University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

Machine Learning Framework for Customer Churn Prediction in E-Commerce

RAHUL YADAV

Student Registration ID: 2321406

PRID: YADAV20005

Email: ry23416@essex.ac.uk

Supervisor: **Dr. Gao Tao**

January 6, 2025
Colchester

Contents

List of Figures

List of Tables

Abstract

Fast e-commerce growth has brought in opportunities for convenience and accessibility, but it has also increased competition and the costs of customer acquisition. Retaining existing customers not only is cheaper than acquiring new ones but is also a sustainable way to look at growth and hence makes studying customer behavior important to predict and reduce churn. While traditional churn analysis has mainly relied on transactional data, this study takes a broader approach by considering behavioral, demographic, and engagement features in developing comprehensive predictive models.

The research utilizes different machine learning techniques in predicting customer churn, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, and XGBoost. Of these, the most effective were Random Forest and XGBoost, with both scoring 99.5% and 99.4% accuracy, respectively, and corresponding ROC-AUC values of 0.9999 and 0.9975. These models were able to capture complex patterns in customer behavior, showing a significant improvement over simpler models like SVM, which scored an accuracy of 78.1% and a ROC-AUC of 0.8536.

These results underline the strong effectiveness of ensemble-based methods like Random Forest and XGBoost for churn prediction and offer actionable insights into the factors driving customer retention. The present study will add to the literature by equipping e-commerce businesses with tools to perform focused retention strategies that reduce the rate of churn and foster long-term loyalty among customers.

List of Abbreviations

- **AUC** - Area Under the Curve
- **ROC** - Receiver Operating Characteristic
- **ML** - Machine Learning
- **SVM** - Support Vector Machine
- **EDA** - Exploratory Data Analysis
- **CR** - Churn Ratio
- **LI** - Loyalty Index
- **AES** - App Engagement Score
- **CE** - Cashback Effectiveness
- **SMOTE** - Synthetic Minority Oversampling Technique
- **TP** - True Positives
- **TN** - True Negatives
- **FP** - False Positives
- **FN** - False Negatives
- **TPR** - True Positive Rate
- **FPR** - False Positive Rate
- **GPU** - Graphics Processing Unit
- **TPU** - Tensor Processing Unit

CHAPTER 1

Introduction

The e-commerce industry has experienced unprecedented growth in recent years, revolutionizing the way consumers shop and engage with businesses. The proliferation of smartphones and the widespread availability of high-speed internet have made online shopping more accessible and convenient than ever before. In fact, global e-commerce sales surpassed a staggering $5.7 trillion in 2023, reflecting a robust shift in consumer behavior [1]. Market analysts predict that this upward trend will continue, with sales expected to rise further in the coming years as technological advancements and changing consumer habits propel the industry forward [2].

However, this rapid expansion has intensified competition among e-commerce platforms. To maintain their customer base, businesses must invest considerable effort and resources into retaining existing customers, as acquiring new ones can be significantly more expensive. Research indicates that the cost of gaining new customers can be 5 to 7 times higher than the expenses associated with retaining loyal ones [3]. This underscores the importance of customer retention strategies, which are not only more economical in the long run but also contribute to enhanced profitability. Loyal customers tend to spend more over time and often act as brand advocates, spreading positive word-of-mouth and generating additional revenue through referrals and recommendations [4].

Despite their efforts, many e-commerce platforms face challenges in retaining customers. Customers often leave due to dissatisfaction with service, better deals offered by com-

petitors, or changes in their personal preferences. For instance, during the COVID-19 pandemic, many customers shifted toward platforms offering superior delivery services and better inventory management. Smaller e-commerce platforms struggled to compete, resulting in a noticeable migration of customers to larger, well-established platforms like Amazon, which consistently offered faster shipping times and a broader product selection [5]. This trend underscores the critical need for businesses to adapt to changing market dynamics and customer expectations to remain competitive.

1.1 Understanding Customer Churn

Customer churn is a critical metric representing the number of customers who discontinue their purchasing relationship with a business over a defined period. In the e-commerce context, understanding the factors contributing to churn is essential for designing effective retention strategies.

One significant cause of customer churn is related to product quality and service reliability. Customers receiving subpar products or experiencing delays in delivery are likely to become frustrated and seek alternatives [6]. Problems such as faulty items, stock unavailability, or inconsistent service can erode customer confidence and loyalty.Pricing also plays a pivotal role in customer retention. In the fast-paced e-commerce landscape, customers are constantly seeking the best deals. Competitors offering similar products at lower prices or attractive discounts can incentivize customers to switch platforms [7]. Maintaining competitive pricing is crucial to retaining the customer base.

Another contributing factor to churn is insufficient customer engagement. When customers feel undervalued or neglected, they are more likely to explore competitors. Effective engagement strategies, such as personalized communication, loyalty programs, and prompt responses to customer feedback, can foster stronger relationships between businesses and their customers [8].For instance, when customers cancel subscriptions to services like Netflix or Spotify, this is categorized as churn [9]. Similarly, in e-commerce, churn occurs when customers cease making purchases from a platform or shift their loyalty to competitors. Understanding these dynamics enables businesses to develop targeted strategies to minimize churn and improve retention.

1.2 Importance of Retaining Customers

Customer retention is a cornerstone of success for e-commerce businesses. Studies reveal that increasing customer retention rates by just 5% can lead to profit increases ranging from 25% to 85% [10]. This underscores the substantial financial impact of fostering customer loyalty.Loyal customers bring multiple benefits to a business. First, they are more likely to spend more frequently and in larger amounts as they become familiar with the brand. This increased spending enhances their lifetime value to the business [11].

Second, satisfied customers often recommend the brand to others, creating organic word-of-mouth marketing. This not only generates new customers but also does so without incurring additional marketing costs [12].Lastly, loyal customers provide valuable insights and suggestions that businesses can use to refine their products and services, ensuring alignment with customer needs and preferences [13].

A notable example of effective customer retention is Starbucks' loyalty program. Customers earn points with every purchase, redeemable for rewards, which incentivizes repeat visits and fosters a stronger sense of loyalty [1].Focusing on customer retention is not merely a strategy but a fundamental aspect of building a sustainable and thriving e-commerce business. Beyond financial gains, it creates a loyal community that significantly contributes to long-term success.

1.3 Reducing Churn in E-Commerce

Reducing customer churn is critical for long-term success in e-commerce, as acquiring new customers is considerably more expensive than retaining existing ones. Companies must identify and address the underlying reasons for churn while leveraging technology, such as machine learning (ML), to predict and mitigate it proactively [14].Personalized marketing is one of the most effective strategies to reduce churn. By analyzing browsing and purchase history, businesses can send targeted offers and recommendations, enhancing the customer experience. For example, Netflix uses sophisticated algorithms to suggest content based on users' viewing patterns, ensuring continued engagement

and loyalty [2].

Loyalty programs also play a vital role in retention. Programs such as Sephora's Beauty Insider allow customers to accumulate points redeemable for benefits, encouraging repeat purchases. These initiatives not only foster loyalty but also strengthen the emotional connection between customers and the brand [4].Additionally, providing enhanced customer support is essential for retention. Businesses must offer multiple channels for support, including live chat, email, and phone, to resolve issues efficiently. Prompt and effective service can turn potentially negative experiences into positive ones, increasing customer satisfaction and loyalty [7].

CHAPTER 2

Literature Review

Customer churn prediction in e-commerce has become a critical area of research due to its direct impact on business profitability and sustainability. With the exponential growth of e-commerce, businesses face increasing competition and seek innovative methods to retain customers while mitigating the high costs of acquiring new ones. Reichheld [3] highlights that a 5% increase in retention can lead to profit growth of 25% to 85%, emphasizing the importance of customer loyalty in sustainable business growth. Similarly, Reichheld and Sasser [10] stress the value of zero defections in improving service quality, a principle still relevant for modern e-commerce platforms. Statista [1] reports global e-commerce sales surpassing $5.7 trillion in 2023, driven by technological advancements and changing consumer habits, further underlining the importance of retention strategies. Personalized marketing strategies, such as tailored offers and targeted communication, have been shown to significantly enhance customer retention in the competitive digital commerce landscape [4].

Understanding customer churn is central to developing effective retention strategies. Wu and Meng [5] demonstrate the effectiveness of segmentation and AdaBoost for churn analysis, showing that advanced machine learning models outperform traditional statistical approaches in handling complex e-commerce data.Machine learning (ML) has revolutionized churn prediction, offering tools to analyze large datasets and uncover hidden patterns. While Logistic Regression remains a popular baseline for its simplicity and interpretability [10], it often struggles with non-linear interactions, necessitating

advanced methods. Ensemble methods like Random Forest and XGBoost excel in capturing complex relationships and handling imbalanced datasets [6]. Freund and Schapire's [7] introduction of boosting laid the groundwork for AdaBoost and XGBoost, which iteratively improve prediction accuracy. Support Vector Machines (SVMs) are also effective for churn prediction, although their scalability is limited by computational complexity and sensitivity to feature scaling [8]. Deep learning approaches, such as neural networks, have gained traction for their ability to model intricate patterns. Feng and team. [9] demonstrate significant improvements in prediction accuracy by combining neural networks with sentiment analysis.

Decision Trees are considered to be quite simple and interpretable, and therefore widely used in practice for churn prediction in e-commerce. By recursively partitioning the dataset depending on the feature values, Decision Trees present a comprehensible and intuitive illustration of the decision-making process. This kind of transparency is of high value to business stakeholders as it helps provide actionable insights about customer behavior. Decision Trees handle both numerical and categorical variables with ease, making them versatile over a wide range of data. However, they can be prone to overfitting, which ensemble methods like Random Forest can reduce. In terms of churn prediction, Decision Trees have been used mostly as baseline models that will identify major predictors of churn, such as transaction frequency or customer satisfaction scores. [15] This is especially useful in a churn prediction setting to understand why customers defect and work out corresponding churn prevention strategies. The split rules of the decision tree are optimized according to the EMPC metric using an evolutionary algorithm.

Logistic regression is also widely used-a benchmark, even-since it can estimate the probabilities of churn while maintaining high simplicity and interpretability. As expected, if the underlying relation of features with a target variable turns out to be pretty linear, the results obtained would be excellent. It works fine for getting insight into data that will compare bench complex models and be more interpretable. It explains quantitatively how every different feature level differently influences the chance of churn in an observation. It can, for example, identify from logistic regression whether variables such as tenure or satisfaction score significantly raise or lower the chances

of churn. Although it cannot model nonlinear relationships, Logistic Regression is computationally efficient and helpful in situations where quick, interpretable results are needed. It's the one machine-learning algorithm that is not a black box model. Normally black box models are complex but the logistic regression tells what it does actually. Logistic regression can be binary, nominal or ordinal . [16]

Feature engineering plays a pivotal role in churn prediction. Behavioral variables such as order frequency, transaction value, and satisfaction scores enhance model performance, as emphasized by Chen and team. [2]. Derived metrics like Loyalty Index and Churn Ratio provide additional insights into customer behavior [5]. Lu and team. [11] advocate for RFM (Recency, Frequency, Monetary) analysis to improve customer segmentation, while Kumar and team. [4] suggest combining RFM with machine learning models for greater granularity. Composite metrics like the Loyalty Index, incorporating tenure, satisfaction, and coupon usage, offer a comprehensive view of customer engagement and retention potential.

Addressing data imbalance is another critical challenge in churn prediction. Datasets often skew heavily toward non-churned customers, necessitating techniques to balance the classes. Dhote and team. [12] propose hybrid geometric sampling and AdaBoost-based deep learning for improved robustness. Oversampling methods like SMOTE generate synthetic samples for the minority class, further mitigating bias [6]. Agrawal and team. [14] integrate deep learning with behavioral pattern analysis to address imbalance while maintaining a balance between sensitivity and precision for actionable predictions.

External factors also influence churn dynamics. Chen and team. [2] analyze the impact of COVID-19 on e-commerce trends, revealing shifts in customer behavior and increased reliance on digital platforms. Huang [13] highlights the role of community-driven engagement in social e-commerce, stressing the importance of adapting prediction models to changing market conditions and external economic factors.Personalization and loyalty programs are effective in reducing churn. Smith and Johnson [4] advocate for personalized marketing strategies leveraging customer data to create tailored offers and recommendations, enhancing engagement. Wang and team. [8] emphasize loyalty programs effectiveness, highlighting incentives like cashback and reward points as

critical retention tools. Kumar and team. [4] discuss the role of digital technologies in fostering loyalty, demonstrating how machine learning and personalized engagement strategies can significantly improve retention rates.

The literature reviewed underlines the pivotal role of machine learning in tackling issues related to customer churn within e-commerce. Traditional analytical methods, such as Logistic Regression, lay a foundation for understanding customer behavior, but most of these methods have been developed and, therefore, failed to capture the intricacies of customer interaction. On the other hand, some modern methods using ensemble learning and deep neural networks show much better performance in churn prediction and hence provide much better insights for companies w.r.t. customer retention and attrition.

