A Smart Model to Enhance the Customer Retention

[1]Norihan N. Abdelgawad
[2] El-Henawy, I.M
[3] Aya M. Mostafa

Abstract

Customer Relationship Management is essential for company's success, focusing on enhancing customer retention through early detection of churn. Due to the Exponential growth of data volume in e-commerce, it is essential to develop techniques to identify customer churn. Machine learning and CRM will effectively perform this task. Additionally predicting rush time linking optimized Employees distribution and minimizing technical issues to achieve customer retention. We applied K-means based on the RFM model and addressed imbalanced data using SMOTE-TOMEK. The Results revealed that XG-Boost outperformed other algorithms, including decision tree and random forest, achieving an accuracy of 87.37%, precision of 99.99%, recall of 74.68%, F1-score of 85.5%, and an AUC of 88%. Additionally, the analysis using z-score showed that rush times between 4:57 AM and 6:04 PM accounted for approximately 68% of all purchases.

Keywords: Customer retention, Customer churn prediction, E-commerce, Ensemble learning algorithms, Recency Frequency Monetary (RFM) model, Rush time prediction, Z-score.

[1] Researcher and Teaching assistant in October High Institute for Engineering & Technology. Researcher at Faculty of Commerce and Business Administration- Helwan University.

[2] Professor of Computer Science at the Faculty of Computer Science, Zagazig University.

[3] Assistant Professor of Information Systems, Information Systems Department, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt.

نموذج ذكي لتحسين الاحتفاظ بالعملاء

مستخلص

إن ادارة علاقات العملاء ضرورية لنجاح أى شركة مع التركيز بشكل خاص على الأحتفاظ بالعملاء من خلال الكشف المبكر عن رحيل العملاء .

فبسبب النمو المتزايد لحجم البيانات فى التجارة الألكترونية فمن الضرورى تطوير تقنيات تستطيع الكشف المبكر عن رحيل العملاء .

يستطيع أن يحظى التعلم الألى و إدارة علاقات العملاء بشكل فعال فى تأدية هذه المهمة بالأضافى إلى ذلك ، يؤدى التنبؤ بوقت الذروة إلى تحسين توزيع العاملين و تقليل المشاكل الفنية لتحقيق الأحتفاظ بالعملاء .

لقد طبقنا الخوارزمية التصنفية بناء على أداة تحليل الحداثة و التكرار و القيم النقدية ، كما تنولنا معالجة البيانات الغير متوازنة بستخدام تقنية SMOTE-TOMEK .

و قد كشفت النتائج أن خوارزميXG-Boost تتفوق على الخوارزميات الأخرى بما فى ذلك شجرة القرارات و الغابة العشوائية , محققة دقة بنسبة 87.37%، Precision 99.99، Recall بنسبة 74.68%، F1-score بنسبة 85.5% و معيار قياس المساحة تحت المنحني بنسبة 87% بالاضافة الي ذلك, اظهرت الاحصائيات بأستخدام Z-Score أن اوقات الذروة بين الساعة الخامسة صباحاً و حتي السادسة مساءاً تمثل حوالي 68% من إجمالي المشتريات.

الكلمات المفتاحية: الإحتفاظ بالعملاء, التنبؤ برحيل العملاء، التجارة الإلكترونية ، خوارزمية التعلم الجماعي، نموذج الحداثة- التتكرار- النقدي ، التنبؤ بوقت الذروة.

1-Introduction

Customers represent a fundamental asset for enterprises, significantly influencing market competitiveness and overall performance [1].

Customer retention is crucial for companies as it enables a focus on satisfying existing customers over acquiring new ones. Satisfied customers contribute to positive word-of-mouth, while unhappy customers may share negative experiences. Long-term customers are more loyal and cost-effective to serve due to a better understanding of their preferences and demands [2].

Customer churn occurs when a customer discontinues making purchases from a company, impacting various organizational processes and overall profitability. Accurate prediction of customer churn is crucial for areas such as marketing, sales, and acquisition. Identifying customer churn is a key requirement for implementing effective retention activities [3].

Churn prediction, a critical aspect within the Big Data domain, is a demanding use case with significant effects on business health and growth. It serves as a key indicator for businesses of any size and those operating through various sales channels, enabling specialists to estimate the number of customers likely to discontinue product or service subscriptions within a specified time frame [4].

E-commerce data is referred to as "Big data" and it is a big challenge. It requires careful analysis to uncover hidden patterns for better decision-making. The complexity increases as we need to choose the right algorithms to improve overall performance [5].

The e-commerce marketplace, also known as online retail, functions as a platform or website where customers can browse a wide range of product brands displayed alongside those of other vendors, businesses, or individuals. The marketplace owner is responsible for attracting customers and facilitating transactions, while third-party providers handle the manufacturing and shipment aspects [6].

Nowadays, online shopping has become a widely adopted global trend, with approximately 2 billion people utilizing online platforms for their daily product purchases [7].

Today, e-commerce is essential for various reasons, offering simplicity, diverse products, flexible payment methods, and the ease of remote shopping. Technological advancements in mobile devices and widespread Internet access have greatly improved user experiences, directly impacting the revenue of e-commerce companies [8].

E-commerce platforms offer crucial insights into modern consumer behavior, preferences, and social interactions. Leveraging this data, businesses can develop strategies to enhance customer satisfaction and loyalty through personalized marketing and optimized shopping experiences. Promptly addressing customer concerns based on engagement metrics fosters stronger relationships and long-term success [9].

Machine learning, a subset of artificial intelligence (AI), involves the creation of algorithms and models designed to automatically learn, identify patterns, and make predictions or decisions from data [10].

The benefits of integrating machine learning with CRM, including enhanced social media data management, improved consumer comprehension, tailored sales strategies through predictive customer behavior, automated communication processes for marketing campaigns, and increased productivity and efficiency in handling customer data [11].

In the domain of big data and advanced analytics, machine learning is essential for customer churn analysis. These techniques enable businesses to identify patterns, predict churn behavior, and engage proactively with at-risk customers. The result is the implementation of targeted retention strategies to optimize customer lifetime value [12].

