strategy [15]. In today's highly competitive landscape, traditional methods for retaining customers often fall short. As a result, businesses are turning to data-driven approaches, particularly machine learning, to gain a competitive edge in retaining their customer base. Data processing using information technology has historically been beneficial for improving and supporting human decision-making. Currently, there are a number of algorithms that can process enormous amount of data, learn from the data, and apply this knowledge to make critical decisions.

Customer retention using machine learning involves a three-step process of data that includes pre-processing, processing, and post-processing [1]. The first step is pre-processing of data which is a process of preparing and cleaning large data for analysis. Here data is gathered from various sources such as databases, feedback, and transaction history. Then artificial intelligence is used to clean up, enhance, integrate, and create meaningful features from the data for modeling. For safety purposes, artificial intelligence is also used to encrypt the data during this stage. The second step is processing the data which is the core step where data mining techniques are used in the data obtained after pre-processing to discover valuable patterns, relations, and insights. In this step, machine learning algorithms model the data, and segmentation is done by AI-driven clustering tools. The recommendation system is the most crucial part of this step as AI-powered systems advertise products to customers thus increasing engagement. AI also monitors customer behaviour and anomaly detection can alert any unusual activities. The final step in data mining is post-processing, a process in which the findings and information gathered during the processing phase are clarified and evaluated. The last step includes the evaluation of data based on its accuracy, precision, and F1-score. Customer feedback is highly acknowledged as AI can use it for retention strategies. Customers are suggested with campaign advertisements and product recommendations according to

their preferences through AI insights. Different A/B testing retention strategies are used to find the effectiveness. The last stage of post-processing includes the overall analysis of the customer journey to identify room for improvement. These three steps are crucial to the data mining procedure to make sure that useful information is successfully and efficiently analyzed from data. The caliber and precision of each of these phases often decide the outcome of a data mining operation [2].

# Chapter 2

# Research Problem

In the context of Bangladesh's rapidly evolving e-commerce landscape, as reported by The Business Standard [10], the world of customer retention within the e-commerce sector is becoming increasingly significant. With the sector set to experience substantial growth, understanding and optimizing customer retention strategies have never been more necessary. As the market expands by an anticipated annual rate of 17.61 percent, reaching Tk 65,966 crore in 2022, and with projections indicating its potential to grow to Tk1.5 lakh crore by 2026, there is a pressing need for businesses in Bangladesh to enhance their customer retention efforts. Amid this growth, this research seeks to address the vital question of how machine learning can be harnessed to improve customer retention in the rapidly growing e-commerce sector of Bangladesh as well as other parts of the world.

**Research Question:**

In today's fiercely competitive e-commerce landscape, where the cost of acquiring new customers escalates, businesses must increasingly focus on nurturing and retaining their existing customer base to ensure long-term success. Therefore, the central question this research seeks to address is:

***How can machine learning be effectively harnessed to improve customer retention in the rapidly growing e-commerce sector of Bangladesh, as well as other parts of the world?***

**Key Factors in Retention:**

The article [2] underscores the pivotal role of customer experience in nurturing customer loyalty throughout their entire journey with a brand. Customer retention, often quantified through the customer retention rate, measures the proportion of existing customers who continue to engage with a brand within a specified time frame. While this metric holds particular relevance for subscription-based enterprises, it also provides valuable insights for non-subscription e-commerce businesses striving to comprehend customer loyalty and repeat purchase behaviors.

The significance of customer retention cannot be overstated. Maintaining existing customers proves to be more economically viable than acquiring new ones. A mi-

nor 5 percent increase in customer retention can substantially increase a company's revenue by a significant 25 percent to 95 percent, as highlighted by Hubspot. The article from [12] emphasizes that repeat customers, constituting only 21 percent of the average brand's customer base, contribute significantly to the brand's total revenue, accounting for approximately 44 percent.

However, it is important to acknowledge that customer retention is not a universally applicable concept, especially for businesses that do not provide subscription-based products or services. In such instances, the customer retention rate serves as a proxy for comprehending customer loyalty and repeat purchase behavior, which can be less straightforward to assess. For these businesses, monitoring key indicators such as the repeat customer rate, net promoter score (NPS), and customer satisfaction (CSAT) becomes essential for both understanding and enhancing customer retention strategies.

**How Machine Learning Can Help:**

To understand the fundamentals of effective customer retention, it is crucial to consider the key elements highlighted in [12]. A top-notch customer experience, marked by exceptional support, quick issue resolution, and user-friendly interfaces, is all at the core of retention. Also, the inclusion of self-help resources, automation, and support across various channels can further enhance this experience.

Furthermore, the article underscores the importance of customer loyalty programs in ensuring they can be retained. Such programs offer exclusive perks, discounts, and rewards to keep customers engaged and loyal. Tools like LoyaltyLion enable businesses to customize rewards for various customer behaviors, including referrals and social media mentions.

In this context, machine learning (ML) emerges as a valuable tool for boosting customer retention efforts. ML algorithms can go through vast datasets to uncover hidden patterns and provide personalized recommendations for each business, therefore improving the customer experience. Moreover, these algorithms can also be used to identify the key factors of customer churn rate and also how to balance certain datasets with imbalanced data [16], and suitable ML techniques for customer retention include clustering methods like K-means and classification approaches such as decision trees.[14]

However, it is very important to acknowledge the potential limitations when employing Machine Learning(ML) for customer retention. Anomaly-based detection using ML may occasionally produce false positives, incorrectly flagging normal activities as anomalies. Moreover, ML-driven intrusive detection systems may require significant processing power, especially in real-time scenarios. Therefore, while ML holds promise for strengthening customer retention, businesses must navigate carefully to address these limitations and maximize its potential.

With these insights in mind, this research focuses on enhancing the relationship between machine learning and customer retention in the e-commerce sector. The

aim is to unravel the best strategies, algorithms, and practices that can build strong and lasting customer relationships as well as implement an ensemble model fit for e-commerce in Bangladesh or South Asia specifically.