The review of different studies shows that despite multiple models being proposed for the problem of customer churn prediction, substantial gaps still exist, especially for real-time prediction models. Such models would have the ability to incorporate new data about customers every time and thus change their forecast. While traditional models, like Logistic Regression, offer simplicity and interpretability, they often fail to capture the complex nonlinear relationships that exist in customer behaviors. More sophisticated ensemble methods, including Random Forests and XGBoost, have been more capable of modeling these complex patterns, thus yielding more accurate predictions of customer churn. Furthermore, deep learning approaches, though promising, also raise issues related to their computational requirements, coupled with inherent difficulties in the interpretation of their results. These are likely to make most of the deep learning algorithms inapplicable in critical business settings where decision-makers require clear visibility into why specific strategies are necessary. This research study will attempt to fill these gaps by leveraging the strengths of XGBoost and other ensemble learning methods for better churn predictions, while maintaining adequate interpretability of the results. This focus is of particular relevance to e-commerce businesses, often working with large and complex datasets. The integration of sophisticated feature engineering with these advanced modeling techniques means the research can deliver more detailed and actionable insights into customer churn dynamics.

This, therefore, finally aims at equipping e-commerce organizations with the ability

to undertake retention strategies with precision. It would, in the end, enable firms to establish long-term relationships with customers by reducing churn rates and, therefore, improve growth and profitability in a fiercely competitive digital marketplace.

CHAPTER 3

About Dataset

In predictive modeling, the quality, structure, and preprocessing of the dataset are very important to ensure that the machine learning models are effective and reliable. This study is using a dataset from Kaggle Dataset Repository, which is known to provide good-quality datasets tailored for research and machine learning purposes. The dataset contains customer information about demographics, behavioral interactions, and transactional information from an e-commerce platform. These features are instrumental in identifying important factors contributing to customer churn and in building robust predictive models.

Figure 3.1 illustrates the methodical conversion of unrefined data into a polished, organized, and model-ready form through the data preparation procedure. This comprehensive approach, covering data cleansing, feature development, and handling of missing values, ensures that the dataset is refined for better performance and accuracy in predictive analytics. Using this systematic framework, the research secures the insights on customer behavior and strategies for retention

3.1 Data Introduction

The table 3.1 highlights the dataset's structure, data completeness, and feature types, which are essential for preprocessing and modeling steps.This table provides an overview of the dataset used for analysis. It contains:

Table 3.1: Dataset Description

#	Column	Non-Null Count	Data Type
1	CustomerID	5630 non-null	int64
2	Churn	5630 non-null	int64
3	Tenure	5366 non-null	float64
4	PreferredLoginDevice	5630 non-null	object
5	CityTier	5630 non-null	int64
6	WarehouseToHome	5379 non-null	float64
7	PreferredPaymentMode	5630 non-null	object
8	Gender	5630 non-null	object
9	HourSpendOnApp	5375 non-null	float64
10	NumberOfDeviceRegistered	5630 non-null	int64
11	PreferedOrderCat	5630 non-null	object
12	SatisfactionScore	5630 non-null	int64
13	MaritalStatus	5630 non-null	object
14	NumberOfAddress	5630 non-null	int64
15	Complain	5630 non-null	int64
16	OrderAmountHikeFromLastYear	5365 non-null	float64
17	CouponUsed	5374 non-null	float64
18	OrderCount	5372 non-null	float64
19	DaySinceLastOrder	5323 non-null	float64
20	CashbackAmount	5630 non-null	float64

- **Column Names:** Lists the 20 features in the dataset.

- **Non-Null Count:** Indicates the number of non-missing values for each feature. Most columns have complete data (5630 rows), while some have missing values (e.g., *Tenure* and *DaySinceLastOrder*).

- **Data Types:** Specifies the type of data for each column, such as `int64` for integers, `float64` for decimal numbers, and `object` for categorical data.

3.2 Dataset Overview

The dataset you provided appears to contain a variety of features related to customer behavior in an e-commerce setting, with a total of 20 columns. These columns include demographic information, customer activities, and transaction-related metrics. Here's a breakdown of the dataset:

- **CustomerID:** A unique identifier for each customer.
- **Churn:** A binary flag indicating whether the customer has churned or not.
- **Tenure:** The duration of the customer relationship in months.
- **PreferredLoginDevice:** The device that the customer most frequently uses to log in.
- **CityTier:** The tier of the customer's city (example, 1 for major cities, 2 for smaller cities).
- **WarehouseToHome:** The distance between the warehouse and the customer's home.
- **PreferredPaymentMode:** The most frequently used payment method by the customer.
- **Gender:** The gender of the customer.
- **HourSpendOnApp:** The number of hours the customer spends on the e-commerce app.
- **NumberOfDeviceRegistered:** The total number of devices linked to the customer's account.
- **PreferedOrderCat:** The category of products the customer orders most frequently.
- **SatisfactionScore:** The customer's satisfaction rating.
- **MaritalStatus:** The marital status of the customer.

- **NumberOfAddress:** The number of addresses the customer has saved in the system.
- **Complain:** A binary flag indicating whether the customer has made a complaint.
- **OrderAmountHikeFromlastYear:** Percentage increase in order amount compared to last year.
- **CouponUsed:** The number of coupons the customer used in the last month.
- **OrderCount:** The total number of orders the customer has placed in the last month.
- **DaySinceLastOrder:** The number of days since the customer's last order.
- **CashbackAmount:** The average cashback amount the customer received in the last month.

3.3 Data Summary

The dataset used for this analysis represents customer behavior and engagement metrics in an e commerce platform. It includes 20 attributes and 5630 entries, covering both numerical and categorical features. The primary goal of this dataset is to understand customer churn whether a customer discontinues their engagement with the platform.

3.4 Summary Statistics of Dataset Attributes

In table 3.2 we removed **MaritalStatus** and **PreferedOrderCat** from the table because these columns lack meaningful numerical data for summary statistics such as mean, standard deviation, minimum, and maximum values. Both columns are categorical, with their values being nominal categories (e.g., "Married", "Single" for MaritalStatus or product categories for PreferedOrderCat). Numerical summary statistics (like mean and standard deviation) do not apply to such data types. Including these columns in a table summarizing numerical attributes does not provide meaningful insights. Removing these columns improves the clarity of the table by focusing only on numerical attributes, making it more concise and easier to interpret.

Table 3.2: Summary Statistics of Dataset Attributes

Statistic	CustomerID	Churn	Tenure	CityTier	WarehouseToHome	HourSpendOnApp
Count	5630.0	5630.0	5366.0	5630.0	5379.0	5375.0
Mean	52815.5	0.168	10.19	1.65	15.64	2.93
Std Dev	1625.39	0.37	8.56	0.92	8.53	0.72
Min	50001.0	0.0	0.0	1.0	5.0	0.0
25%	51408.25	0.0	2.0	1.0	9.0	2.0
Median	52815.5	0.0	9.0	1.0	14.0	3.0
75%	54222.75	0.0	16.0	3.0	20.0	3.0
Max	55630.0	1.0	61.0	3.0	127.0	5.0

Statistic	NumberOfDeviceRegistered	SatisfactionScore	NumberOfAddress	Complain
Count	5630.0	5630.0	5630.0	5630.0
Mean	3.68	3.07	4.21	0.28
Std Dev	1.02	1.38	2.58	0.45
Min	1.0	1.0	1.0	0.0
25%	3.0	2.0	2.0	0.0
Median	4.0	3.0	3.0	0.0
75%	4.0	4.0	6.0	1.0
Max	6.0	5.0	22.0	1.0

Statistic	OrderAmountHikeFromLastYear	CouponUsed	OrderCount	DaySinceLastOrder	CashbackAmount
Count	5365.0	5374.0	5372.0	5323.0	5630.0
Mean	15.71	1.75	3.01	4.54	177.22
Std Dev	3.67	1.89	2.94	3.65	49.21
Min	11.0	0.0	1.0	0.0	0.0
25%	13.0	1.0	1.0	2.0	145.77
Median	15.0	1.0	2.0	3.0	163.28
75%	18.0	2.0	3.0	7.0	196.39
Max	26.0	16.0	16.0	46.0	324.99

3.5 Dataset Insights

The dataset was systematically analyzed to understand its structure, completeness, and unique characteristics. This steps ensures data readiness for prepossessing and modeling.Below Figure 3.1 is the workflow of the Data.

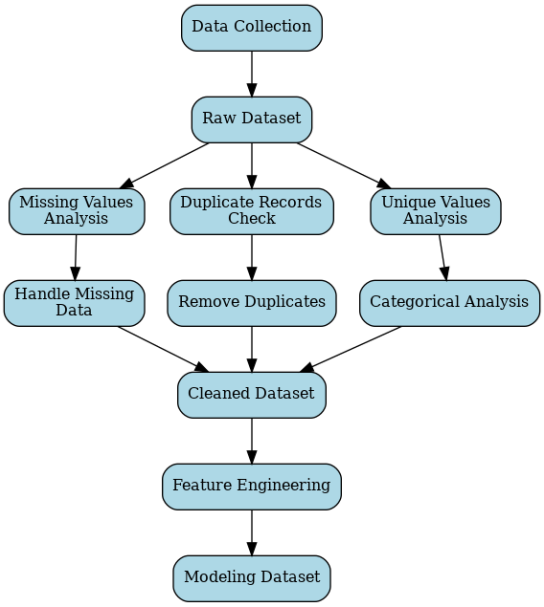

Figure 3.1: Data Description Workflow

3.5.1 Missing Values Analysis

Several columns were found to contain missing values, necessitating imputation strategies to maintain dataset integrity. Key columns with missing data include:

- **Tenure:** 264 missing values.
- **WarehouseToHome:** 251 missing values.
- **HourSpendOnApp:** 255 missing values.
- **DaySinceLastOrder:** 307 missing values.

Proper handling of these gaps is crucial to avoid biases in the analysis. Techniques such as mean or median replacement or more sophisticated methods like K-Nearest Neighbors (KNN) imputation can be applied, depending on the distribution of each variable.

3.5.2 Duplicate Records

The dataset was found to be free of duplicate rows, confirming its uniqueness and eliminating the need for deduplication.

3.5.3 Unique Value Analysis

The dataset utilized for this e-commerce churn analysis comprises various features that provide comprehensive insights into customer behavior and patterns. Key attributes include **CustomerID**, a unique identifier for each customer with 5,630 unique values, ensuring no duplication in records. The target variable, **Churn**, is binary, representing two classes: 0 for non-churned customers and 1 for churned customers, facilitating the predictive analysis of customer retention.The dataset also features **Tenure**, with 36 unique values that reflect the duration of customer relationships, enabling the identification of patterns related to long-term engagement. **CityTier**, with three distinct categories, classifies customers based on their urban location, providing geographical context to the analysis.

Additionally, **PreferredLoginDevice** outlines customer preferences for accessing the platform, encompassing categories such as *Mobile Phone*, *Phone*, and *Computer*, which offer insights into technology usage trends. The **PreferredPaymentMode** feature highlights seven distinct payment behaviors, including *Debit Card*, *Credit Card*, and *Cash on Delivery*, providing valuable information on customer payment preferences.By analyzing these attributes, the study uncovers critical patterns and behaviors, forming a solid foundation for building predictive models and designing targeted retention strategies.

3.5.4 Unique Values in Categorical Columns

Analyzing unique values in categorical columns is essential for understanding, cleaning, and utilizing the data effectively. This process ensures the dataset is prepared for predictive modeling and enables actionable insights for retention strategies, personalization, and enhancing the customer experience. The categorical variables in the dataset provide vital context for segmentation and analysis.

The column PreferredLoginDevice highlights customers preferred devices for accessing the service, with categories such as Mobile Phone, Phone, and Computer. This provides insights into technology usage trends. PreferredPayingMode reflects diverse payment behaviors, with categories including Debit Card, UPI, Cash on Delivery, E Wallet, and Credit Card. The Gender column enables demographic segmentation with two

categories: Male and Female. The PreferredOrderCat column captures customers dominant purchase preferences, including categories such as Mobile, Laptop ,Accessory and Grocery. Lastly, MaritalStatus provides demographic profiling opportunities with categories like Single, Married, and Divorced.

These insights derived from the categorical columns form the foundation for effective customer segmentation, enabling predictive modeling and supporting retention strategies tailored to customer preferences.

CHAPTER 4

Methodology

This Research was conducted using Google Colab, a cloud-based platform that offers an accessible, collaborative, and efficient environment for developing and executing Python code. Google Colab is particularly well-suited for data science and machine learning projects due to its integration with powerful computing resources, including free GPU and TPU support, which significantly accelerates model training and evaluation.

4.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an essential ingredient in any data-driven project and even more so in e-commerce churn analysis. EDA visually and statistically examines the dataset to identify critical characteristics, detect anomalies, and uncover relationships among variables. This step is crucial to ensure the data is well-prepared for modeling, and key insights can be put to good use.