In the competitive e-commerce sector, developing a reliable predictive model for customer churn and rush time prediction is essential for effective retention strategies. Analyzing customer behavior, predicting churn and rush time prediction are important research areas, shaping strategies within the e-commerce platform. these tasks will be performed using machine learning and statistical method.

The rest of this paper is organized into 6 parts. The next section is the literature review. Section 3 provides proposed model for enhancing customer retention, detailed descriptions of data set used in this proposed model and the data pre-processing process. Section 4 introduces Performance evaluation, Model validation and The Experimental Results. Section 5 describes the implementation of rush time prediction, and finally, section 6 is the conclusion.

2-Literature Review

In Research [13], Hoang Tran and colleagues employed various machine learning models, such as decision tree, random forest, KNN, support vector machine, and linear regression, for predicting customer churn in banking. They addressed class imbalance using SMOTE and assessed the impact of customer segmentation on churn prediction accuracy. The results favored Random Forest, achieving 98.81% accuracy, 98.46% precision, 99.07% recall, and a 98.77% F1-score. Additionally, findings suggested that customer segmentation has a limited impact on prediction accuracy, influenced by the dataset and chosen models.

Manal Loukili and her colleagues in research [14] compared four machine learning algorithms for churn prediction in a telecom company, namely Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression, and Random Forest. The results revealed that the Support Vector Machine (SVM) model outperformed the other models. Adjusting the hyperparameters via cross-validation resulted in better accuracy. Specifically, the SVM accuracy increased from 83.46% to 96.92%, and the error rate reduced from 16.54% to 3.08%.

In researches [15][16], presented customer retention prediction models for telecommunications companies. In [15], they used XG Boost with four oversampling methods and found that XG Boost without oversampling outperformed SVM, Random Forests, Logistic Regression, and SGD Classifier in the initial experiment. The second part revealed the highest F-measure (around 84%) with the SMOTE method at a 20% oversampling ratio. In [16], focusing on the same objective and sector, Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Ada-Boost, XG-Boost, and Decision Tree were employed. The results showed that XG-Boost and Ada-Boost achieved the highest accuracy (81.71%, 80.8% respectively) and the highest AUC (84%) among all models.

Levent CALLI and Sena KASIM, in their research [17], presented a model for predicting customer churn in a software company, focusing on the active usage of software. They employed six machine learning algorithms, including Random Forest, K Nearest Neighbor (K-NN), Naïve Bayes, Logistic Regression, Neural Network, and Decision Tree. Notably, Random Forest achieved the highest accuracy among them. The research findings highlighted that key elements in customer churn analysis include the number of customers, products, and invoices.

Sayeda Farjana Shetu and her colleagues in [18] introduced a model for predicting customer satisfaction in the online banking system in Bangladesh. Six machine learning algorithms were employed, namely Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, Neural Network, and K-nearest neighbor (K-NN). Recursive Features Elimination (RFE) was used to reveal the top 10 features, of which Random Forest selected 7. The results showcased Random Forest's superior performance in accuracy, precision, recall, and F1 score. Finally, they identified the bank with the highest service quality that achieved customer satisfaction.

In their Research [19], Skandar Zul Putera Hamdan and Muhaini Othman employed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology and three algorithms—Decision Tree, Random Forest, and Logistic Regression—to predict customer loyalty in a hotel company. CRISP-DM consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The findings revealed that the decision tree algorithm was the most effective, achieving 71.44% accuracy, 73.01% precision, 74.81% recall, and a 74.00% F1-score when analyzing the hotel booking dataset.

Yuechi Sun et al. in [20], proposed a customer segmentation model based on customer lifetime value under the condition of noncontractual relationship due to the limitation and difficulty of using single data mining method in this condition.

They combined RFM model with machine learning algorithm, introduces the knowledge discovery method in data mining then constructed the model. They formulated the corresponding CRM strategy according to the customer group using an accurate CRM strategy and achieve customer retention.

Sabina-Cristiana Necula in research [21], analyzed the click stream data of consumers while shopping online by using machine learning algorithms to identify the factors that affect customer’s decision to make a purchase. The Results revealed that the time spent reading product related information, combined with other factors such as bounce rates, exit rates, and customer type significantly influences a customer’s purchasing decision.

Xiancheng Xiahou and Yoshio Harada in research [22], proposed a model to segment customers into three categories based on different times and then predict customer churn in E-commerce used two machine learning algorithms namely Ada Boost and BP neural network. The Results showed that Ada boost achieves the highest accuracy, clustering before prediction improve the prediction accuracy such as Ada Boost accuracy before segmentation was 94% but after segmentation increased to 95% and also found out that cluster I important to analyze because non churn rate was higher compared to cluster II and III customers.

Anne-Nee Wong and Booma Poolan Marikannan in [23], highlighted the importance of customer feedback in e-commerce success. They introduced a model that analyzed factors affecting customer satisfaction using four classification algorithms—Random Forest, Support Vector Machine (SVM), Decision Tree, and Artificial Neural Network—applied to a Brazilian E-commerce dataset. The findings highlighted Random Forest's highest accuracy, identifying meeting the estimated delivery date and delivery duration as the two most significant factors impacting customer satisfaction.

Traditional methods face challenges in predicting e-commerce customer churn because of the non-contractual nature of customers and high-dimensional, unbalanced data. The suggested PCA-AdaBoost model in [24] addresses these issues by introducing novel features into RFM analysis, minimizing data dimensions through principal component analysis, and employing adaptive boosting to manage unbalanced data. It improves performance metrics of churn prediction in e-commerce contexts and achieves an accuracy of 98%, precision of 98, recall of 99, and G-mean of 98%.

In researches [25][26], the primary goal is to detect customer churn in the telecom sector. Both studies reveal that Random Forest outperforms alternative models, including Decision Tree, Neural Network, Logistic Regression, ADA-Boost, and Naive Bayes. In paper [25], Random Forest achieves an accuracy of 88.8%, precision of 89%, F-measure of 88%, and an ROC area of 94.7%. Similarly, in paper [26], Random Forest achieves an accuracy of 85.71%, error rate of 14.29%, precision of 77%, recall of 100%, and an F-measure of 87%.