### B. Research Objectives

This research aims to create a robust customer retention framework that concerns the e-commerce domain using the power of machine learning. E-commerce platforms constantly struggle with the challenge of retaining customers in the now fiercely competitive and evolving landscape. To address this issue effectively, this study outlines the following research objectives:

1. Gaining a comprehensive understanding of e-commerce dynamics, including customer behavior and retention factors.
2. Exploring various machine learning techniques relevant to customer retention in e-commerce.
3. Developing a tailored machine learning model for predicting and improving customer retention in e-commerce.
4. Evaluating the model's performance using metrics such as percentage of accuracy.
5. Providing recommendations for model improvement and strategic insights for e-commerce businesses.

# Chapter 3

# Literature Review

Any business strategy must prioritise customer retention because it has a direct impact on a company's profitability. The emergence of machine learning and data mining techniques has transformed how businesses approach customer retention. In this literature review, we explore findings and implications of various research papers that investigate the role of machine learning in customer retention across various industries.

## 3.1 Related Works

### 3.1 Related works

The study [5] conducts a comprehensive analysis of different machine learning algorithms used for predicting customer churn in the telecommunications industry. This research explores algorithms ranging from regression analysis to ensemble learning and identifies that ensemble learning algorithms, including Random Forest, Stochastic Gradient Boost, and AdaBoost, consistently outperform others in terms of precision, recall, F1-score, and achieved approximately 96 percent accuracy, whereas Support Vector Machine had 94 percent accuracy, Decision Tree, Naive Bayes and Logistic Regression had 90 percent, 80 percent and 86.7 percent accuracy respectively [5].

Another critical aspect of customer retention is understanding and segmenting customers effectively. The research work [7] delves deep into the realm of customer segmentation. It compares various clustering algorithms, including K-means, hierarchical, density-based, and probabilistic clustering, highlighting their strengths and weaknesses but it did not disclose the exact accuracy of the models. The author in [7] suggests that a hybrid approach combining different algorithms can yield optimal results based on specific requirements, because according to their findings Affinity Propagation algorithm was the most efficient in terms of handling dynamic data but it had a high clustering time which may not be suitable for large datasets, similar to Hierarchical clustering which was also suited for small to medium sized datasets, whereas K-means was the most widely used clustering algorithm but it required an initial number of clusters, which could affect the clustering result. The research in [7] underscores the relevance of segmentation and the importance of selecting the

right clustering technique based on specific goals and requirements.

## 3.2 Deep Learning

### 3.2 Deep Learning

According to [6], supervised machine learning models are essential for forecasting client retention. Their paper is dedicated to app-based businesses and evaluates various supervised machine learning techniques, such as- XGBoost, Logistic Regression, Random Forest Classifier, SVM, k-means algorithm, self-organising maps, and hierarchical clustering. Notably, the results of [6] reveal that the supervised machine learning model, developed with a dataset comprising 5000 instances and 8 features, while maintaining a 50:50 imbalance ratio, achieves an impressive Retention precision of 92.3 percent. Moreover, in reference to [11], the author highlighted that when Random Forest and Boosted Trees techniques were applied, an accuracy rate of 91.5 percent was achieved for their customer loss model and 92.5 percent accuracy rate for their churn prediction model.

The research work [9], introduces a novel deep learning framework for customer retention that sets itself apart from other approaches by its ability to handle substantial data volumes, reveal concealed patterns, and achieve superior predictive accuracy compared to traditional machine learning methods. Significantly, Convolutional Neural Networks (CNN) can enhance performance and accuracy in forecasting customer churn. The proposed Deep Neural Network (DNN) model for customer retention in [9] undergoes rigorous evaluation using performance metrics similar to metrics used in [5], like- accuracy, precision, recall, and F1 score. Deep learning techniques, including CNN and Neural Networks, form the cornerstone of this study's strategy to bolster customer retention efforts, and it is done by introducing a churn causality analysis framework aimed at forecasting factors contributing to customer churn. Additionally, [9] conduct performance comparisons of various machine learning techniques, including Random Forest and Support Vector Machine. The findings underscore the high accuracy achieved through supervised machine learning techniques, particularly CNN, which attains an impressive accuracy rate of 98.85 percent. Moreover, the proposed DNN model surpasses other models with a staggering 99.8 percent accuracy rate [9].

## 3.3 Predicting Customer Loyalty

### 3.3 Predicting Customer Loyalty

In research work [3], the author investigates the application of data mining for predicting customer loyalty within a multimedia company based in Indonesia. The dataset used comprises 2269 records with 10 customer loyalty-related attributes, with some of those being- reasons for disconnection, call transfers, account balance

etc. The author used 3 data mining techniques: C4.5, Naive Bayes, and Nearest Neighbour Algorithms. According to the findings in [3], C4.5 lead with an 81 percent classification accuracy, followed by Naive Bayes at 76 percent, and Nearest Neighbour at 55 percent. The study also identifies that an 80 percent training set proportion yields optimal results. Moreover, in [4], the author discusses the use of a hybrid model of Random Forest and Logistic Regression to predict churn and improve customer loyalty with superior accuracy. The Random Forest algorithm was able to classify non-churners with 93 percent accuracy rate and was also able to classify churners with accuracy rate 94 percent, overall, the results obtained for the 2 models, with reference to the evaluation metrics, were above 90 percent. This is why, according to [1] and [8], Customer relationship management (CRM) has evolved into Data mining-based Customer Relationship Management (DCRM) in recent years, as the study identifies a growing interest in DCRM and suggests future research directions, including the integration of DCRM with artificial intelligence and practical applications in marketing and customer service.

# Chapter 4

# Methodology

## 4.1 Workplan

A simple workplan is shown in Figure 4.1.