EDA begins by assessing the completeness of the dataset, such as detecting missing values and understanding their potential effects on analyses. One common problem in many real-world datasets is missing data, and addressing it properly ensures the validity of subsequent analyses. In this study, features such as *Tenure*, *WarehouseToHome*, and *DaySinceLastOrder* were examined for missing values, and applicable imputation methods were systematically applied.Additionally, the distribution of the target variable, *Churn*, was closely examined to address potential imbalances that could affect predictive

modeling. The dataset exhibited a notable class imbalance, with 83% of customers categorized as non-churned. Addressing this imbalance was a primary focus of the EDA to ensure the robustness and fairness of the model's performance.

Bar charts, histograms, and pie charts were extensively utilized to develop intuition about customer behavior. For example, analyzing the distribution of payment methods, login devices, and satisfaction scores provided valuable insights into customer preferences and engagement levels. Heatmaps were employed to check pairwise correlations among numerical variables, revealing highly related features. For instance, *DaySinceLastOrder* and *Churn* exhibited an inverse relationship. These visual analyses assisted in identifying numerous potential predictors of churn.Furthermore, the distribution and relationship of categorical variables with *Churn* were analyzed. Features such as *PreferredPaymentMode* and *CityTier* were examined for patterns that could reveal customer tendencies toward churning. This analytical approach enhanced the understanding of customer behavior and informed feature engineering and variable selection in the context of predictive modeling.

In summary, the EDA conducted during this project transformed raw data into actionable insights. It provided a concrete foundation for data preprocessing, feature engineering, and ultimately the development of predictive models capable of accurately assessing churn risk and informing retention strategies.

4.2 Churned versus retained customers

Churn	Frequency	Percentage (%)
0 (Non-Churned)	4682	83.16
1 (Churned)	948	16.83

Table 4.1: Distribution of Churn Variable in the Dataset

The analysis of the Churn variable provided essential insights into the class distribution within the dataset.As shown in the table 4.1 Calculating the churn percentages revealed a notable imbalance, where the majority of customers (83.16%) are non-churned, while only 16.83% represent churned customers. This class imbalance is critical and poses

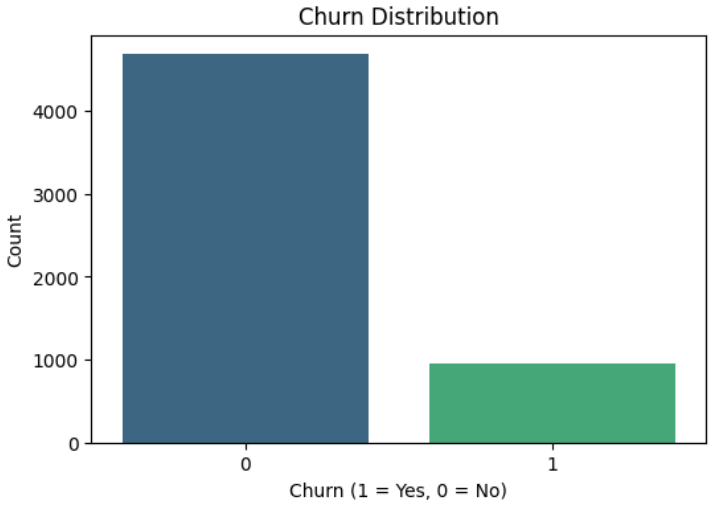

Figure 4.1: Churn Distribution

challenges for predictive modeling, as machine learning models might tend to favor the majority class (non-churned) over the minority class (churned). This table 4.1 highlights the distribution of churn, emphasizing the need for careful handling during the data preprocessing and modeling phases to mitigate bias and achieve better classification results.

4.3 Insights from the Distribution of Features with Churn

The insights from the feature-wise churn analysis as shown in figure 4.2 reveal specific patterns that highlight opportunities for reducing churn and enhancing customer retention. These insights emphasize practical strategies for improving customer satisfaction and loyalty.One key takeaway is the need to enhance customer engagement through personalized experiences. Customers who spend more time on the app or use multiple devices to access the platform tend to stay loyal. This suggests that by offering features like personalized product recommendations, interactive interfaces, and tailored communication, businesses can make customers feel more valued and engaged. Initiatives such as curated shopping suggestions or regular updates about products customers love can encourage longer app usage and reduce the likelihood of churn.

Another area requiring attention is service improvement in Tier 2 and Tier 3 cities, where higher churn rates were observed. Customers in these regions often face issues like

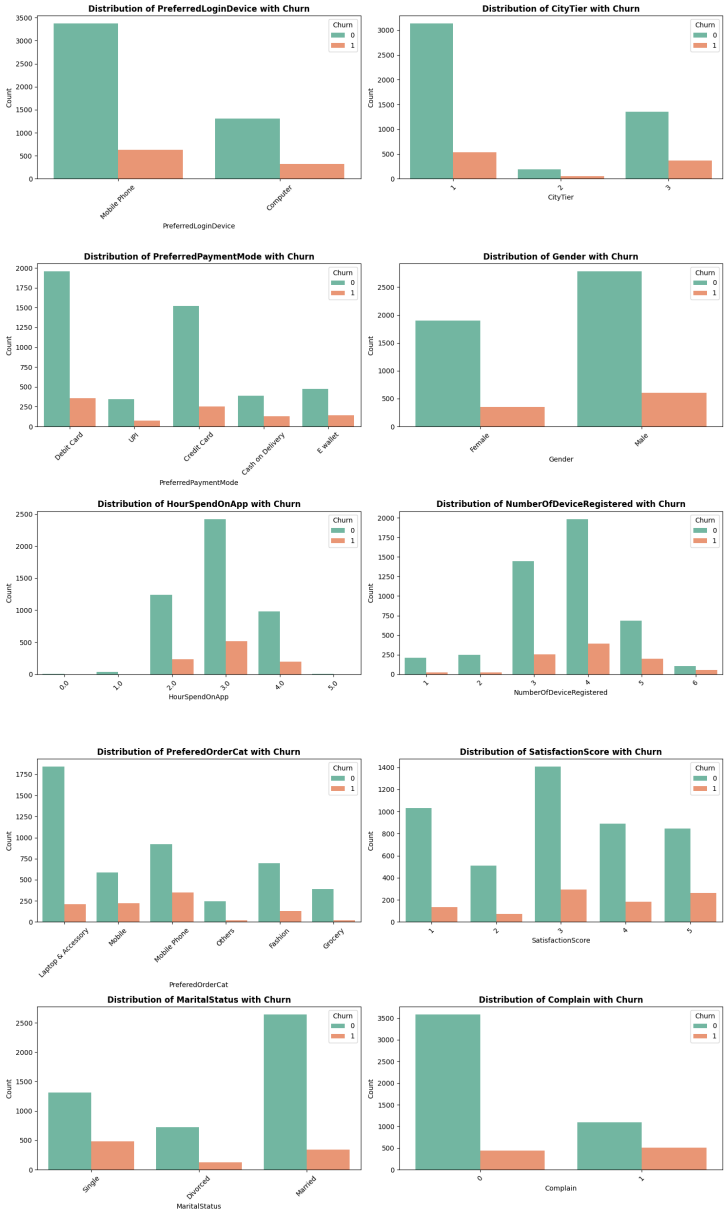

Figure 4.2: Distribution of Features with Churn

delayed deliveries or limited product options. To address this, businesses could invest in better logistics infrastructure, such as opening warehouses closer to these areas or optimizing delivery routes. Offering special promotions or loyalty rewards to these customers can further make them feel prioritized and appreciated.Payment methods also play a significant role in customer loyalty. The data indicates that customers using digital payment methods like credit or debit cards are less likely to churn compared to those who rely on cash-on-delivery (COD). Encouraging the use of digital payments through discounts, cashback, or reward points can reduce churn while streamlining the payment process for both the customer and the company.

Lastly, the way customer complaints are handled significantly impacts retention. Customers who feel their complaints are not resolved effectively are much more likely to churn. To tackle this, businesses should focus on building a responsive and efficient customer support system. This could include real-time chat support, 24/7 helplines, and proactive follow-ups to ensure every issue is addressed. Quick and empathetic resolutions can transform a negative experience into a positive one, fostering trust and loyalty.

By focusing on these strategies, businesses can create a more customer-centric approach that not only reduces churn but also builds lasting relationships with their customers, ensuring long-term success and growth.

4.4 Correlation Matrix Analysis

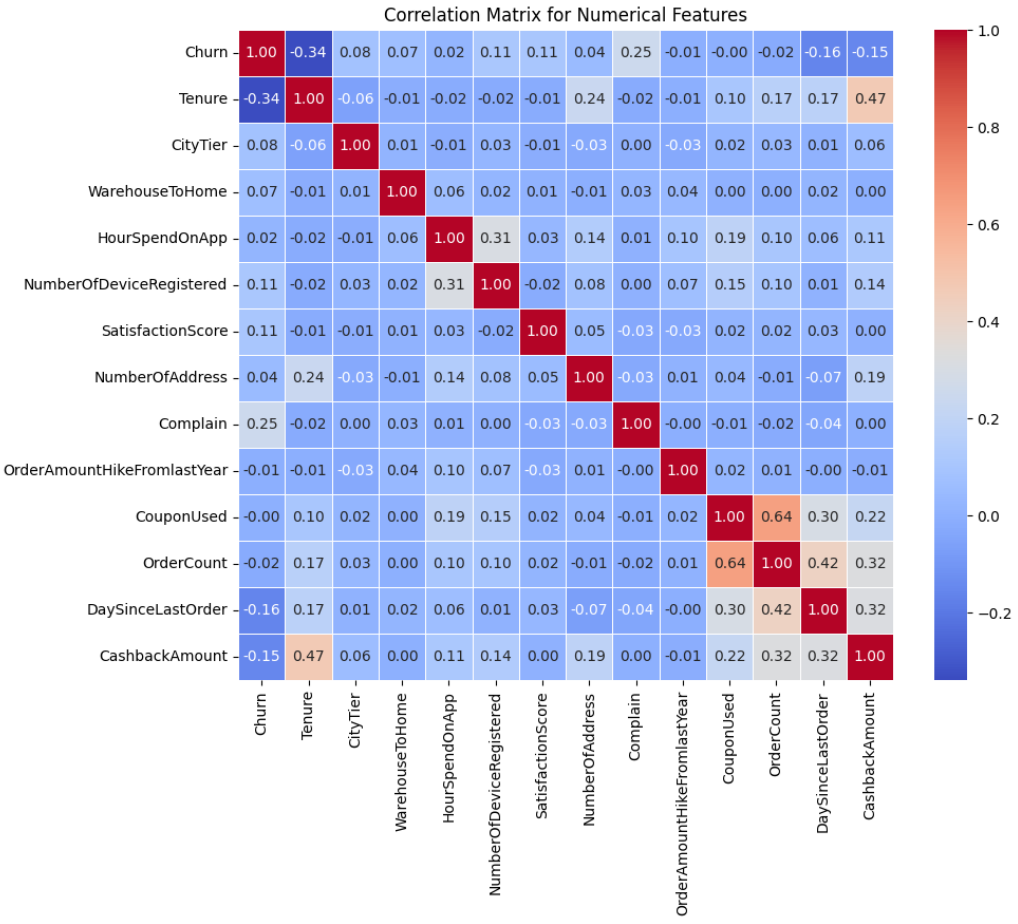

Figure 4.3: Correlation Matrix for Numerical Features

The correlation matrix provides valuable insights into the relationships between numerical features and customer churn, forming a crucial foundation for identifying churn drivers and guiding feature selection in predictive modeling. Among the key findings,as shown in figure 4.3 the negative correlation of -0.34 between *Tenure* and *Churn* indicates that customers with longer durations of association with the platform are less likely to churn. This emphasizes the importance of fostering loyalty over time, suggesting that implementing loyalty programs and targeted retention strategies can be effective in reducing churn.Similarly, a moderate negative correlation of -0.25 between *Complaint* and *Churn* reveals that customers who register complaints are more prone to leave the platform. This finding highlights the need to improve customer service systems and address grievances efficiently, such as through real-time complaint resolution and proactive follow-ups.

The analysis also shows a correlation of -0.16 between *DaySinceLastOrder* and *Churn*, suggesting that customers who have recently made purchases are less likely to churn. This underscores the importance of maintaining regular engagement with customers through personalized offers, reminders, and re-engagement campaigns to encourage consistent interaction with the platform.Inter-feature relationships further reveal critical patterns. A positive correlation of 0.47 between *CashbackAmount* and *Tenure* suggests that cashback incentives effectively encourage long-term customer retention. This validates the role of cashback programs as a critical retention strategy. Additionally, the strong positive correlation of 0.64 between *OrderCount* and *CouponUsed* indicates that customers who frequently use coupons tend to place more orders, reinforcing the effectiveness of promotional campaigns in driving customer activity and retaining interest in the platform. Similarly, a positive relationship between *SatisfactionScore* and *HourSpendOnApp* indicates that satisfied customers spend more time on the app, highlighting the need to enhance user experience and satisfaction through personalization and interactive features.