Takuma Kimura in [27], proposed a merging hybrid resampling such as SMOTE-ENN and SMOTE-Links as a more effective way to handle imbalanced data and ensemble learning algorithms for customer churn prediction model in a telecom company used classification algorithms and each algorithm was applied to four datasets: the imbalanced dataset, one resampled by SMOTE, SMOTE Tomek Links, and SMOTE-ENN. The Results discovered out that this hybrid resampling balance the data more effective and the Boosting algorithms outperformed the traditional classification algorithms.

In their research [28], Nikhil Patel and Sandeep Trivedi examined the impact of artificial intelligence applications on customer loyalty across 910 organizations worldwide. The study focused on AI features such as customer service, predictive modeling, ML-powered personalization, and natural language processing. Using six machine learning algorithms, including Support vector machine, Random Forest, Decision

Tree, Ada Boost, K-Nearest Neighbor, and Logistic Regression. The results highlighted that combining machine learning and natural language processing enhances customer loyalty, with Random Forest and Ada Boost showing superior accuracy.

In research [29], introduced a predictive model for customers' satisfaction and churn in an e-commerce platform. The study employed four machine learning algorithms, including Random Forest, Logistic Regression, Support Vector Machine, and Gradient Boosting. The findings revealed that the Gradient Boosting algorithm outperformed the others, achieving the highest accuracy for satisfaction 88%, satisfaction recall 69.8%, churn accuracy 95%, and churn recall 96%.

Peiyi Song and Yutong LIU in research [30], proposed a model to predict customer’s purchasing behavior used XG Boost and random forest to train and predict a real e-commerce platform user data. They discovered that XG Boost algorithm can effectively improve the accuracy of prediction and also discovered that exit rates the most important feature to predict customer purchasing behavior high rate means that the web content is less attractive to users so need to improve the web content.

Jesmi Latheef and S.Vineetha in research [31], proposed a model to predict customer churn and identify the factors that cause churn in bank. They used ensemble learning algorithm including Linear Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, Ada Boost, XG Boost and compared the accuracy between ensemble learning and each individual model. The Results found out that ensemble learning algorithm achieves highest accuracy 85% rather than each individual model and AUC-ROC 83%

In researches [32] [33] share a common goal of predicting customer churn in the telecom sector. Both studies conclude that ensemble learning algorithms outperform other methods. In

[32], Orange company data was utilized, employing individual algorithms such as Random Forest, ADA Boost, and Gradient Boosting. The stacking ensemble approach, combining these algorithms, resulted in a robust predictive model with notable metrics: accuracy 97.65%, precision 97.80%, recall 97.80%, F1-score 96.98%, and AUC 92.45%. In paper [33], 900,000 customer personal characteristics and historical data were utilized. The study employed an Ada Boost dual-ensemble learning model with Random Forest as the base learner. Additionally, four other models—Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Naive Bayes—were individually employed. The results demonstrated that the ensemble learning model, incorporating Random Forest as the base learner, outperformed the other models.

In research [34], Logistic Regression and Random Forest models were applied to predict customer churn in an e-commerce platform. Results showed that Random Forest outperformed logistic regression, emphasizing the effectiveness of ensemble learning. The study identified transaction frequency, customer tenure, and customer support contacts as key factors essential for improving customer retention in the e-commerce environment.

In research [35], Maha Alkhayrat and her colleague employed Principal Component Analysis (PCA) and an innovative Auto Encoder Neural Network approach to reduce a telecom dataset's high dimensionality (220 features for 100,000 customers). Clustering was performed on the original dataset and on datasets reduced by PCA and Encoder Neural Network. Results indicated that the reduced datasets enhanced clustering accuracy, with the best outcomes achieved by the Encoder Neural Network, reducing the original 220 features to just 20.

Zhang et al. in [36], enhanced XG Boost's classification performance on unbalanced data by employing mixed sampling technology and ensemble learning (SVM-SMOTE over-sampling and Easy Ensemble under-sampling). The approach balances data at both the data and category levels, using SVM-SMOTE and Easy Ensemble. At the algorithm level, XG Boost is trained with the Bayesian optimization algorithm for optimal parameter selection. Experimental results revealed the proposed model's superior effectiveness compared to the original XG Boost and other representative classification models (RUS Boost, CAT Boost, and Light GBM).

In research [37], the research introduced a multi-behavior RFM (MB-RFM) model, departing from traditional RFM models that focused solely on purchase behavior. Leveraging the self-organizing map (SOM) algorithm, the MB-RFM model considered interactions like clicking, favoriting, and adding to cart. By analyzing the weight relationship between user behaviors and items, the model classified customers using SOM neural networks. Experimental results on real-world datasets demonstrated the model's effectiveness in obtaining valuable information for developing marketing strategies (e.g., pricing, promotions, personalized services) and enhancing customer retention.

Mahaboob Basha in [38], highlighted that customer satisfaction and loyalty are essential for customer retention. The study indicated a strong positive correlation between factors like ease of use, service quality, security, and reliability in online shopping and customer loyalty. These findings underscore the importance of improving these aspects to maintain long-term customer retention.

Yuran Dong et al. in [39], presented a model for predicting customer behavior using Random Forest and the Mean Decrease in Impurity (MDI) method to identify impactful features on online purchase actions. The results highlighted user ID and user session as the most crucial features for predicting customer behavior. Moreover, product price, discount rate, and product ID were identified as significant factors influencing customer decisions. The study concluded that appropriate pricing, free shipping, and discount rates were the most important factors influencing customers to make frequent purchase orders.

SHULI WU and his colleagues in [40], three datasets were used to predict customer churn in the telecom sector, employing six machine learning algorithms and SMOTE for addressing unbalanced data. Ada Boost achieved the highest f1-score (63.11%) and AUC (84.52%) for the first dataset, while Random Forest excelled with the highest f1-score (77.20%) and AUC (91.40%) for the second dataset. Multi-layer Perceptron and Logistic Regression achieved the highest f1-score (42.84%) and AUC (58.66%) for the third dataset. These results highlight the importance of selecting algorithms based on the nature of the data set.