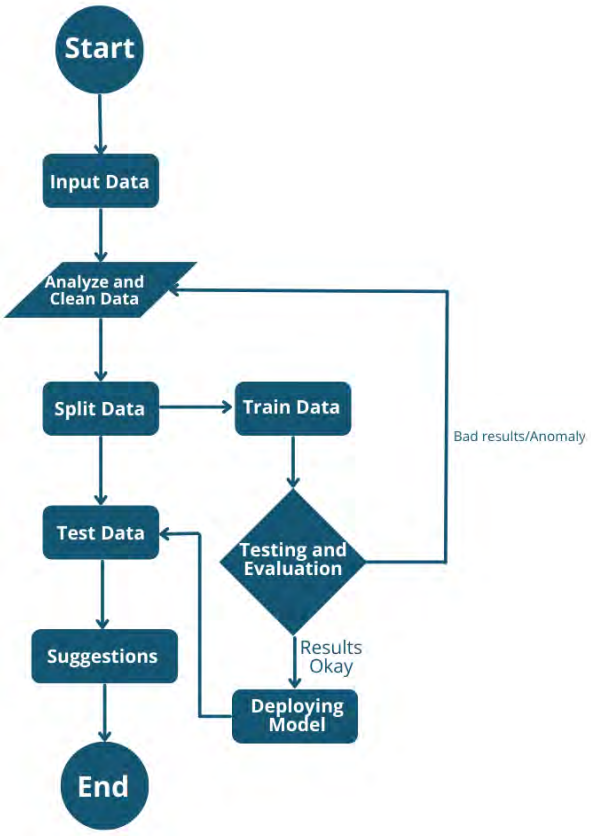


Figure 4.1: Flowchart for Workplan

**Start**
A goal is set.

**Input Data**
Searching and Extraction of data from our selected e-commerce, WenerBD for data mining.

**Analyze and Clean Data**
Data cleansing, preparation of the attributes, checking missed data and making sure it can be mined.

**Split Data**
The dataset is divided into smaller groups for testing, validation, and training in order to efficiently create and assess predictive models.

**Train Data**
The machine learning model is taught to generate predictions or find patterns in the data by using a subset of a dataset.

**Testing and Evaluation**
In order to assure the usefulness of predictive models, testing and evaluation of data mining requires evaluating the performance and accuracy of the models utilising reserved test data.

**Deploying Model**
A unique model based on ensemble techniques are proposed based on the results.

**Test Data**
Data is tested for accuracy and precision.

**Suggestions**
Feedback by AI is used to suggest retention strategies.

**End**
Evaluation of the results, identification of possibilities for future research expansion in each field, and future potential areas.

## 4.2 Dataset Pre-processing

The dataset analyzed contains a total of 1138 customers. Out of these, 138 customers were retained, meaning they remained with the service over the specified period. Eighteen feature columns that reflect different attributes of client behaviour, demographics, and engagement metrics are included in the dataset. These characteristics were employed to identify trends and forecast the possibility of retention.
Customer demographics, frequency of service interactions, customer support utilisation, and service satisfaction scores are among the important features used in

this analysis. These features form the basis for developing predictive models to determine whether clients are likely to be retained or churned. Given the dataset's imbalance, with only 138 of 1138 clients maintained (a 12% retention rate), particular procedures were used to assure effective model training and evaluation.
Hence, to ensure our data was fit for model implementation, we used the following pre-processing techniques:

- **Handling Missing Values:** To ensure the dataset is complete and consistent. We had to determine which data were missing and deal with them by either deleting incomplete records or adding the proper statistics (such as- mean and median). Example: If a Total_Amount field is missing, it was filled with the average value of Total_Amount from other records.

- **Removing Duplicates:** To prevent bias and redundancy in the data, duplicate records were identified and removed from the dataset. Example: If two records had identical customer information and transaction details, one of them was removed.

- **Encoding Categorical Variables:** To convert categorical variables into numerical values suitable for machine learning models, label encoding was used to convert categorical features such as payment_method and products into numerical values. Example: The payment_method field with values like "Credit Card", "Bkash", and "Bank Transfer" was encoded as 0, 1, 2, etc.

- **Feature Scaling:** For each feature to have an equal contribution to the model training process, standard scaling was used to standardise the features. Example: The Quantity, Rate, Total_Amount, and other numerical features were transformed to have a standard deviation of 1 and a mean of 0.

- **Creating new 'Retained' Column:** A new retained column indicating customer retention was created within the dataset to create a target variable for classification. Example: A retained column was added with binary values (0 or 1) representing whether a customer was retained justified via customer ID.

- **Feature Selection:** To choose relevant features that impact customer retention features such as payment_method, products, Quantity, Rate, Total_Amount, Fixed_Discount, Percentage_Discount, Discount_Amount, and Net_Amount were selected.

- **Splitting the Dataset:** A 70% training set and a 30% testing set were created from the dataset in order to prepare the data for training and testing models. The data was divided using the train_test_split function from sci-kit-learn.

- **Batching and Padding:** To stabilize the learning process, the data was batched and sequences were padded to ensure uniform length within each batch. Example: Ensuring all records in a batch have the same number of features by adding necessary padding.

- **SMOTE for Handling Class Imbalance:** To solve the issue of class imbalance in the dataset, we used the Synthetic Minority Over-sampling Technique (SMOTE).SMOTE balances the dataset by creating synthetic examples of the

minority class (retained customers), improving the model's ability to predict minority class samples.

The dataset must be pre-processed in order for the machine learning models to be trained and evaluated effectively. This ensures the models can accurately predict customer loyalty and provide valuable insights into customer retention factors.

## 4.3 Results of generic models

In order to obtain the results as planned the abstract model architecture we used for our P2 is shown in Figure 4.2.