While some features, such as *WarehouseToHome* and *NumberOfDeviceRegistered*, exhibit weaker correlations with churn, their potential indirect effects should not be overlooked. For example, the distance from the warehouse could influence delivery times, which may indirectly affect customer satisfaction and churn. These weaker correlations may

hold value when combined with other features, providing additional context for churn prediction.The findings from this correlation analysis also inform feature engineering for the project. Variables such as *Tenure*, *Complaint*, and *CashbackAmount* should be prioritized in churn prediction models due to their significant correlations. Additionally, relationships like those between *OrderCount* and *CouponUsed* could be captured through derived interaction features to improve model accuracy. Exploring indirect effects of weaker correlations, such as those involving *WarehouseToHome*, could further enhance the predictive power of the model.By leveraging these insights, this analysis not only facilitates robust feature engineering but also guides the development of actionable strategies to improve customer retention and reduce churn effectively.

4.5 Derived Metrics for Customer Churn Analysis

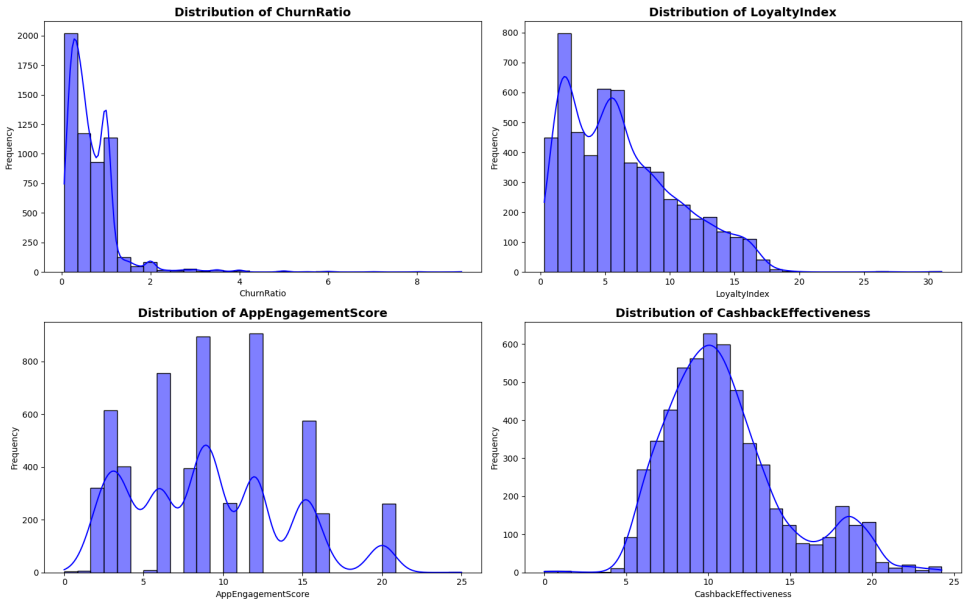

Figure 4.4: Distributions of Derived Metrics for Customer Churn Analysis.

In the analysis of customer churn, four critical derived metrics were meticulously designed and computed to significantly enhance the predictive modeling process.As shown in Figure 4.4 These metrics were formulated by leveraging and combining existing features in the dataset, enabling a deeper and more nuanced understanding of customer behavior and retention dynamics. By transforming raw data into actionable insights, these metrics offer an enriched perspective, facilitating the identification

of key churn drivers and the development of targeted retention strategies. Below is a comprehensive explanation of each metric, including its purpose, mathematical formulation, and the observed distributions from the dataset:

4.5.1 Derived Metrics

1. **Churn Ratio (CR)**:

   - The Churn Ratio captures customer activity by measuring the frequency of orders relative to the time elapsed since the last order. A higher Churn Ratio indicates consistent purchasing behavior and lower churn likelihood. This metric helps identify customers at risk based on declining activity levels.
   - **Equation:**

$$CR = \frac{\text{OrderCount}}{\text{DaySinceLastOrder} + 1}$$

2. **Loyalty Index (LI)**:

   - The Loyalty Index combines three key factors: the duration of association with the platform (Tenure), customer satisfaction levels, and coupon usage. This composite metric offers a holistic view of customer loyalty. A higher Loyalty Index reflects stronger customer retention potential.
   - **Equation:**

$$LI = 0.5 \cdot \text{Tenure} + 0.3 \cdot \text{SatisfactionScore} + 0.2 \cdot \text{CouponUsed}$$

3. **App Engagement Score (AES)**:

   - This metric evaluates customer engagement by correlating the time spent on the app with their satisfaction level. A higher App Engagement Score implies active and satisfied customers, reducing the probability of churn. It helps assess the effectiveness of the platform in retaining user attention.
   - **Equation:**

$$AES = \text{HourSpendOnApp} \cdot \text{SatisfactionScore}$$

4. **Cashback Effectiveness (CE)**:

- This metric assesses the impact of cashback incentives on customer retention. A higher Cashback Effectiveness score indicates that cashback offers effectively incentivize purchases, contributing to customer retention. It helps measure the return on investment for cashback programs.
- **Equation:**

$$CE = \frac{\text{CashbackAmount}}{\text{OrderAmountHikeFromLastYear} + 1}$$

4.5.2 Observed Values

ID	ChurnRatio	LoyaltyIndex	AppEngagementScore (	CashbackEffectiveness
1	0.1667	2.8	6.0	13.33
2	1.0000	5.4	9.0	7.56
3	0.2500	5.4	8.0	8.02
4	0.2500	5.0	5.58	5.58
5	0.2500	1.7	15.0	10.80

Table 4.2: Derived Metrics and Sample Observed Values

The derived metrics in the e-commerce customer churn prediction project provide a nuanced understanding of customer behavior, enabling data-driven decision-making to improve retention strategies.

The **Churn Ratio** identifies customers whose purchasing activity has declined, serving as an early warning system for potential churn. This metric enables businesses to proactively implement targeted re-engagement strategies, such as personalized offers or reminders, to retain these customers.

The **Loyalty Index** consolidates key factors such as tenure, satisfaction, and coupon usage into a single score, offering a comprehensive view of customer loyalty. High Loyalty Index scores can guide the implementation of tiered rewards or exclusive benefits, while lower scores signal the need for tailored retention initiatives, such as additional incentives or improved service experiences.

The **App Engagement Score** highlights the effectiveness of the platform in engaging customers. A high score indicates strong customer interaction and satisfaction, em-

phasizing the importance of maintaining an intuitive, feature-rich app experience. For customers with lower scores, businesses can introduce app-specific incentives, gamified features, or push notifications to enhance engagement.

The **Cashback Effectiveness** metric evaluates the impact of cashback programs on customer retention and activity. High scores validate the success of such programs and identify customers who respond well to cashback incentives. Conversely, low scores suggest the need for alternative promotional strategies to optimize resource allocation.

Integrating these metrics into the predictive modeling process enhances churn prediction accuracy and informs actionable strategies. They enable a deeper understanding of the drivers behind customer behavior, facilitating targeted interventions to foster loyalty, improve satisfaction, and ultimately reduce churn. These metrics not only refine the retention framework but also support the platform's long-term profitability and customer-centric growth.

4.5.3 Outcome of Derived Metrics

The derived metrics have significantly enhanced the predictive modeling process by providing deeper insights into customer behavior, enabling more targeted and effective strategies for customer retention. These metrics were developed to capture subtle, yet impactful, patterns in customer data that standard features might overlook. By integrating these insights into machine learning models, the project achieved a comprehensive understanding of the factors contributing to customer churn, allowing for precise identification of at-risk customers. The **Churn Ratio** allowed us to detect early signs of disengagement by quantifying purchase activity relative to time since the last order. This metric provided a straightforward yet powerful way to highlight customers needing immediate attention, enabling proactive retention efforts. The **Loyalty Index** consolidated key indicators like tenure, satisfaction scores, and coupon usage to offer a holistic measure of customer loyalty. This not only guided strategic decisions on rewards and offers but also prioritized high-value customers for personalized engagement. The **App Engagement Score** underscored the importance of customer interaction with the platform, revealing strong correlations between app usage and customer satisfaction. This metric validated investments in platform enhancements, such as intuitive

navigation and personalized content, which in turn strengthened customer engagement and loyalty. The **Cashback Effectiveness** metric quantified the return on investment for cashback initiatives. By highlighting which cashback programs resonated most with customers, this metric ensured resources were allocated effectively, maximizing both customer satisfaction and business profitability. When integrated into the machine learning models, these metrics improved the accuracy and interpretability of predictions. They provided actionable insights, translating complex customer data into meaningful strategies that addressed churn at its roots. Beyond statistical improvement, the practical application of these insights has led to a more customer-centric approach, fostering trust and loyalty among the user base.

In summary, these derived metrics transformed the project by bridging the gap between analytical rigor and practical utility. They empowered the predictive model to not only identify churn risks more effectively but also enabled the creation of personalized, impactful interventions. This balance of data-driven insights and actionable strategies has proven instrumental in driving long-term customer retention and profitability for the e-commerce platform.

4.6 Data Preprocessing

Before building the models, several preprocessing steps were carried out to ensure optimal performance:

Data Splitting The dataset was split into training and testing sets to facilitate model training and evaluation. The training set comprised 75% of the total dataset, resulting in 2,830 rows, while the testing set contained 25% of the data, amounting to 944 rows.

Feature Engineering Feature engineering was performed to refine the predictors for the models. The `CustomerID` column, which did not provide predictive value, was removed. Categorical predictors were encoded into numerical representations to make them compatible with machine learning algorithms. Additionally, predictors with zero or low variability were excluded from the dataset to avoid redundancy and improve model efficiency.

Handling Class Imbalance To address the class imbalance in the target variable `Churn`, an Oversampling Technique was applied. This technique generated synthetic samples for the minority class using the nearest neighbors, ensuring a balanced distribution of classes in the training dataset. This preprocessing step helped improve the robustness and fairness of the predictive models.

4.7 Oversampling

In the context of this project, oversampling was employed to address the imbalance in the target variable, `Churn`, within the dataset. Imbalanced datasets, where one class significantly outnumbers another, pose a challenge for machine learning models. Models trained on imbalanced data tend to favor the majority class, leading to poor predictive performance for the minority class, which in this case represents churned customers (`Churn = 1`). This imbalance skews the decision boundaries and reduces the model's ability to generalize effectively, making it critical to balance the classes before training.

The dataset used for this project had a significant class imbalance, with the majority class (non-churned customers, `Churn = 0`) comprising 4,682 records, while the minority class (churned customers, `Churn = 1`) had only 948 records. This imbalance made it imperative to use oversampling to prevent the model from being biased toward predicting non-churned customers and to ensure meaningful insights into customer churn behavior.

4.7.1 Steps Taken for Oversampling

1. **Label Encoding:** Since machine learning models require numerical input, categorical columns were converted into numeric representations using `LabelEncoder`. For instance, categories like `Male` and `Female` were encoded as 0 and 1, respectively. This step ensured that all data was in a compatible format for further processing.

2. **Separating Classes:** The dataset was divided into two subsets: one for the majority class (`Churn = 0`) and another for the minority class (`Churn = 1`). This

segregation was necessary to apply oversampling exclusively to the minority class.

3. **Oversampling the Minority Class:** The minority class was oversampled by randomly duplicating its rows until its size matched that of the majority class. The use of the `replace=True` parameter allowed rows to be duplicated, thereby increasing the size of the minority class to 4,682 records.

4. **Combining Classes:** After oversampling, the augmented minority class was combined with the original majority class to create a balanced dataset. This new dataset contained an equal number of records for both `Churn = 0` and `Churn = 1`.

5. **Shuffling:** The combined dataset was shuffled to ensure random distribution of the oversampled rows throughout the dataset. This step eliminated any order bias that could potentially affect the model's learning process.

6. **Verification:** The class distribution in the balanced dataset was confirmed, showing equal representation of 4,682 records for both `Churn = 0` and `Churn = 1`. This validated the success of the oversampling process.

Oversampling effectively addressed the class imbalance in the dataset, ensuring an equal representation of both churned (`Churn = 1`) and non-churned (`Churn = 0`) customers. This preprocessing step eliminated the bias toward the majority class, allowing machine learning models to learn patterns and relationships from the minority class more effectively. By balancing the dataset, the models were better equipped to accurately predict churn, resulting in improved classification performance and enhanced reliability.

Before oversampling, the dataset had a significant class imbalance, with 2,357 retained customers and only 473 churned customers. After applying oversampling, the training data was balanced, with 2,357 samples in each class. This ensured that machine learning models were not biased toward the majority class and could effectively learn patterns associated with churn.

CHAPTER 5

Models Implementation

The model-building phase begins by splitting the dataset into training and testing subsets, followed by the development of five machine learning models to predict customer churn: Decision Tree, Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost.