Reviewed previous studies utilizing machine learning algorithms for predicting customer churn, particularly in the E-commerce sector. We explored the data pre-processing methodologies adopted by each study. Additionally, we reviewed the challenges faced with imbalanced data and the corresponding strategies implemented for resolution. Furthermore, we reviewed the features and dataset sizes utilized in each study. Moreover, we reviewed the machine learning algorithms employed in these studies, highlighting the algorithm that achieved the highest performance. Finally, we clarified the performance metrics associated with the superior algorithm, as detailed in Table (1).

Table 1 Previous studies for customer churn on E-commerce

Paper	Objective	Data Pre-processing	Problems & How solve it	Data set (size, features)	Used Algorithms	High Performance	Performance metrices
[20]	To build a customer segmentation model based on customer lifetime value in a noncontractual relationship	- data cleaning	–	541,909 8 Features -Invoice no. -Stock code -Description -Invoice date -Unit price -Customer ID -County -Quantity	-KNN -Logistic regression -SVM -Decision tree -Random forest -AdaBoost -Gradient boosting decision tree -Naïve Bayes -Multilayer perceptron	-Gradient boosting decision tree	-Accuracy 93.80% -AUC 95.12%
[1]	-To predict customer churn and compare the performance of the prediction model before and after segmentation	-Feature selection (random forest)	-unbalanced data solved by SMOTE	-8,156 customers 17 features reduced into only 4 features -Night Buy -PM buy -Night PV -PM PV	-SVM -Linear regression	-SVM	The results improved after seg. - Accuracy 0.91 -Precision 0.97 -Recall 0.97
[21]	-To identify the factors that affect customer's decision to make a purchase	–	-unbalanced data solved by ADASYN algorithm	12,330 user sessions 17 features reduced into 10 features -Administrative -Administrative_ Duration -Informational - Informational_ Duration -Month	Logistic Regression -SVM -Random forest -Decision Tree	-SVM	-ROC-AUC 0.94 -Accuracy 0.90 -Recall 0.90

Paper	Objective	Data Pre-Processing	Problems & How Solve it	Data Set (size, features)	Used Algorithms	High Performance	Performance metrices
				-Product Related -Product Related_ Duration -Exit Rate -Pages Values -Visitor Type			
[22]	-To predict customer churn	-Feature Selection (Random forest	–	987,994 customers 17 features reduced into 12 features -Night Buy -PM Buy -Night PV -PM PV -AM PV -AM Buy -Categories -Daybreak PV -Night Cart -Daybreak Buy - PM Cart -AM Cart	-Ada Boost -Bp neural Network	-Ada Boost	The results improved after seg. -Accuracy 0.95 -Precision 0.90 -Recall 0.93
[23]	-To identify the key factors that influence customer satisfaction	-Feature Selection (decision tree) -Feature Scaling (normalization)	Unbalanced data (compare the accuracy of 4 methods SMOTE and Oversampling with the same high accuracy	112,000 orders 20 features reduced into 5 features -Delivery Performance -Purchase_	-Decision tree -Random forest -SVM -Artificial neural network	-Random forest	-Accuracy 87.6% -Sensitivity 98.0% -Specificity 29.5% -F1-Score 0.93

Paper	Objective	Data Pre-Processing	Problems & How Solve it	Data set (size, feature)	Used Algorithms	High Performance	Performance metrices
				delivery date -Order_ Qty -Sum_ pay _ value -Customer _ state			
[30]	To predict customer’s purchasing behavior	-Encoding (one-hot encoding) -Feature selection (Random forest and XG-Boost)	–	12,330 users 14 Features reduced into 9 features -Administrative -Product Related -Product Related_ Duration -Bounce Rates -Exit Rates -Page values -Month_ Category _Nov -Visitor Type_ Category_ Returning_ Visitor - Visitor Type_ Category_ New_ Visitor	-Random Forest -XG- Boost	-XG-Boost	-Accuracy 0.90 -Positive Precision 0.59 -Negative Precision 0.96 -Positive Recall 0.73 -Negative Recall 0.93 -Positive F1 Score 0.65 -Negative F1 Score 0.94
[34]	-To predict customer churn	-Handling missing values -Categorical encoding -Feature Scaling	–	30,15,356 customers data 8,37,868 products data	-Logistic Regression -Random Forest	-Random Forest	-Accuracy 84% -Precision 85% -Recall 96% -F1-score 90% -ROC AUC 87%