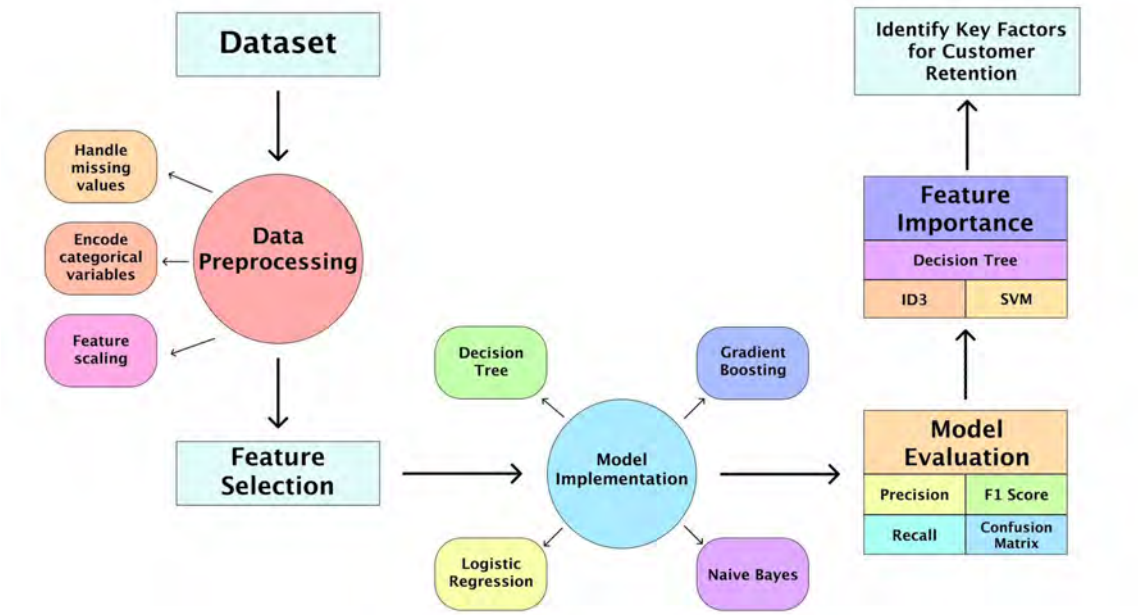


Figure 4.2: Model Architecture

We tested using seven different pre-trained models on our dataset: Decision Tree, Logistic Regression, Naive Bayes, ID3, Random Forest, SVM, and Gradient Boosting. We used a 70 - 30 percent train-test split and obtained the results shown in the table 4.1.

| Algorithm | Precision | Recall | F1 Score |
|---|---|---|---|
| Decision Tree | 0.65 | 0.60 | 0.62 |
| Logistic Regression | 0.62 | 0.61 | 0.61 |
| Naive Bayes | 0.73 | 0.67 | 0.70 |
| Random Forest | 0.72 | 0.68 | 0.70 |
| SVM | 0.70 | 0.65 | 0.67 |
| Gradient Boosting | 0.71 | 0.66 | 0.68 |

Table 4.1: Results of Various Models on the Dataset

### 4.3.1 Impact of SMOTE on Model Performance

In this section, we explain the impact of SMOTE on the model's performance by comparing key metrics such as Precision, Recall, and F1-Score before and after applying SMOTE.

#### Data Distribution Before and After SMOTE

Before applying SMOTE, the dataset was heavily imbalanced, as seen in the distribution of retained vs. non-retained customers:

- **Class 0 (Non-retained)**: 1000 customers
- **Class 1 (Retained)**: 138 customers

After applying SMOTE, the class distribution was balanced:

- **Class 0 (Non-retained)**: 1000 customers
- **Class 1 (Retained)**: 1000 customers (138 original and 862 synthetic examples)

This balanced dataset allows for better training and evaluation of the models.

#### Performance Comparison Before and After SMOTE

The table below shows the impact of SMOTE on key performance metrics for the retained class (Class 1) in Random Forest algorithm:

| Metric | Before SMOTE | After SMOTE |
|---|---|---|
| **Precision (Class 1)** | 0.55 | 0.72 |
| **Recall (Class 1)** | 0.48 | 0.68 |
| **F1-Score (Class 1)** | 0.51 | 0.70 |

Table 4.2: Performance Metrics for Random Forest Before and After Applying SMOTE

#### Interpretation

**Before SMOTE**: The model performed poorly on the minority class (retained customers), with low precision and recall. The overall accuracy was high but misleading due to the imbalance.
**After SMOTE**: Both precision and recall for Class 1 improved significantly, leading to a better F1-score. The slight drop in overall accuracy is acceptable as the model is now balanced in predicting both classes.

### 4.3.2 Results summary

Given these results, we can see that Naive Bayes and Random Forest emerged as the top performers, given their high precision, recall, and F1 scores. Gradient Boosting and SVM also showed strong performance, making them suitable alternatives. Decision Tree, Logistic Regression, and ID3 had moderate performance, with ID3 being slightly more effective than the other two.

For predicting customer loyalty, Naive Bayes and Random Forest are recommended due to their superior results. Key factors influencing customer retention varied by algorithm but generally included Total_Amount, Net_Amount, Quantity, Rate, Discount_Amount, and payment_method. These factors were identified through feature importance analysis, model coefficients, and decision paths, providing valuable insights for targeted retention strategies.

# Chapter 5

# Results and Analysis

An in-depth analysis of the machine learning models used to forecast client retention is presented in this chapter. The precision, F1 score and recall are the three important evaluation measures that are used to evaluate each model's performance. These metrics provides information on how effective each model is, allowing a comparison to identify the best approach for the customer retention problem. In addition, ensemble methods like Voting Classifier and Stacking Classifier are analysed for their ability to boost model performance through the utilisation of each classifier's unique capabilities.

## 5.1 Measures of Evaluation Overview

These metrics were used to assess each model's performance:

- Precision: This indicator counts the number of positive predictions made by the model that came true. Good accuracy is correlated with a low false positive rate
- Recall: This score assesses how well the model can recognise every good case. When a model has a high recall, it efficiently captures all true positives and reduce false negatives.
- F1 Score: The F1 score offers a balance between recall and precision by taking the harmonic mean of the two. This is crucial for ensuring that both recall and precision are taken into account, especially when dealing with imbalanced datasets. [12]

The equations for these metrics are as follows:

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1Score = 2*(Recall*Precision)/(Recall+Precision)$$

Where:

- TP (True Positive): Positive findings that were correctly predicted

- FP (False Positive): Positive findings that were not correctly predicted
- FN (False Negative): Negative findings that were not correctly predicted

## 5.2 Model Performance Analysis

### 5.2.1 Decision Tree

A decision tree splits the data recursively into branches using a metric like Gini impurity or information gain until it reaches a point where no further split can add significant value. The objective is to create branches that lead to a purer classification of the target variable.