Logistic Regression

Logistic Regression is a reliable baseline model for binary classification tasks like churn prediction, where the target variable has two classes: churned and non-churned. Its simplicity and efficiency make it a valuable starting point for understanding the relationship between predictors and churn. By normalizing predictors and removing highly correlated variables, the logistic regression model avoids redundancy and improves performance. This model is particularly effective for identifying linear relationships, making it useful for initial insights into the key drivers of churn.

The logistic regression model was developed by David R. Cox in 1958. He introduced the method as a way to model binary outcomes using a logistic function. However, the logistic function itself (the sigmoid curve) has a much older history and was first used by Pierre François Verhulst in the 19th century to model population growth.
David Cox is often credited for formalizing the logistic regression as a statistical method for classification, and it has since become a foundational technique in fields like machine learning, data science, and medical statistics. [17]

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \tag{5.1}$$

- $P(Y = 1|X)$: Probability of the customer churning ($Y = 1$).
- e: Euler's number (base of the natural logarithm).
- β_0: Intercept term.
- $\beta_1, \beta_2, \ldots, \beta_n$: Coefficients for each predictor $X_1, X_2, \ldots, X_n$.

Random Forest

Random Forest is an ensemble method that overcomes the limitations of Decision Trees by combining predictions from multiple trees. It excels at handling imbalanced datasets, such as the one in this project before oversampling, by using class weighting or sampling techniques. Random Forest's robustness and ability to model complex non-linear relationships make it highly effective in capturing subtle patterns in customer behavior, such as the interplay between tenure, satisfaction scores, and app engagement. Additionally, its feature importance scores provide valuable insights into the relative influence of each predictor on churn, which is crucial for business strategy.

While the majority voting technique is a simple yet powerful method for combining multiple models, its integration into ensemble learning was significantly advanced by Leo Breiman with his introduction of Bagging in 1996. Subsequent methods like Boosting by Robert Schapire further leveraged voting mechanisms to enhance model performance. These contributions, along with earlier exploratory research, established majority voting as a cornerstone technique in ensemble methods used widely in machine learning, data science, and various applied fields today. [18]

$$f(X) = \text{MajorityVote}\left(T_1(X), T_2(X), \ldots, T_m(X)\right) \tag{5.2}$$

- $f(X)$: Final prediction for input X.
- $T_i(X)$: Output from the i^{th} decision tree.
- m: Number of decision trees in the forest.

Support Vector Machine (SVM)

SVM is *particularly useful* in this project for its ability to handle both linear and non-linear relationships through the use of kernel functions. In an e-commerce setting, where customer behavior can exhibit complex patterns, SVM is advantageous for its capacity to create an optimal decision boundary (hyperplane) in high-dimensional feature space. Despite being computationally intensive, SVM's effectiveness in achieving high precision makes it an ideal choice for accurately identifying churned customers, *ensuring minimal false negatives*, which is critical for targeted retention strategies.

Frank Rosenblatt (1958) introduced the decision function with the Perceptron, making it one of the first linear classifiers.Vladimir Vapnik and Alexey Chervonenkis (1963) advanced this with the development of Support Vector Machines (SVM), which also rely on a similar linear decision function but with a focus on maximizing the margin between classes. [19]
SVM aims to find the hyperplane that maximizes the margin between two classes. The decision function is defined as:

$$f(X) = \text{sign}\left(w^T X + b\right) \tag{5.3}$$

- w: Weight vector.
- b: Bias term.
- X: Feature vector.
- sign: Returns $+1$ or -1 depending on the side of the hyperplane.

In case of non-linear data, SVM uses kernel functions (ϕ) to map data to a higher-dimensional space:

$$f(X) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(X, X_i) + b\right) \tag{5.4}$$

- $K(X, X_i)$: Kernel function (e.g., linear, polynomial, or RBF).
- α_i: Lagrange multipliers.
- y_i: Class label for the i^{th} data point.

Decision Tree

The Decision Tree algorithm is a simple yet powerful model that divides the data into branches based on feature values. It is particularly suitable for this project because it can handle both numerical and categorical data, which is essential given the diverse types of features in the dataset, such as tenure, order counts, and preferred payment modes. Additionally, its interpretability allows for clear *visualization* of the decision-making process, helping stakeholders understand the key factors driving churn. However, care is taken to prevent overfitting by using techniques like pruning and setting maximum tree depth.

John Ross Quinlan (1986): Quinlan is credited with developing the ID3 algorithm, which uses Information Gain as a criterion for selecting the feature that best separates the data at each node. This method of feature selection based on Information Gain is fundamental to decision tree learning. [20]
At each decision node, a split is made based on a feature that maximizes information gain (IG):

$$IG = H(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} H(S_i)$$

- $H(S)$: Entropy of the dataset S.
- S_i: Subset of S after splitting.
- $|S_i|$: Size of subset S_i.
- k: Number of subsets.

For classification, entropy is defined as:

$$H(S) = -\sum_{i=1}^{c} P_i \log_2(P_i)$$

- P_i: Proportion of samples belonging to class i.

XGBoost

XGBoost is a high-performance *gradient boosting* algorithm that is well-suited for large datasets like the one in this project. It stands out for its ability to handle missing values, an important consideration given the presence of missing data in features like tenure and order counts. XGBoost also incorporates *regularization* techniques, such as L1 and L2 penalties, to prevent overfitting, making it a reliable choice for highly predictive yet *generalizable* models. Its speed and accuracy ensure that the model can scale effectively, allowing for real-time churn prediction and dynamic customer retention strategies.

XGBoost (Extreme Gradient Boosting) was developed by Tianqi Chen and Carlos Guestrin in 2016. XGBoost introduced several improvements to traditional gradient boosting, including a more efficient implementation, regularization techniques, and handling of missing data. The equation you provided, with the regularization term,XGBoost became widely popular due to its performance and efficiency, winning numerous machine learning competitions. [21]

$$\Omega(f_k)$$

is part of the XGBoost framework, where both loss and regularization are optimized together to reduce overfitting.

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

- $L(\theta)$: Objective function.
- $l(y_i, \hat{y}_i)$: Loss function (e.g., log loss for classification).
- $\Omega(f_k) = \frac{1}{2}\lambda\|w\|^2 + \gamma T_k$: Regularization term.
- T_k: Number of leaves in tree f_k.
- λ, γ: Regularization parameters.

Prediction is the sum of tree outputs:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i)$$

- f_k: The k^{th} decision tree.
- K: Total number of trees.

5.1 Model Evaluation and Validation

Some key metrics used to evaluate the performance of the models in this e-commerce churn prediction project are **accuracy**, **precision**, **recall**, and **F1-score**. These metrics play a vital role in assessing the effectiveness of the churn prediction models, especially in the context of imbalanced datasets, where the number of customers who have churned is significantly lower than those who have not.

Accuracy

Accuracy is the ratio of correctly predicted instances (both churned and non-churned) to the total number of predictions. It is calculated using the formula:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Predictions (TP + TN + FP + FN)}}$$

While accuracy provides a general measure of how well the model performs, it can be misleading when applied to imbalanced datasets. In this project, where non-churned customers outnumber churned customers, a model could achieve high accuracy by simply predicting the majority class (non-churned) more frequently, without correctly identifying churned customers. Therefore, while accuracy is considered, it is not the sole metric for evaluating model performance.

Precision

Precision measures the ratio of true positives (correctly predicted churners) to the total number of positive predictions (both true positives and false positives). It is given by:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

In churn prediction, high precision ensures that the model is effectively identifying customers who are truly at risk of churning. This is crucial because it allows companies to allocate retention resources only to customers who are likely to churn, avoiding

unnecessary costs associated with false positives. For example, if the model inaccurately predicts non-churners as churners, the company might waste resources targeting customers who were never at risk of leaving.

Precision is especially critical in this project because the dataset initially exhibited class imbalance. Feature selection, which incorporates key predictors such as tenure, complaints, and preferred payment mode, enhances precision by focusing on features that truly contribute to churn. Models like Random Forest and XGBoost, which rank feature importance, are particularly effective at improving precision. These techniques ensure the predictions are targeted and cost-effective, supporting better retention strategies.

Recall

Recall, also known as sensitivity, measures the proportion of actual positives (churners) that were correctly identified by the model. It is calculated as:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

High recall ensures that the largest possible portion of churners is identified, reducing the risk of overlooking at-risk customers. In the e-commerce context, where retaining customers is vital to profitability, recall is important to ensure that the majority of potential churners are identified, even if it comes at the expense of slightly lower precision. For this project, recall highlights the effectiveness of the model in capturing churn patterns and addressing potential blind spots in retention strategies.

F1-Score

The F1-score is the harmonic mean of precision and recall, balancing the trade-off between these two metrics. It is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is particularly useful in this project because it accounts for both false positives and false negatives, providing a balanced measure of the model's performance. For churn prediction, where both identifying true churners and minimizing false positives are critical, the F1-score ensures that the model achieves a reasonable trade-off. A high

F1-score signifies that the model is both precise and sensitive, effectively identifying churners without excessive false alarms.

5.2 Importance of Precision in Churn Prediction

Precision is one of the most critical metrics for this project because it ensures that retention efforts are focused only on customers who are truly at risk of churning. This prevents unnecessary spending on customers who are unlikely to churn (false positives), optimizing the allocation of resources. For example, if the model predicts a high number of false positives, the company might waste resources targeting customers who were not planning to leave the platform, which can reduce the overall cost-effectiveness of retention campaigns.

Incorporating advanced feature engineering techniques, such as derived metrics like loyalty index, churn ratio, and cashback effectiveness, enhances precision by improving the relevance of predictors. These metrics capture deeper insights into customer behavior, such as how cashback offers influence loyalty or how the frequency of orders relates to churn risk. Additionally, ensemble methods like Random Forest and XGBoost, which leverage feature importance rankings, further improve precision by prioritizing the most significant predictors.

5.3 Practical Application in This Project

By focusing on precision and other key metrics, this project successfully developed a churn prediction model that balances predictive accuracy with actionable insights. Key outcomes include:

- **Improved Precision:** Ensures retention efforts are directed at true churners, reducing wasteful expenditures on false positives.
- **High Recall:** Maximizes the identification of churners, ensuring proactive measures are taken for at-risk customers.
- **Optimized Retention Strategies:** Feature engineering and advanced algorithms enable cost-effective and targeted interventions, enhancing customer satisfaction

and profitability.

By leveraging precision alongside recall and the F1-score, the project has not only achieved robust model performance but also provided a practical framework for reducing churn and improving business outcomes in the competitive e-commerce domain.

CHAPTER 6

Results and Evaluation

6.1 Model Evaluation Results

The evaluation of the machine learning models for churn prediction was carried out using key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score. These metrics provided insights into how well each model performed in identifying churners while minimizing errors. Below is the detailed explanation of the results and how they were achieved for each model.

1. Logistic Regression

The Logistic Regression model achieved an accuracy of **82.3%** and a ROC-AUC score of **0.886**, indicating decent discrimination between churners and non-churners. The model's precision was **84%** for non-churners (class 0) and **81%** for churners (class 1), while its recall was **80%** for non-churners and **84%** for churners. These metrics suggest that the model effectively identified a significant portion of churners without overpredicting them.

The results of figure 6.1 were achieved by preprocessing the data through normalization and removing highly correlated features to ensure that the model captured the relationships between the predictors and the target variable effectively. The confusion matrix showed that while **789 churners** were correctly predicted (true positives), **147 churners**

were misclassified as non-churners (false negatives). This highlights the importance of recall in ensuring that at-risk customers are not overlooked.

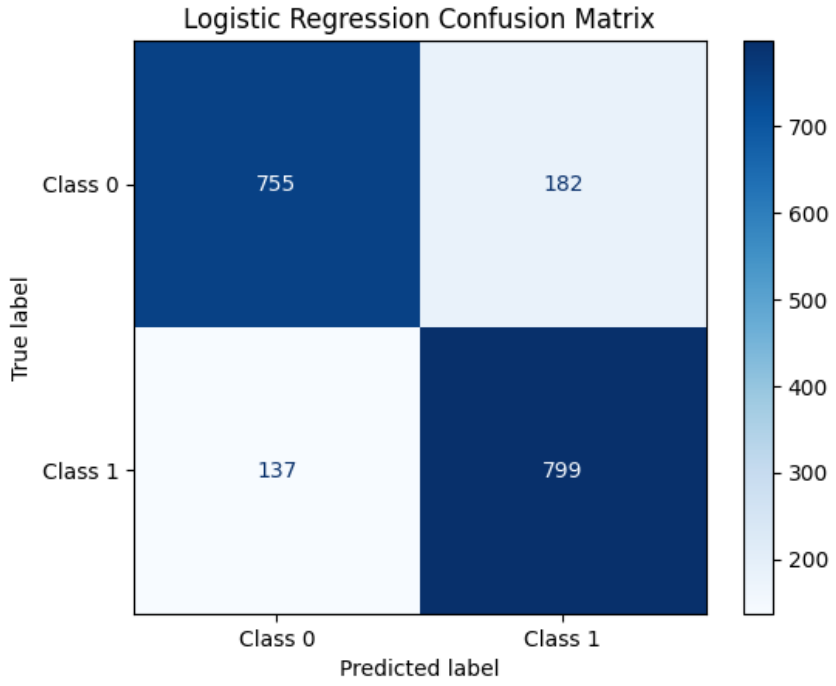

Figure 6.1: Logistic Regression Confusion Matrix.