3-Propsed model for Enhancing Customer Retention

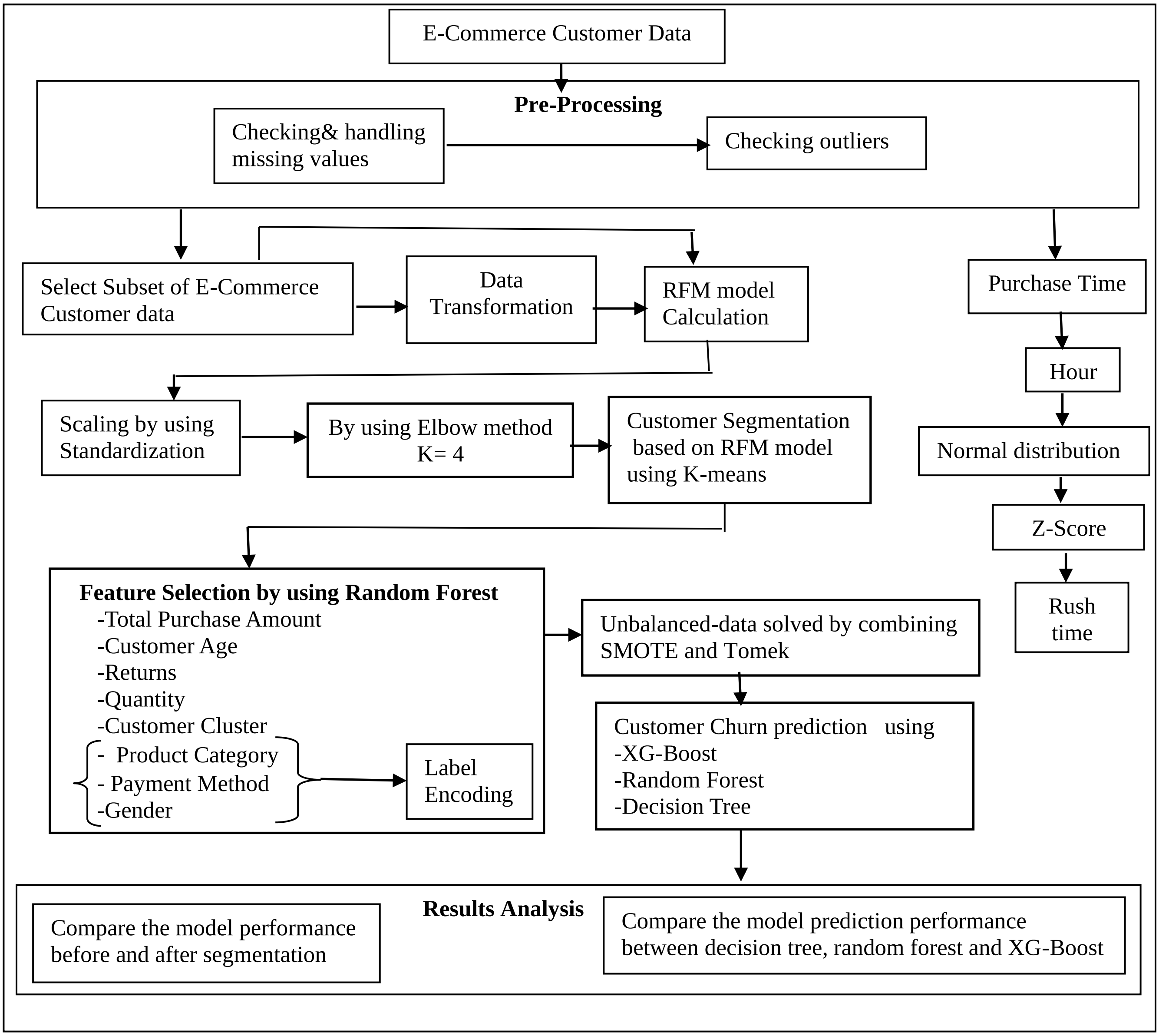

E-Commerce Customer Data
Pre-Processing
Checking& handling missing values
Checking outliers
Select Subset of E-Commerce Customer data
Data Transformation
RFM model Calculation
Purchase Time
Hour
Normal distribution
Z-Score
Rush time
Scaling by using Standardization
By using Elbow method K= 4
Customer Segmentation based on RFM model using K-means
Feature Selection by using Random Forest
-Total Purchase Amount
-Customer Age
-Returns
-Quantity
-Customer Cluster
- Product Category
- Payment Method
-Gender
Label Encoding
Unbalanced-data solved by combining SMOTE and Tomek
Customer Churn prediction using
-XG-Boost
-Random Forest
-Decision Tree
Results Analysis
Compare the model performance before and after segmentation
Compare the model prediction performance between decision tree, random forest and XG-Boost

Figure 1 Proposed model for enhancing customer retention

Data Set Description

An "E-commerce Customer Behavior and Purchase Dataset" is publicly available on Kaggle. This Data Set capturing various aspects of customer behavior and purchase history within a digital marketplace. The dataset contains 250,000 instances and 13 attributes. The last attribute denote churn or not, of which 200,000 are not churners and 50,000 are churners.

Data dictionary: Each row represents a customer, and each column contains the customer’s attributes, as described below in Table 2.

Table 2 data set used in the proposed model

Feature	**Description**
Customer ID	A unique identifier for each customer
Purchase Date	The date of each purchase made by the customer
Purchase Time	The Time of each purchase made by the customer
Product Category	The category or type of the purchased product
Product Price	The price of the purchased product
Quantity	The quantity of the purchased product
Total Purchase Amount	The total amount spent by the customer in each transaction.
Payment Method	The method of payment used by the customer (e.g., credit card, PayPal).
Customer Age	The age of the customer
Returns	Whether the customer returned any products from the order (binary: 0 for no return, 1 for return)
Customer Name	The name of the customer
Gender	The gender of the customer

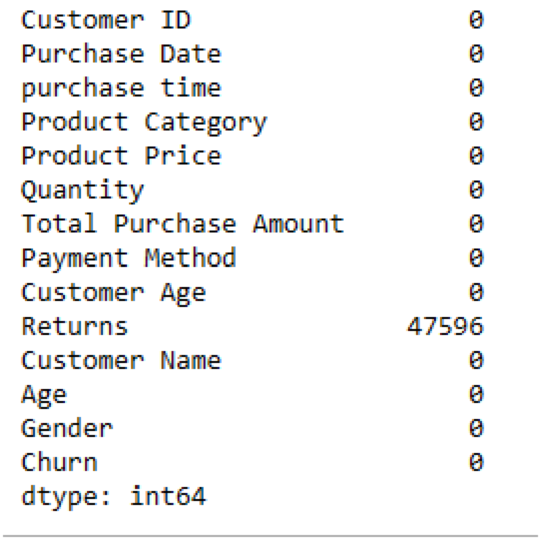
Churn	A binary column indicating whether the customer has churned (0 for retained, 1 for churned).

Data Pre-Processing

Data pre-processing involves the preparation and adjustment of raw data to make it suitable for analysis. This process typically includes cleaning, transforming, and organizing data to enhance its quality and usability. It is essential for improving the accuracy and effectiveness of machine learning models and other data analysis techniques.

Data Cleaning

This step involves removing or handling the missing data and also checking for outliers and also eliminating noise. Figure 2 illustrates that the dataset contains 47,596 missing values for the variable "Returns". we handle missing values in the 'Returns' column, which contains categorical data (0 or 1), by filling missing values with the mode. For customers with existing returns, missing values are replaced with the mode of their returns, maintaining the categorical nature. For customers with no recorded returns, any remaining missing values are filled with 0, ensuring consistency across the column.



```
Customer ID                  0
Purchase Date                0
purchase time                0
Product Category             0
Product Price                0
Quantity                     0
Total Purchase Amount        0
Payment Method               0
Customer Age                 0
Returns                  47596
Customer Name                0
Age                          0
Gender                       0
Churn                        0
dtype: int64
```

Figure 2 checking missing values

After checking the numerical features presented in figures 3, 4, 5, and 6, it was determined that no outliers were detected in the dataset.