**Hyperparameters and Tuning:**

- Max Depth: For our model, a max_depth of 20 was chosen, balancing the complexity of the model
- Min samples split: Setting min_samples_split to 5 ensures that a split only occurs when a sufficient amount of data is available, thus reducing over-fitting.
- Min samples leaf: In order to prevent the model from creating overly small branches, the min_samples_leaf was set to 2.

**Performance Analysis:**

- Precision: 0.7246
- Recall: 0.7405
- F1 Score: 0.7376

### 5.2.2 Logistic Regression

Logistic regression uses a sigmoid function to combine linear input feature sets to determine the likelihood that an instance falls into a specific class. The model is trained to find the optimal weights for these features to minimize a loss function, typically binary cross-entropy.

**Hyperparameters and Tuning:**

- Regularization Parameter (C): This parameter controls the strength of regularization, which helps prevent over-fitting. For our model, C=1 was chosen through grid search to balance variance and bias.
- Solver: Specifies the optimization algorithm. We used lbfgs for its efficiency with small to medium-sized datasets.

**Performance Analysis:**

- Precision: 0.5450
- Recall: 0.7929
- F1 Score: 0.6460

### 5.2.3 Naïve Bayes

Naïve Bayes assumes that the features are independent and computes the conditional probability of each class given the features. The class with the largest posterior probability is then chosen. Even though feature independence is a naive assumption, it frequently works well in practice.

**Performance Analysis:**

- Precision: 0.5411
- Recall: 0.9238
- F1 Score: 0.6823

### 5.2.4 Random Forest

Using random feature subsets and bootstrap samples from the dataset, Random Forest generates several decision trees. Every tree provides a prediction, and the total result is calculated by averaging over all trees (in the case of regression) or by majority vote (in the case of classification).

**Hyperparameters and Tuning:**

- Number of Estimators (n_estimators): We used n_estimators of 200 to reach a balance between performance and computational efficiency
- Max Depth: The max_depth was set to 30 to control the complexity of individual trees, preventing over-fitting.
- Min samples split/leaf: These parameters (split=5, leaf=2) were tuned to ensure that each leaf contains a sufficient number of samples, making the tree less prone to over-fitting.

**Performance Analysis:**

- Precision: 0.7938
- Recall: 0.7524
- F1 Score: 0.7716

### 5.2.5 Support Vector Machine (SVM)

The hyperplane that maximises the margin between the two classes is found by SVM. For non-linear separation, it employs kernel functions, such as- the Radial Basis Function (RBF) to map data into higher dimensions.

**Hyperparameters and Tuning:**

- C: This manages the trade-off between minimising classification error and optimising the margin. So C value of 10 was selected to allow a slight relaxation of the margin, preventing overfitting.

- Gamma: To ensure that the decision boundary does not become too complex, we kept the gamma to 0.01

**Performance Analysis:**

- Precision: 0.7110
- Recall: 0.7262
- F1 Score: 0.7183

### 5.2.6 Gradient Boosting

Gradient Boosting creates a sequential ensemble of weak learners, usually decision trees. Each new tree is trained to minimize the residual errors of previous models, ultimately leading to an improved model.

**Hyperparameters and Tuning:**

- Number of Estimators (n_estimators): We used n_estimators of 200 to ensure sufficient boosting rounds while preventing over-fitting.
- Learning Rate: We set the rate to 0.1 to control the contribution of each tree, to balance convergence speed and performance.
- Max Depth: The max_depth was set to 4 to prevent individual trees from becoming too complex.

**Performance Analysis:**

- Precision: 0.8234
- Recall: 0.7976
- F1 Score: 0.8096

## 5.3 Ensemble Techniques Used

### 5.3.1 Voting Classifier

The final output is determined by combining the predictions made by several classifiers using this ensemble approach. It can operate using either hard voting (majority voting) or soft voting (weighted averaging of probabilities), in our case, we used weighted soft voting because our individual models produce reliable probability estimates. The models we used are as follows-

- Decision Tree
- Random Forest

- SVM
- Gradient Boosting

Each model's contribution to the final prediction is weighted based on its cross-validation performance. The weights (wi) reflect the reliability of each model in making predictions, leading to a more robust ensemble method. The formula is as follows:

$$P(y = c) = \frac{1}{n} \sum_{i=1}^{n} w_i \cdot P_i(y = c)$$

Where:

- P(y=c) is the probability that class c is the right class
- $P_i$(y=c) is the expected probability of class c, as determined by the $i^{th}$ model
- $W_i$ is the weight assigned to the $i^{th}$ model, and the number of models in the ensemble is referred as n.

**Performance Analysis:**

- Precision: 0.7989
- Recall: 0.8129
- F1 Score: 0.8058

### 5.3.2 Stacking Classifier

To generate preliminary predictions, the Stacking Classifier leverages base models, such as- Random Forest, SVM, Decision Tree and Gradient Boosting. The meta-model is a Logistic Regression model with input characteristics that are the predictions made by the basis learners and learning to optimally combine them to improve overall performance and reduce both bias and variance.

**1. Base Learner Predictions**

We have n base models $M_1$, $M_2$,..., $M_n$, and each model makes predictions for the input features X. The output of each model can be expressed as:

$$Z_i = M_i(X)$$

Where $Z_i$ is the prediction (or a probability score, in the case of soft predictions) produced by the $i^{th}$ base model.