Confusion Matrix Analysis for Logistic Regression:

- **True Negatives (TN):** 755 customers were correctly identified as non-churners.
- **True Positives (TP):** 799 churners were accurately identified.
- **False Positives (FP):** 182 customers were misclassified as churners, leading to unnecessary retention efforts.
- **False Negatives (FN):** 137 churners were missed, indicating room for improvement in recall.

Logistic Regression displayed a balanced performance but struggled with False Negatives (missed churners), which could affect retention strategies. Addressing these False Negatives through improved feature engineering or alternative models can significantly enhance customer retention efforts.

2. Random Forest

The Random Forest model emerged as the top-performing algorithm with an accuracy of 99.5% and a ROC-AUC score of 0.9999. The model demonstrated nearly perfect performance, with a precision of 100% for non-churners and 99% for churners. Its recall was 99% for non-churners and 100% for churners, leading to an impressive F1-score of 1.00 for both classes.

These results shown in figure 6.2 were achieved by leveraging the ensemble nature of Random Forest, which aggregates predictions from multiple decision trees. This approach mitigates overfitting and effectively captures complex, non-linear patterns in the data. Feature importance analysis revealed that variables like tenure, cashback effectiveness, and complaints played a crucial role in driving the model's accuracy. The confusion matrix confirmed that the model had minimal errors, misclassifying only 9 non-churners as churners.

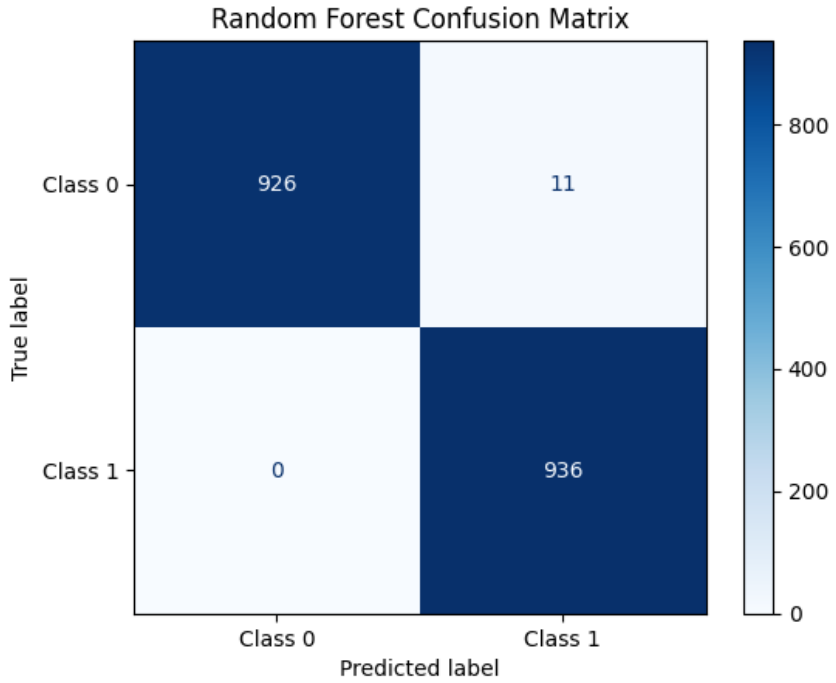

Figure 6.2: Random Forest Confusion Matrix

Confusion Matrix Analysis for Random Forest:

- **True Negatives (TN):** 926 non-churners were correctly classified.
- **True Positives (TP):** All 936 churners were identified accurately.

- **False Positives (FP):** Only 11 non-churners were incorrectly flagged as churners.
- **False Negatives (FN):** None, as the model captured all churners.

The Random Forest model performed exceptionally well, minimizing both FP and FN, making it highly reliable for churn prediction.

3. Support Vector Machine (SVM)

As shown in figure 6.3 the SVM model achieved an accuracy of 78.1% and a ROC-AUC score of 0.8536, indicating moderate performance. The model's precision was 82% for non-churners and 75% for churners, while its recall was 71% for non-churners and 85% for churners. The F1-scores were 77% for non-churners and 79% for churners, reflecting a reasonable balance between precision and recall.The SVM model performed well in identifying churners (high recall for class 1), which is critical for churn prediction. However, its performance was slightly hampered by its computational complexity and sensitivity to feature scaling. Kernel functions were applied to handle non-linear relationships, but the confusion matrix revealed that the model misclassified 268 non-churners and 142 churners, indicating room for improvement in precision.

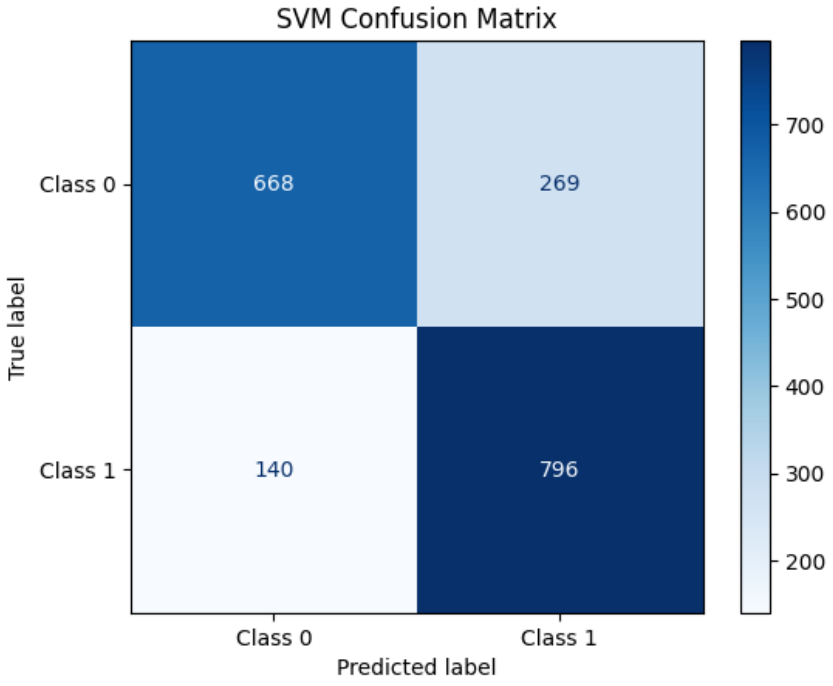

Figure 6.3: Confusion Matrix for SVM Model

Confusion Matrix Analysis for SVM Model

- **True Negatives (TN):** 668 non-churners were correctly classified.
- **True Positives (TP):** 796 churners were accurately identified.
- **False Positives (FP):** 269 non-churners were mistakenly flagged as churners.
- **False Negatives (FN):** 140 churners were missed.

The Support Vector Machine (SVM) model struggled with higher false positives (FP) and false negatives (FN) compared to Random Forest and XGBoost. While it successfully captured a substantial number of churners, achieving good recall, it was less effective in minimizing false positives, leading to unnecessary retention costs. These results highlight the need for further optimization of the SVM model to improve its precision and overall effe

4. Decision Tree

As shown in figure 6.4 the Decision Tree model achieved an accuracy of 99.2% and a ROC-AUC score of 0.9919, demonstrating strong performance. The model's precision and recall were 100% for non-churners and 98% for churners, resulting in a high F1-score of 0.99 for both classes. These metrics indicate that the Decision Tree was highly effective in classifying churners and non-churners.
The Decision Tree performance was enhanced by careful tuning of hyperparameters, such as maximum depth and minimum samples per leaf, to prevent overfitting. However, despite its strong performance, the confusion matrix showed that the model misclassified 15 non-churners, which could impact the precision of targeted retention strategies.
Confusion Matrix Analysis for Decision Tree:

- **True Negatives (TN):** 920 non-churners were correctly classified.
- **True Positives (TP):** All 936 churners were identified.
- **False Positives (FP):** Only 17 non-churners were flagged as churners.
- **False Negatives (FN):** None, as the model identified all churners.

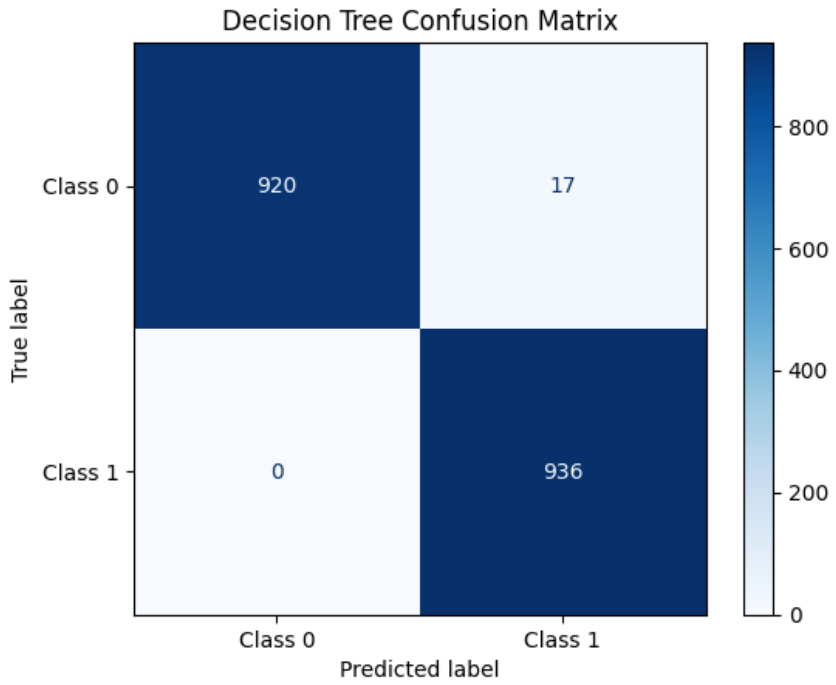

Figure 6.4: Confusion Matrix for the Decision Tree Model.

The Decision Tree model performed nearly as well as the Random Forest model, with very few false positives and no false negatives. This demonstrates its capability to classify churn effectively while maintaining interpretability. The combination of high precision and recall ensures that the model is both accurate and actionable for retention strategies.

5. XGBoost

As shown in figure 6.5 the XGBoost model achieved an accuracy of 99.4% and a ROC-AUC score of 0.9975, making it one of the top-performing algorithms in the project. Its precision was 100% for non-churners and 99% for churners, while its recall was 99% for non-churners and 100% for churners. These results translated to an F1-score of 0.99 for both classes.

XGBoost strong performance can be attributed to its ability to handle missing data and prevent overfitting through regularization techniques such as L1 and L2 penalties. The algorithm's gradient boosting framework efficiently captured the underlying patterns in the data, making it highly effective for churn prediction. The confusion matrix indicated only 10 misclassifications, underscoring its reliability.

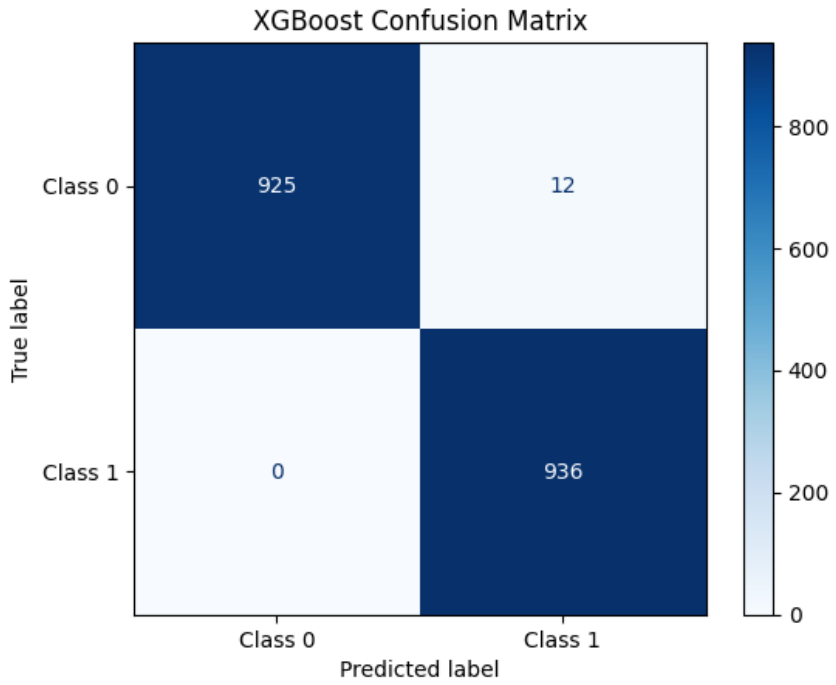

Figure 6.5: XGBoost Confusion Matrix

Confusion Matrix Analysis for XG Boost:

- **True Negatives (TN):** 925 non-churners were correctly classified.
- **True Positives (TP):** All 936 churners were identified.
- **False Positives (FP):** 12 non-churners were misclassified as churners.
- **False Negatives (FN):** None, as the model captured all churners.