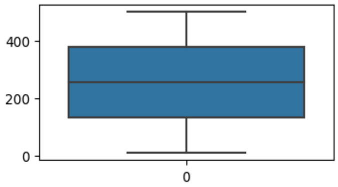


Figure 3 checking outliers for product price

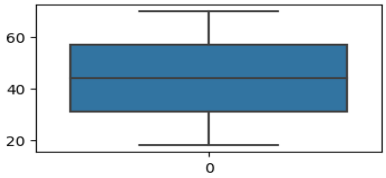
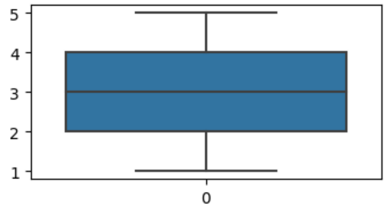


Figure 4 checking outlier for customer age



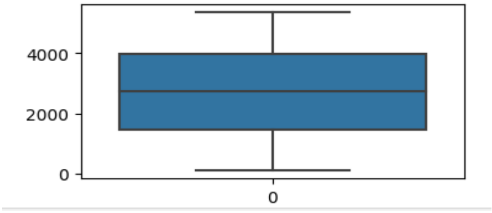


Figure 5 checking outliers for quantity

Figure 6 checking outliers for total purchase amount

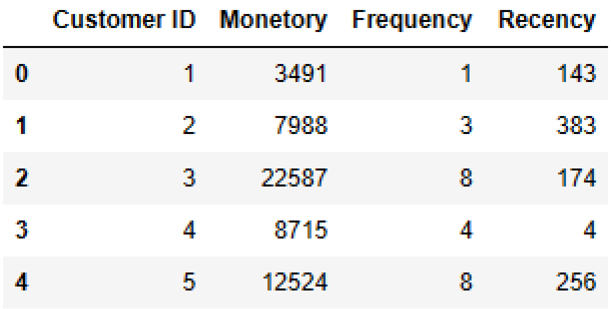
Feature Engineering

The features are preprocessed through the following steps, which are determined based on their data types.

Label encoding is applied to the following characteristics: "Product Category" and "Payment Method", each containing four variables, as well as "Gender" with two variables.
Label encoding is particularly suitable for handling categorical features with a small number of unique values, such as in this case where the maximum unique values for categorical features is four. Therefore, we employed label encoding to ensure that even with weighted algorithms, the data remains unaffected.

"Purchase Date" and "Purchase Time" are initially converted into string format. Then, they are combined into datetime objects. These datetime objects are merged into a single column named "Datetime" within the data frame. This preprocessing step enables easier manipulation and analysis of date and time data within the dataset and is crucial for facilitating the calculation of RFM (Recency, Frequency, Monetary) metrics. RFM analysis involves quantifying customer behavior based on how recently they made a purchase (Recency), how often they make purchases (Frequency), and how much they spend (Monetary).

The recency value in this study represents the difference between the current date and each customer's latest transaction date. This calculation involves setting the current date as the maximum date in the dataset, grouping transactions by customer ID, computing recency for each customer, and storing the results in a Data Frame named 'recency'. The frequency value is determined by counting how many items each customer has purchased. To do this, we first group transactions by customer ID. Then, within each customer group, we count the total number of items, indicated by the 'Quantity' column, to calculate the frequency for each customer. The monetary value is calculated by summing up the 'Total Purchase Amount' for each customer, which is achieved by grouping transactions by customer ID and calculating the sum of purchase amounts for each group. as shown in figure 7 RFM calculations.



	Customer ID	Monetory	Frequency	Recency
0	1	3491	1	143
1	2	7988	3	383
2	3	22587	8	174
3	4	8715	4	4
4	5	12524	8	256

Figure 7 RFM Calculation

After applying the RFM model for data transformation, it was observed that the dataset had a wide range, which could potentially impact the success of the clustering process. To address this concern, data standardization was implemented using the Standard Scaler technique, as it is essential to ensure uniformity and comparability for accurate analysis.

Customer segmentation using K-means

Customer churn management involves categorizing customers by shopping behavior, identifying valuable core customers and those at risk, allowing for customized marketing strategies suited to each group's needs. In this paper we employed the k-means clustering algorithm based on the RFM model, we chose to employ K-means in this study after checking outliers in the dataset and finding none, as K-means is appropriate in cases of normal distribution. We should understand the data itself to select the appropriate technique. The determination of the optimal number of clusters (K) was achieved using the elbow method as shown in figure 8, K=4.

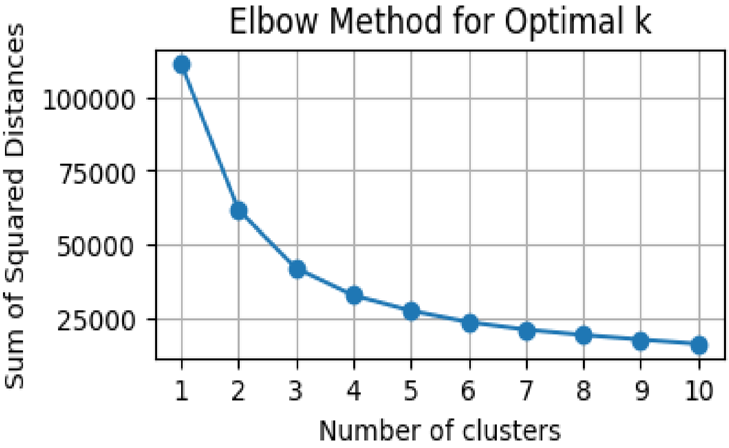


Figure 8 Elbow method

Then we classified customers into 4 groups based on the K-means clustering algorithm applied to the RFM model as shown in figure.

Data Balancing

Churned customers total 200,000 and non-churned customers amount to 50,000 as shown in figure 9.