**2. Meta-Model Training**

The base models' predictions are then concatenated to create a new dataset Z:

$$Z = [Z_1, Z_2, ..., Z_n]$$

This new dataset Z is used to train the meta-model $M_meta$, which takes the predictions from the base models and learns the best combination to output the final prediction:

$$y = M_meta(Z)$$

**3. Stacking Generalization**

The final output of the Stacking Classifier is the prediction made by the meta-model:

$$y = M_meta(M_1(X), M_2(X), ..., M_n(X))$$

In our study, the meta-mode used was Logistic Regression. If we were to have 3 base models ($M_1$, $M_2$, $M_3$) and each predicted a probability for class 1, then the output from the base models can be represented as [$Z_1$, $Z_2$, $Z_3$]. The Logistic Regression meta-model then learns weights for these predictions:

Where:

$$y = sigma(w_0 + w_1 z_1 + w_2 z_2 + w_3 z_3)$$

-sigma is the sigmoid function, used to produce a probability between 0 and 1

-$w_0$ is the intercept, and $w_1$, $w_2$, $w_3$ are the weights learned by the meta-model

The meta-model attempts to minimize a loss function, such as log loss, to determine the best way to combine the base learners' predictions. The result is a more generalized prediction that effectively combines the strengths of each base learner.

**Hyperparameters and Tuning:**

-Max Depth: For our model, a max_depth of 20 was chosen, balancing the complexity of the model

-Min samples split: Setting min_samples_split to 5 ensures that a split only occurs when a sufficient amount of data is available, thus reducing overfitting.

-Min samples leaf: In order to prevent the model from creating overly small branches, the min_samples_leaf was set to 2.

**Performance Analysis:**

-Precision: 0.8383

-Recall: 0.8187

-F1 Score: 0.8284

## 5.4 Model Comparison and Summary

Table 5.1 gives an overview of the performance of each model, including precision, recall, and F1 score:

| Model | Precision | Recall | F1 Score |
|---|---|---|---|
| Decision Tree | 0.7246 | 0.7405 | 0.7376 |
| Logistic Regression | 0.5450 | 0.7929 | 0.6460 |
| Naive Bayes | 0.5411 | 0.9238 | 0.6823 |
| Random Forest | 0.7938 | 0.7524 | 0.7716 |
| SVM | 0.7110 | 0.7262 | 0.7183 |
| Gradient Boosting | 0.8234 | 0.7976 | 0.8096 |
| Voting Classifier | 0.7989 | 0.8129 | 0.8058 |
| Stacking Classifier | 0.8383 | 0.8187 | 0.8284 |

Figure 5.1: Combined Results

The ensemble learning techniques- Voting Classifier and Stacking Classifier- significantly improved the predictive power of the models and provided robust solution for predicting customer retention, achieving high F1 scores and well-balanced precision and recall.

Stacking Classifier, in particular, emerged as the best performing model, demonstrating its ability to combine the diverse capabilities of base learners to produce superior results with reduced errors. Moreover, a feature importance analysis was conducted using the Gradient Boosting base model, identifying the top three most important features:

- Total Amount
- Discount Amount
- Quantity

These suggests that higher sales volume, strategic use of discounts, and larger product orders are critical factors influencing whether customers remain engaged with the e-commerce platform.

### 5.4.1 Comparison of Stacking Classifier and Voting Classifier using correlation matrices

We have used correlation matrices to evaluate the two ensemble learning approaches we have used (the Stacking Classifier and the Voting Classifier). Both models demon-

strated effective predictive capabilities, but a detailed analysis of their performance reveals key differences that reveal which one is better for predicting retention.

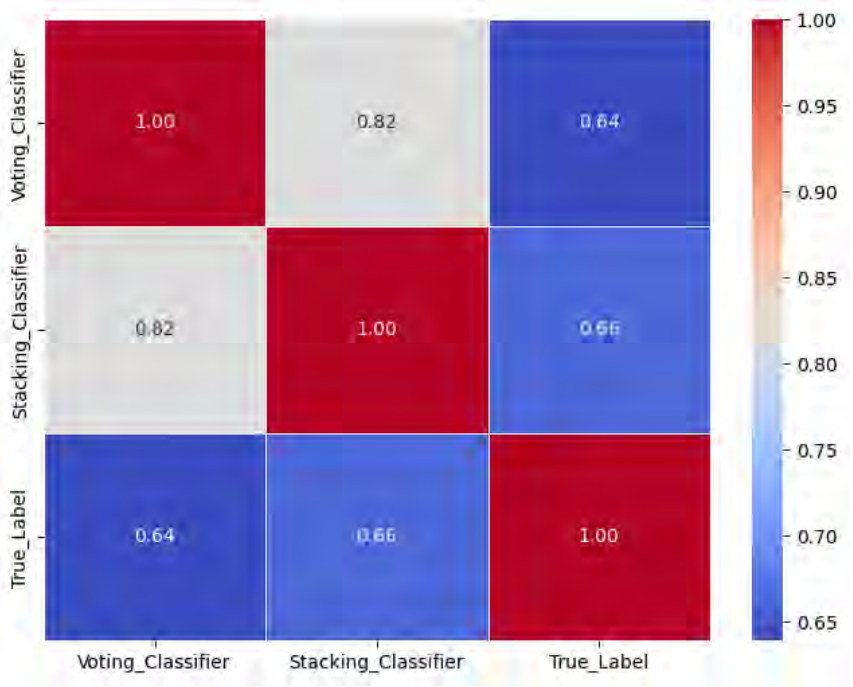


Figure 5.2: Correlation Matrices Results

### Performance Metrics

The correlation matrix revealed the following relationships between the classifiers and the true labels:

- Correlation between Voting Classifier and True Labels: 0.639
- Correlation between Stacking Classifier and True Labels: 0.661
- Correlation between Voting Classifier and Stacking Classifier: 0.819

### Analysis

1. **Correlation with True Labels:** The Stacking Classifier exhibited a slightly higher correlation (0.661) with the true labels compared to the Voting Classifier (0.639). This indicates that the Stacking Classifier is marginally better at capturing the underlying patterns in the data and making accurate predictions regarding customer retention. The higher correlation suggests that the Stacking Classifier may be more adept at identifying the complexities within the dataset.