XGBoost demonstrated near-perfect performance, with minimal FP and no FN, making it one of the best models for this project.

6.2 Model Performance Analysis Based on ROC Curves and AUC Scores

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is a vital metric that measures the ability of a model to distinguish between classes (churners and non-churners in this case). A higher AUC indicates better model performance, with a value of 1.00 representing perfect classification. The ROC curve

itself visually represents the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) across different classification thresholds, making it a crucial tool for evaluating classification models.

6.2.1 Logistic Regression

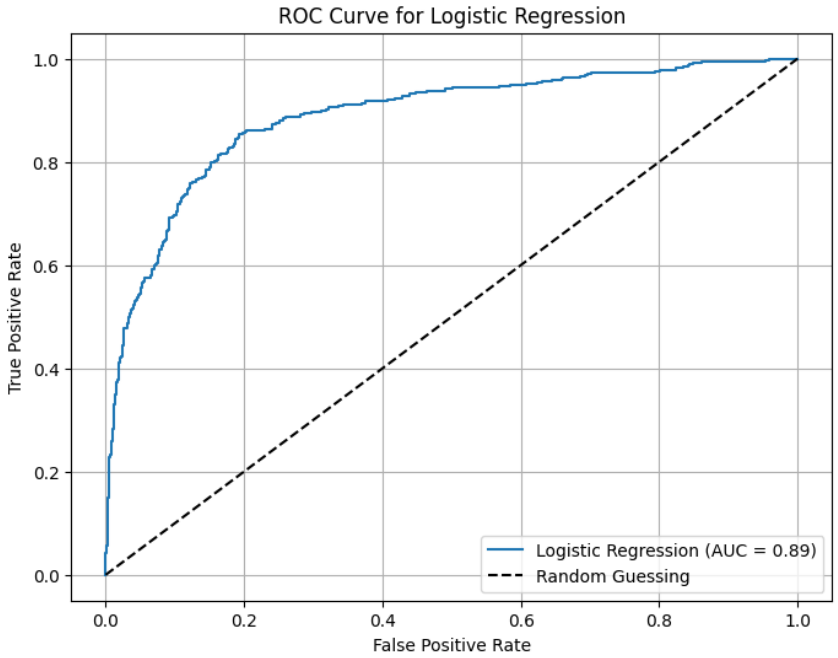

Figure 6.6: ROC Curve for Logistic Regression

The ROC curve for Logistic Regression (Figure 6.6) shows a steady increase in the True Positive Rate (TPR) as the False Positive Rate (FPR) rises. However, the curve does not closely align with the top-left corner of the graph, indicating moderate performance in separating churners from non-churners. Logistic Regression achieved an AUC score of 0.89, reflecting good but not outstanding performance in churn prediction.

As a linear model, Logistic Regression works effectively for datasets with linearly separable classes. It provides interpretable results, making it suitable as a baseline model. However, its inability to capture non-linear relationships limits its effectiveness in datasets with complex patterns, such as customer behavior in e-commerce. While Logistic Regression can deliver quick results and offers simplicity in understanding the influence of individual features, it is less competitive when compared to ensemble models like Random Forest and XGBoost, which excel in handling intricate data

structures.

6.2.2 Random Forest

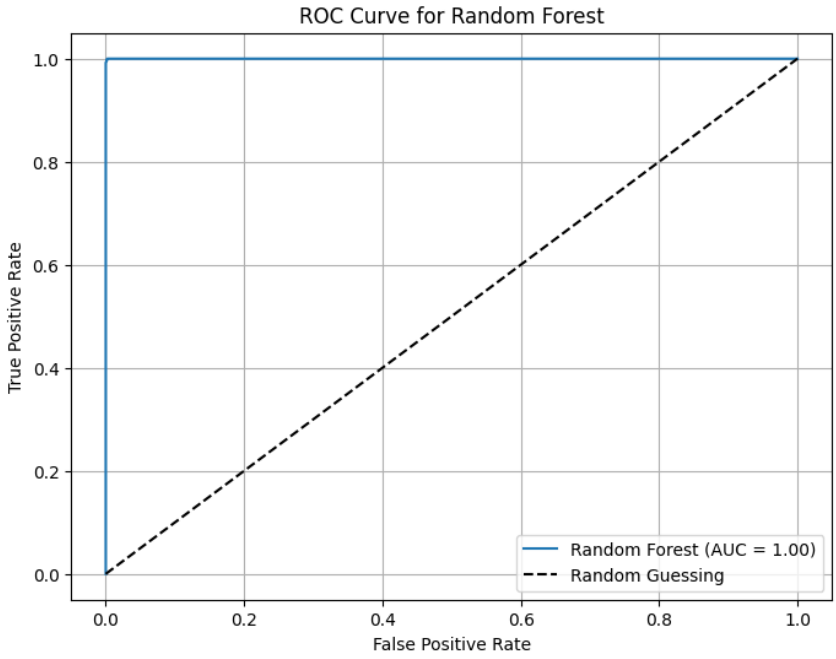

Figure 6.7: ROC Curve for Random Forest

Random Forest (Figure 6.7) demonstrates exceptional performance, achieving a perfect AUC score of 1.00. Its ROC curve aligns almost perfectly with the top-left corner, highlighting its outstanding capability to distinguish between churners and non-churners at all threshold levels. This ensemble model combines the predictions of multiple decision trees to improve accuracy and generalization while reducing the risk of overfitting.

The strength of Random Forest lies in its ability to handle complex, non-linear relationships within the dataset, making it a robust choice for churn prediction. Furthermore, it effectively addresses class imbalance through bootstrapping and provides feature importance rankings, which can offer actionable insights into the key drivers of customer churn. However, Random Forest can be computationally intensive, requiring more resources and longer training times compared to simpler models. Despite this limitation, its near-perfect classification performance and ability to generalize well to unseen data make it an excellent tool for e-commerce platforms aiming to implement data-driven retention strategies.

6.2.3 Decision Tree

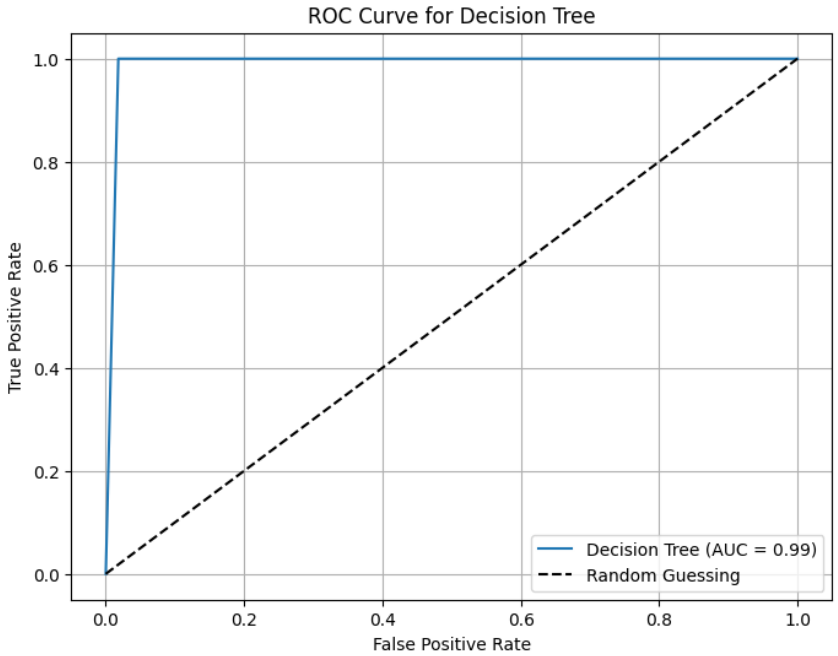

Figure 6.8: ROC Curve for Decision Tree

The Decision Tree model (Figure 6.8) also performed notably well, with an AUC score of 0.99. Its ROC curve approaches the top-left corner, indicating strong classification performance. Decision Trees are interpretable and computationally efficient, making them a valuable choice for understanding the hierarchy of decisions that lead to churn.

However, as a standalone model, Decision Trees are prone to overfitting, which can limit their generalization capabilities. This limitation is addressed in ensemble methods like Random Forest, which aggregate multiple decision trees to improve robustness. While the Decision Tree performs exceptionally well in this case, it may not be as reliable as Random Forest or XGBoost for large-scale, high-dimensional datasets commonly found in e-commerce. Despite this, its interpretability and simplicity make it a valuable tool for gaining initial insights into customer churn dynamics.

6.2.4 Support Vector Machine (SVM)

The Support Vector Machine (SVM) model (6.9) achieved an AUC score of 0.85, reflecting moderate performance in churn prediction. Its ROC curve rises gradually and does not

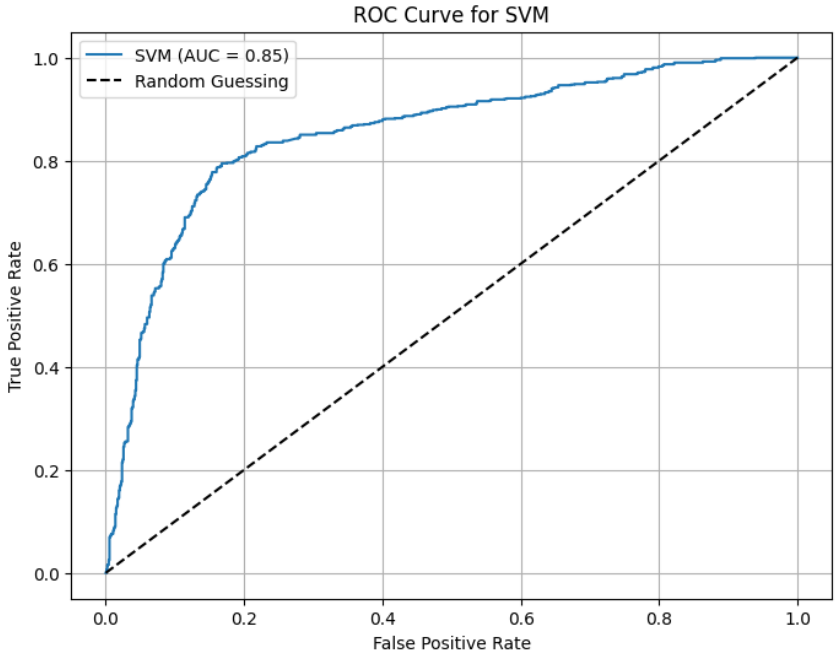

Figure 6.9: ROC Curve for SVM

closely approach the top-left corner, suggesting that the model struggles to effectively classify churners and non-churners at certain thresholds.

SVM is effective for non-linear classification tasks, especially when kernel functions are applied. However, it faces challenges in handling large, high-dimensional datasets due to its computational complexity. Additionally, SVM is sensitive to feature scaling, requiring extensive preprocessing to achieve optimal performance. While SVM provides reasonable results, its scalability issues and lower AUC score make it less practical for e-commerce datasets compared to ensemble methods like Random Forest and XGBoost.

6.2.5 XGBoost

XGBoost (Figure 6.10) is one of the top-performing models, achieving a perfect AUC score of 1.00. Its ROC curve closely aligns with the top-left corner, signifying excellent classification performance across all thresholds. XGBoost uses a gradient boosting framework to iteratively minimize errors and enhance predictive accuracy, making it highly efficient and robust.

XGBoost excels in handling missing data and imbalanced datasets, which are common in customer churn prediction. It also incorporates regularization techniques to prevent

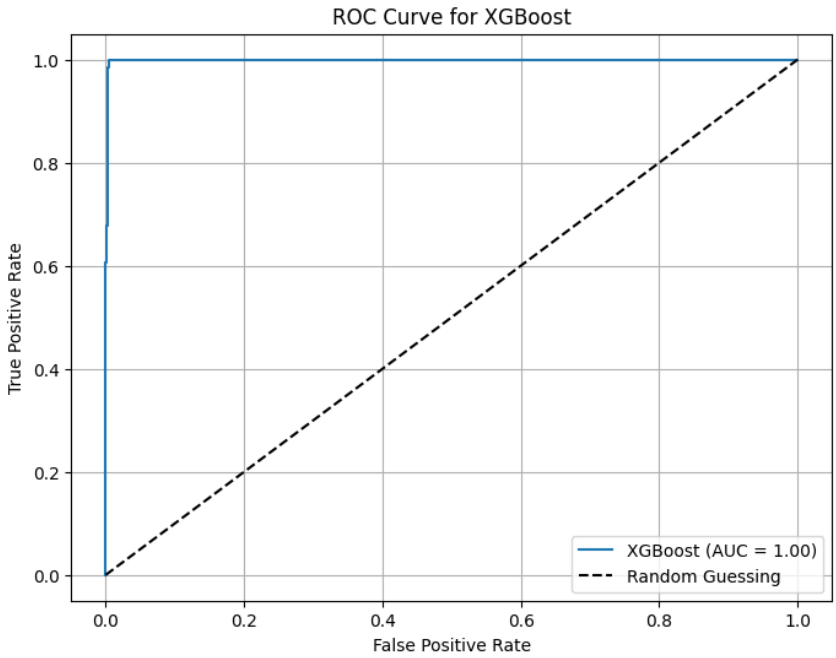

Figure 6.10: ROC Curve for XGBoost

overfitting, further enhancing its ability to generalize to new data. These features make XGBoost particularly suitable for large-scale e-commerce applications, where computational efficiency and accuracy are critical. Its exceptional performance highlights its capability to drive effective, data-driven customer retention strategies.