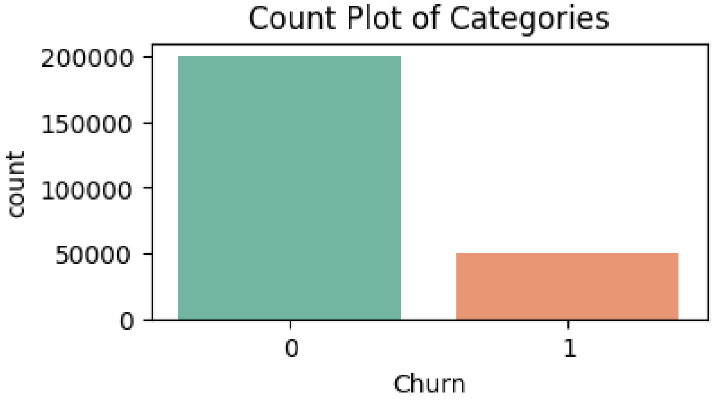


Figure 9 Churn rate

We solved imbalanced data by combining the Synthetic Minority Over-sampling Technique with Tomek links (SMOTE-TOMEK). SMOTE oversamples the minority class, while Tomek links under sample the majority class to reduce overlap and ambiguity. The combined approach aims to enhance dataset balance and improve predictive performance for customer churn. After applying SMOTE-Tomek combination as a resampling method, the class distribution is balanced, with counts of 197,145 instances for each class.

Feature Selection

Analyzing feature importance is crucial for predicting customer churn accurately. In the random forest algorithm, feature importance is evaluated by analyzing each feature's contribution value in every tree and then averaging these values across all trees. This comparison helps determine which features have the most significant impact on predicting customer churn.

In this study, Random Forest was employed, where feature selection naturally occurs during the construction of decision trees within the ensemble. Random Forest evaluates the importance of each feature based on its contribution to the overall performance of the forest.

As shown in Figure 10, it is observed that the first feature, "Total purchase amount," typically ranks highest in importance due to its strong correlation with customer churn.

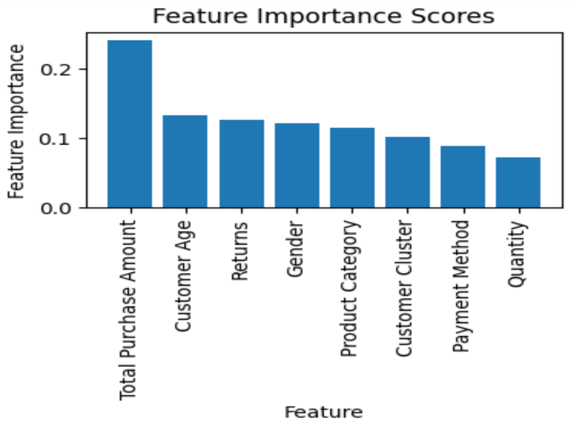


Figure 20 Feature selection by Random Forest

4- Performance Evaluation Matrix

After training the dataset with selected algorithms, the results are evaluated and compared during the evaluation phase. In this study, the proposed model's effectiveness in enhancing customer retention is evaluated using metrics such as accuracy, precision, recall, F1-score, and Receiver Operating Characteristic area (ROC).

-Accuracy represents the proportion of correct predictions out of the total number of predictions as shown in equation (1).

(1) accuracy = (TP+TN) / (TP + TN+ FP+ FN).

Here, "TN" represents True Negative, "TP" represents True Positive, "FN" represents False Negative, and "FP" represents False Positive. The TP Rate, also known as sensitivity, indicates the proportion of data correctly classified as positive.

-Precision is calculated as the number of correctly classified positive samples divided by the total number of samples classified as positive as shown in equation (2).

(2) precision = (TP) / (TP + FP).

-Recall is calculated as the number of positive samples correctly identified divided by the total number of actual positive samples in the testing set as shown in equation (3).

(3) recall = (TP) / (TP + FN).

- F-measure (F1-Score) is calculated as the weighted average of Precision and Recall as shown in equation (4).

(4) F1-score = 2 * ((precision * recall) / (precision + recall)).

-Receiver Operating Characteristic area (ROC) reflects the average performance across all possible cost ratios between false positives (FP) and false negatives (FN). A ROC area value of 1.0 indicates perfect prediction. Conversely, values such as

0.5, 0.6, 0.7, 0.8, and 0.9 correspond to random prediction, poor, moderate, good, and excellent performance, respectively.

Model Validation

Customer churn prediction is approached as a supervised learning classification problem. The dataset is divided into predictor (X) and label (y) subsets. During training, the model learns from the predictor variables to predict the target variable. In testing or validation, the model's performance is evaluated using the target variable to evaluate its generalization to unseen data. This setup enables machine learning algorithms to predict churn based on given features and ensures the evaluation of model performance and comparison across algorithms using unseen data.

K-fold cross-validation, a method utilized during hyperparameter tuning with techniques like "Grid Search" and "Randomized Search". In this study, we employed the Randomized Search method for hyperparameter tuning due to its efficiency, scalability, effectiveness, and cost-effectiveness, and the dataset was partitioned into 80% for training and 20% for testing. Additionally, a k-fold cross-validation approach with k set to 3 was adopted to compare model performance and evaluates model generalizability. Consequently, the dataset was divided into three groups, with 80% of the cases utilized for training in each group. Each model was trained and tested three times, and subsequently, performance metrics were computed as the average of the three folds.

Experimental Results

In this section, the results analysis for all algorithms is conducted, and the findings are presented in Table (3) below. After preprocessing the E-commerce dataset, it is divided into training and testing sets, composing 80% and 20% of the data, respectively, totaling 200,000 and 50,000 instances. Subsequently, the proposed machine learning algorithms, namely Decision Tree, Random Forest, and XG-Boost, are applied using the eight attributes selected by Random Forest and one target label. Figure 11 presents the confusion matrix results for the decision tree algorithm. According to the analysis, the True Negative (TN) value is 31021, the False Positive (FP) value is 8343, the False Negative (FN) value is 7493, and the True Positive (TP) value is 31965. In Figure 12, the Receiver Operating Characteristic area (ROC) for the decision tree algorithm is illustrated as 0.80.