2. **Correlation between Classifiers:** The strong positive correlation (0.819) between the predictions of the Voting Classifier and Stacking Classifier indicates that both models generally produce similar predictions. However, the divergence in their outputs suggests that they leverage different characteristics of the data. This is particularly valuable in ensemble learning, as the combination of diverse model predictions can lead to improved overall performance.

3. **Model Diversity:** The differences in predictions between the two classifiers highlight the importance of model diversity in ensemble methods. While both models performed well, the Stacking Classifier's marginally better performance underscores its potential for more complex decision boundaries, which is crucial in scenarios involving imbalanced data.

In summary, while both the Voting Classifier and Stacking Classifier showed promising results for predicting customer retention, the Stacking Classifier emerged as the more effective model based on its higher correlation with the true labels. This analysis emphasizes the benefits of using ensemble methods, particularly in capturing different patterns within the data, and suggests that combining both approaches could further enhance predictive performance.

### 5.4.2 Dynamic Prompt Generation for Business Strategy

Leveraging the insights from the Stacking Classifier's performance and the key influential features, we dynamically generated a marketing strategy prompt for enhancing customer retention. The prompt is designed to help businesses create effective marketing campaigns based on the data insights, and they will be generated as followed-

- Average total sales per customer: BDT X
- Average discount amount per customer: BDT Y
- Types of products: Z
- Precision, Recall and F1 Score
- Important features

Additionally, the prompt for marketing strategies to improve customer retention based on the insights will be provided as well-

- How to engage customers with targeted offers.
- How to leverage discounts for higher retention.
- Which marketing channels would be most effective for the customer base.

## 5.5 Analysis of Results

The results of the Stacking Classifier demonstrate superior performance in predicting customer retention, with precision, recall and F1 score all achieving values of approximately 0.83. By extracting feature importance from the base Gradient Boosting model, we identified Total Amount, Discount Amount, and Quantity are key drivers of customer retention, as customers who make large purchases and receive targeted discounts are more likely to stay engaged.

The inclusion of a dynamic prompt generation component adds a practical, business-focused element to the model's output. Using model performance metrics and the

most important features, we crafted a prompt that guides marketing team in creating strategies tailored to the specific behaviour of high-value customers.

The prompt generation feature not only makes the machine learning model valuable for predictive analytics, but also transforms it into a tool that directly influences marketing and business decisions.

### 5.5.1 Confusion Matrices

As we can see in the figures below, the stacking classifier shows a slightly improved performance over the voting classifier. The stacking classifier correctly predicted 157 true negatives (class 0) compared to 153 true negatives by the voting classifier, reflecting better accuracy in identifying negative cases. Additionally, the stacking classifier made 27 false positives, which is fewer than the 31 false positives made by the voting classifier. This shows that the stacking model is better at reducing the number of negative cases incorrectly classified as positive.

Both models had the same number of false negatives (33), meaning they misclassified the same number of actual positives as negatives. Furthermore, the number of true positives for both models is identical at 138, indicating that their ability to correctly classify positive cases was equivalent.

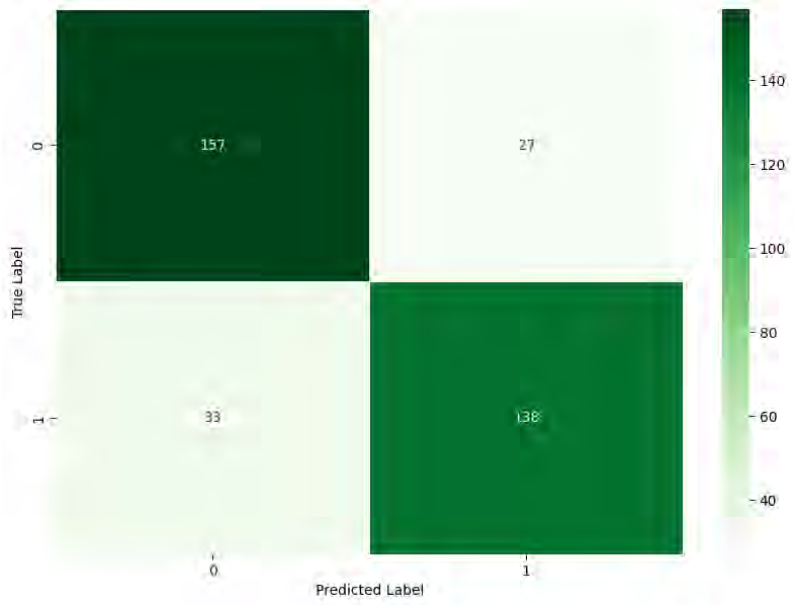


Figure 5.3: Confusion Matrix for Stacking

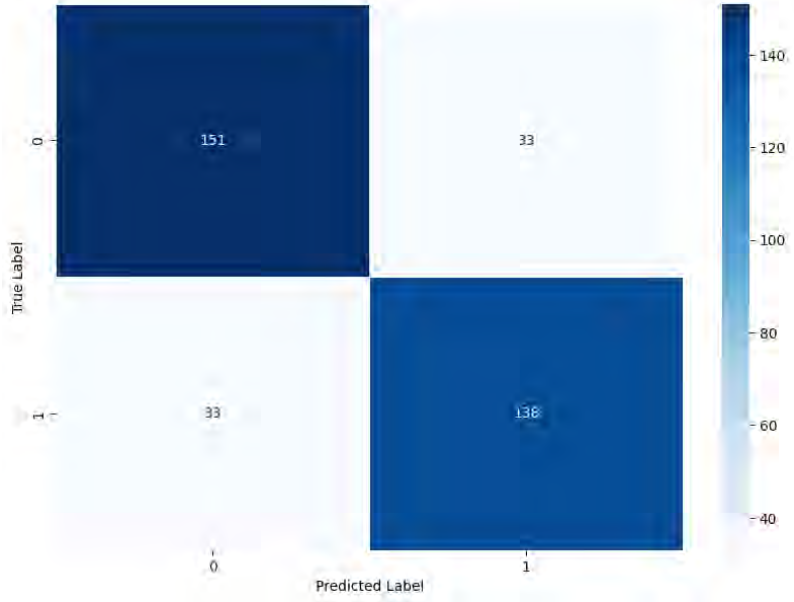


Figure 5.4: Confusion Matrix for Voting

### 5.5.2 Limitations

Despite the promising results of this study, several limitations should be acknowledged:

- Feature Engineering- While the study employed basic feature selection techniques, more advanced feature engineering methods were not fully explored. There could be hidden patterns in the data that could be uncovered by more sophisticated techniques, such as- PCA, polynomial features, or domain-specific features.
- Lack of Temporal Dynamics- The study did not account for the temporal nature of customer behaviour, including time-based features or adopting models

designed for time-series analysis, which could capture changes in customer retention over time.