6.3 Summary of Model Performance

Model	**Accuracy**	**ROC-AUC**
Logistic Regression	82.3%	0.886
Random Forest	99.5%	0.9999
SVM	78.1%	0.8536
Decision Tree	99.2%	0.9919
XGBoost	99.4%	0.9975

Table 6.1: Summary of Model Performance

The evaluation revealed that Random Forest and XGBoost are the most effective models for churn prediction in this project, achieving near-perfect performance across all metrics. These models are ideal for deployment in the e-commerce setting, where

accurately identifying churners is critical for targeted retention efforts. Logistic Regression provided a reliable baseline, while SVM and Decision Tree offered additional insights but were outperformed by the ensemble methods. The robust performance of these models underscores the importance of feature engineering, oversampling, and algorithm selection in achieving high predictive accuracy.

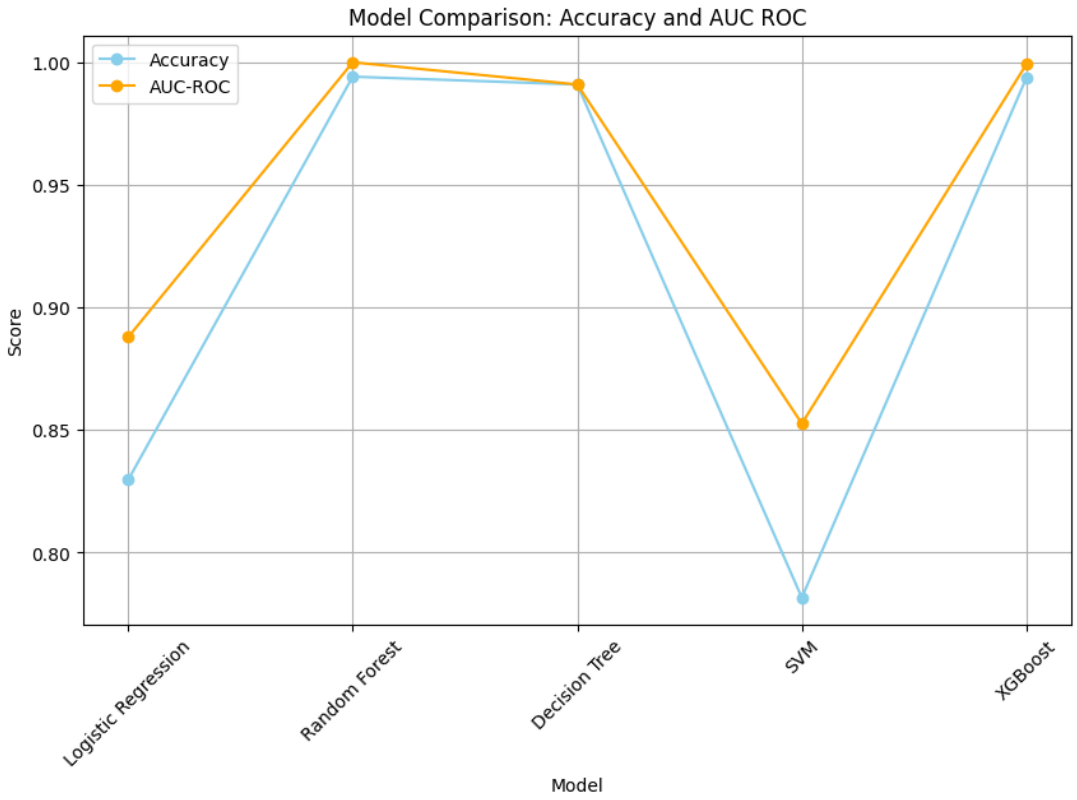

Figure 6.11: Model Comparison: Accuracy and AUC ROC

The evaluation of the machine learning models clearly showed that Random Forest and XGBoost are the top performers for predicting customer churn in this project. These models excelled in nearly every metric, such as accuracy, precision, recall, and F1-score, making them the most reliable options. Their ability to understand and model complex patterns in customer behavior, along with their robustness in handling imbalanced datasets, makes them perfect for deployment in an e-commerce setting. This is particularly important when the goal is to accurately identify customers at risk of churning so businesses can take timely action to retain them.

Logistic Regression, while not as powerful as the ensemble models, proved to be a dependable starting point. It provided a solid baseline by identifying clear, linear

relationships between features and churn. However, its simplicity meant it wasn't as effective in capturing the more intricate customer behavior patterns needed for deeper insights.

The SVM (Support Vector Machine) model performed fairly well, especially in identifying churners, which is critical for customer retention efforts. However, it struggled with a higher number of false positives, meaning it sometimes flagged loyal customers as churners, which could lead to unnecessary retention efforts. Additionally, SVM's computational demands made it less practical for larger datasets.

The Decision Tree model offered strong accuracy and was easy to interpret, which makes it useful for explaining predictions. However, compared to Random Forest and XGBoost, it lacked the added benefit of combining multiple trees to reduce errors, which limited its overall effectiveness.

What stood out in this project was how the preprocessing steps, such as oversampling to balance the dataset and creating meaningful derived metrics like the loyalty index and churn ratio, helped all the models perform better. These steps made the data more informative, giving the models a better chance to succeed.

In the end, Random Forest and XGBoost not only performed the best but also proved to be the most practical choices for implementation. They excel at pinpointing at-risk customers while minimizing unnecessary actions, making them perfect tools for crafting targeted and cost-effective customer retention strategies. This project shows how combining smart preprocessing, thoughtful feature engineering, and powerful algorithms can lead to meaningful and actionable insights.

CHAPTER 7

Conclusion

Customer churn prediction remains a critical challenge for e-commerce platforms, where retaining existing customers proves significantly more cost-effective than acquiring new ones. This study tackled this challenge by employing machine learning models to predict churn and provide actionable insights for designing targeted retention strategies.

A comprehensive dataset capturing customer demographics, engagement metrics, and transaction data was utilized, prepared through exploratory data analysis, feature engineering, and oversampling techniques to address class imbalance. Five machine learning models2̌014Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, and XGBoost2̌014were implemented and evaluated on key performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

The evaluation revealed critical insights into the effectiveness of different models:

XGBoost and Random Forest emerged as the top-performing models, with XGBoost achieving an AUC of 0.9975 and Random Forest achieving a perfect AUC of 1.00. While both models demonstrated near-perfect accuracy (99.4% for XGBoost and 99.5% for Random Forest), XGBoost proved more efficient in terms of training speed and scalability, making it better suited for handling large, high-dimensional datasets typical in e-commerce platforms.

Random Forest, while achieving slightly higher accuracy, required more computational

resources and longer training times due to the aggregation of numerous decision trees. In contrast, XGBoost's gradient-boosting approach efficiently focused on minimizing errors iteratively, enhancing performance on challenging, imbalanced datasets.

Decision Tree delivered robust results with an accuracy of 99.2% and an AUC of 0.9919. However, it lacked the aggregation benefits provided by ensemble methods, leading to slightly reduced robustness compared to XGBoost and Random Forest.

Logistic Regression served as a reliable baseline with an accuracy of 82.3% and an AUC of 0.886. While it effectively captured linear relationships, its inability to handle non-linear patterns limited its predictive power. Support Vector Machine (SVM) achieved moderate performance with an accuracy of 78.1% and an AUC of 0.8536. However, its sensitivity to feature scaling and higher computational complexity rendered it less practical for large-scale datasets. The study highlights the pivotal role of advanced feature engineering, such as derived metrics like Loyalty Index and Churn Ratio, in enhancing predictive accuracy. Moreover, the use of oversampling techniques ensured balanced representation of churned and non-churned customers, further improving the robustness of the models.

The findings validate the effectiveness of ensemble methods, particularly XGBoost, for customer churn prediction. While Random Forest demonstrated exceptional performance, XGBoost's gradient-boosting framework offered distinct advantages in terms of computational efficiency, scalability, and error minimization. These attributes make XGBoost the preferred choice for real-world applications where rapid, accurate predictions are essential.

By leveraging these advanced models, e-commerce platforms can proactively identify at-risk customers, enabling timely and personalized retention strategies. Such interventions enhance customer satisfaction, reduce churn rates, and drive long-term profitability, demonstrating the transformative potential of machine learning in modern business environments.

Future Works

Future research in e-commerce customer churn prediction could focus on enhancing model robustness and interpretability. While this study demonstrated the effectiveness of ensemble learning techniques like Random Forest and XGBoost, exploring advanced deep learning models, such as Long Short-Term Memory (LSTM) networks and Transformers, could enable better handling of sequential customer behavior data and provide real-time churn predictions.

Moreover, integrating external data sources, such as market trends, competitor activities, and macroeconomic indicators, may enrich feature sets and improve model performance. These factors could help models adapt dynamically to changes in consumer behavior, particularly during significant economic shifts like the COVID-19 pandemic.Another promising avenue lies in developing explainable AI (XAI) frameworks to improve stakeholder trust in predictive models. Incorporating SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) can help businesses understand why specific predictions are made, leading to more targeted and actionable retention strategies.

Finally, implementing real-time churn prediction systems using edge computing or cloud-based platforms could enable businesses to act immediately on at-risk customers. This can be coupled with automated retention campaigns tailored to individual customer preferences and behaviors.

Bibliography

[1] "Global e-commerce sales report," 2023. [Online]. Available: https://www.statista.com

[2] J. Chen *et al.*, "Impact of covid-19 on e-commerce trends," *Journal of Business Research*, vol. 124, pp. 456–469, 2021.

[3] F. Reichheld, *The Loyalty Effect*. Harvard Business Review Press, 1996.

[4] A. Smith and L. Johnson, "Personalized marketing strategies for customer retention," *Journal of Digital Commerce*, vol. 15, no. 3, pp. 240–258, 2020.

[5] X. J. Wu and S. S. Meng, "Research on e-commerce customer churn prediction based on customer segmentation and adaboost," *Industrial Engineering*, vol. 20, no. 2, pp. 99–107, 2017.

[6] N. Gordini and V. Veglio, "Machine learning for customer churn prediction," *Expert Systems with Applications*, vol. 185, p. 115632, 2022.

[7] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," in *European Conference on Computational Learning Theory*. Springer, 1995, pp. 23–37.

[8] Z. Wang *et al.*, "Loyalty programs and customer retention in retail," *International Journal of Retail & Distribution Management*, vol. 50, no. 4, pp. 457–478, 2022.

[9] Y. L. Y. Y. X. Feng, C. Wang and H. G. An, "Research on customer churn prediction based on comment emotional tendency and neural network," *Journal of China Academy of Electronics Science*, vol. 13, no. 3, pp. 340–345, 2018.

[10] F. Reichheld and W. Sasser, "Zero defections: Quality comes to services," *Harvard Business Review*, vol. 68, no. 5, pp. 105–111, 1990.

[11] X. W. L. N. Lu and L. Lee, "Research on customer value segmentation of online shop based on rfm," *Computer Knowledge and Technology*, vol. 14, no. 18, pp. 275–284, 2018.

[12] R. P. S. B. S. Dhote, C. Vichoray and P. M. Shakeel, "Hybrid geometric sampling and adaboost-based deep learning approach for data imbalance in e-commerce," *Electronic Commerce Research*, vol. 20, no. 2, pp. 259–274, 2020.

[13] J. Huang, "A comparative study of social e-commerce and traditional e-commerce," *Economic and Trade Practice*, no. 23, pp. 188–189, 2018.

[14] A. G. S. Agrawal, A. Das and S. Dhage, "Customer churn prediction modelling based on behavioural patterns analysis using deep learning," in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 2018, pp. 1–6.

[15] S. Kim and H. Lee, "Customer churn prediction in influencer commerce: An application of decision trees," *Procedia Computer Science*, vol. 199, pp. 1332–1339, 2022.

[16] H. Jain, A. Khunteta, and S. Srivastava, "Churn prediction in telecommunication using logistic regression and logit boost," *Procedia Computer Science*, vol. 167, pp. 101–112, 2020.

[17] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.

[18] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

[19] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[20] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.