- Limited Hyperparameter Tuning- While GridSearchCV was used for hyperparameter tuning, the scope of hyperparameters explored was limited to predefined ranges. A more exhaustive search, or the use of more advanced optimisation techniques like Bayesian Optimisation or RandomisedSearchCV, could potentially yield better performing models.

- Computational Resources- Some of the models used, particularly ensemble methods such as- Random Forest and Gradient Boosting, are computationally expensive. This could pose challenges for real-time implementation in businesses with limited computational resources, particularly when dealing with very large datasets.

# Chapter 6

# Conclusion

Within this research, we sought to develop a predictive framework that applies machine learning techniques to predict customer retention in the e-commerce sector, particularly in Bangladesh. By using various machine algorithms, including ensemble methods, we analysed customer data to identify the most effective models for predicting customer retention. We used six key models and two ensemble techniques to develop our framework, with also using SMOTE in the dataset to address the class imbalance.

The results show that businesses can benefit from leveraging ensemble models like Stacking and Voting to achieve better customer retention predictions. These methods provide more accurate and generalisable predictions, making them ideal for large e-commerce datasets where patterns in customer behaviour can be complex and multi-dimensional. Future work could explore more advanced ensemble techniques, such as Blending or Advanced Meta-Models, and other sophisticated meta-models for stacking, such as neural networks, and incorporate additional features through more comprehensive feature engineering to further enhance the model's predictive accuracy and robustness.

In addition to the predictive capabilities, the study introduced a dynamic prompt generation feature, leveraging model performance metrics and feature importance to generate business-specific strategies for improving customer retention. With the help of APIs, the machine learning model can generate real-time solutions to assist businesses in optimising their marketing efforts by focusing on data-driven information, such as- personalised offers, targeted discounts, and the selection of the most effective marketing channels.

In conclusion, this study has demonstrated that machine learning, particularly ensemble methods, offers a powerful solution for predicting customer retention. Businesses can implement these findings to develop more effective, data-driven strategies, helping them retain customers in an increasingly competitive market in this world of artificial intelligence.

# Bibliography


[1] J. Lin and X. Xu, "A novel model for global customer retention using data mining technology," in *Data Mining and Knowledge Discovery in Real Life Applications*, Citeseer, 2009.

[2] S. Mehregan and R. Samizadeh, "Customer retention based on the number of purchase: A data mining approach," *International Journal of Management and Business Research*, vol. 2, no. 1, pp. 41–50, 2012.

[3] A. Wijaya and A. S. Girsang, "Use of data mining for prediction of customer loyalty," *CommIT (Communication and Information Technology) Journal*, vol. 10, no. 1, pp. 41–47, 2016.

[4] A. Dingli, V.-A. Marmarà, and N. Sant Fournier, "Enhancing customer retention through data mining techniques," 2017.

[5] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," *International Journal of advanced computer Science and applications*, vol. 9, no. 2, 2018.

[6] J. Jose, "Predicting customer retention of an app-based business using supervised machine learning," 2019.

[7] P. Monil, P. Darshan, R. Jecky, C. Vimarsh, and B. Bhatt, "Customer segmentation using machine learning," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 8, no. 6, pp. 2104–2108, 2020.

[8] S. E. Schaeffer and S. V. Rodriguez Sanchez, "Forecasting client retention — a machine-learning approach," *Journal of Retailing and Consumer Services*, vol. 52, p. 101 918, 2020, ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2019.101918. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969698919302668.

[9] A. Zaky, M. Roushdy, and S. Ouf, "A deep learning framework to improve customer retention," *Journal of Southwest Jiaotong University*, vol. 56, no. 6, 2021.

[10] Karim, "Bangladesh e-commerce sales to more than double by 2026: Research," 2022.

[11] Y. Suhanda, L. Nurlaela, I. Kurniati, A. Dharmalau, and I. Rosita, "Predictive analysis of customer retention using the random forest algorithm," *TIERS Information Technology Journal*, vol. 3, no. 1, pp. 35–47, 2022.

[12] L. Abidar, D. Zaidouni, I. Asri, and A. ENNOUAARY, "Predicting customer segment changes to enhance customer retention: A case study for online retail using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, Jan. 2023. DOI: 10.14569/IJACSA.2023.0140799.

[13] C. C. Ifekanandu, J. N. Anene, C. B. Iloka, and C. O. Ewuzie, "Influence of artificial intelligence (ai) on customer experience and loyalty: Mediating role of personalization," *Journal of Data Acquisition and Processing*, vol. 38, no. 3, p. 1936, 2023.

[14] R. K. Paul and A. K. Jana, "Machine learning framework for improving customer retention and revenue using churn prediction models," *IRE Journals*, vol. 7, no. 2, pp. 100–106, 2023, ISSN: 2456-8880.

[15] C. Singh, M. K. Dash, R. Sahu, and A. Kumar, "Artificial intelligence in customer retention: A bibliometric analysis and future research framework," *Kybernetes*, 2023.

[16] A. Khaliq, S. Ajaz, A. Ali, D. Shakir, and K. Baig, "From data to decisions: Predictive machine learning models for customer retention in banking," *The Asian Bulletin of Big Data Management*, vol. 4, no. 3, pp. 74–85, 2024. DOI: 10.62019/abbdm.v4i3.206.