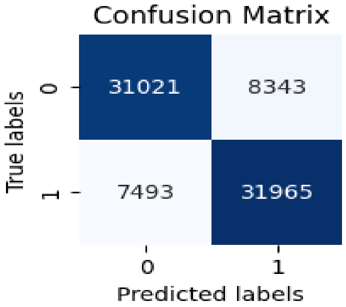


Figure 11 Confusion matrix for decision tree

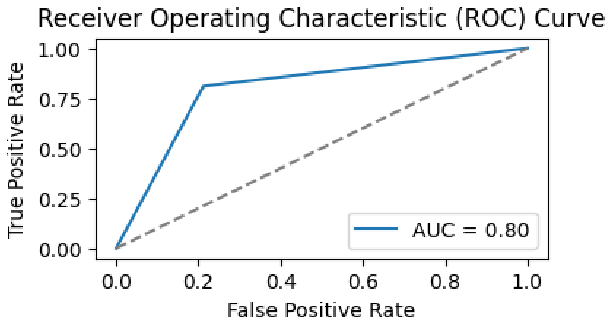


Figure 13 ROC for decision tree

Figure 13 shows the confusion matrix results for the random forest algorithm. Based on the analysis, the True Negative (TN) value is 36969, the False Positive (FP) value is 2395, the False Negative (FN) value is 8982, and the True Positive (TP) value is 30476. In Figure 14, the Receiver Operating Characteristic area (ROC) value for random forest algorithm is presented as 0.88.

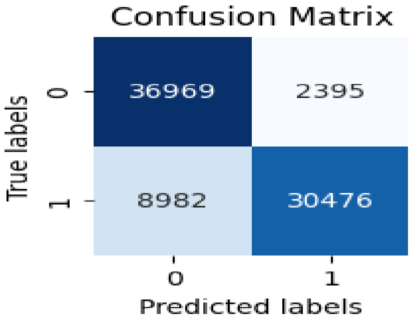


Figure 13 confusion matrix of random forest

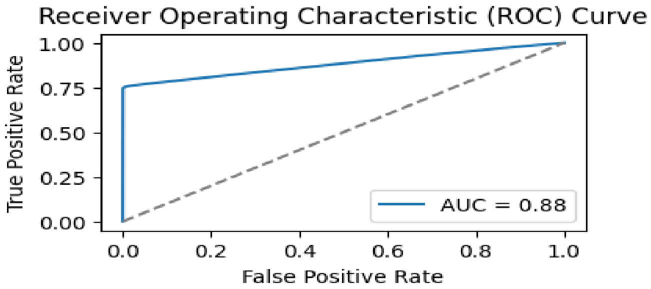


Figure 14 ROC for random forest

In Figure 15, the confusion matrix results for the XG-Boost algorithm are displayed. According to the analysis, the True Negative (TN) value is 39359, the False Positive (FP) value is 5, the False Negative (FN) value is 9730, and the True Positive (TP) value is 29728. Additionally, Figure 16 illustrates the Receiver Operating Characteristic area (ROC) value for XG-Boost, which stands at 0.88.

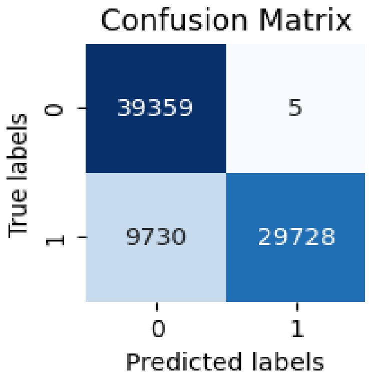


Figure 15 confusion matrix for XG-Boost

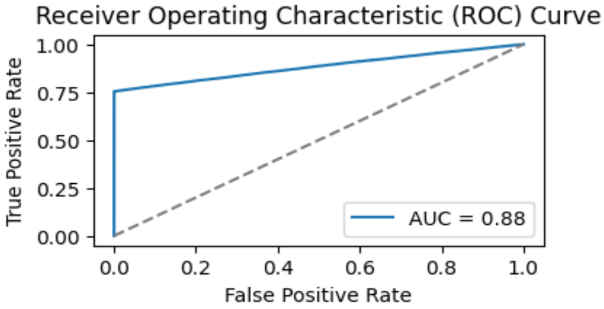


Figure 16 ROC for XG-Boost

Based on table 3, XG-Boost topped with the highest accuracy rate at 87.64%, followed by Random Forest at 85.57%, and Decision Tree at 79.91%. XG-Boost also achieved the highest precision, reaching 99.97%, while Random Forest followed with 92.71%, and Decision Tree with 79.3%. In terms of recall, Decision Tree achieved the highest score of 81.01%, followed by Random Forest at 77.24%, and XG-Boost at 75.33%. XG-Boost achieved in terms of f1-score 85.92%, followed by Random Forest at 84.27%, and Decision Tree at 80.15%. XG-Boost and Random Forest both achieved an equal Receiver Operating Characteristic (ROC) of 88%, while Decision Tree scored 80%.

Table 3 performance metrices of each algorithm

Algorithms	Accuracy	Precision	Recall	F1-Score	Receiver Operating Characteristic
Decision tree	79.91%	79.3%	81.01%	80.15%	80%
Random forest	85.57%	92.71%	77.24%	84.27%	88%
XG-Boost	87.64%	99.97%	75.33%	85.92%	88%

5-Rush time prediction

This paper aims to achieve customer retention by incorporating both customer churn prediction and the prediction of rush times for purchases. This part mainly focuses on the second task the prediction of rush times for purchases. Following the same pre-processing steps used in the earlier task of customer churn prediction. The purchase time feature is broken down into three columns: hour, minute, and second. This separation is crucial for a focused analysis of the 'hour' column, given its central role in this task. After this segmentation, we visualize the data and notice a normal distribution pattern as shown in figure 17.

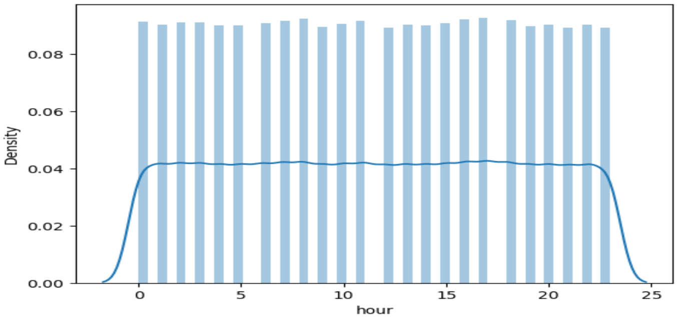
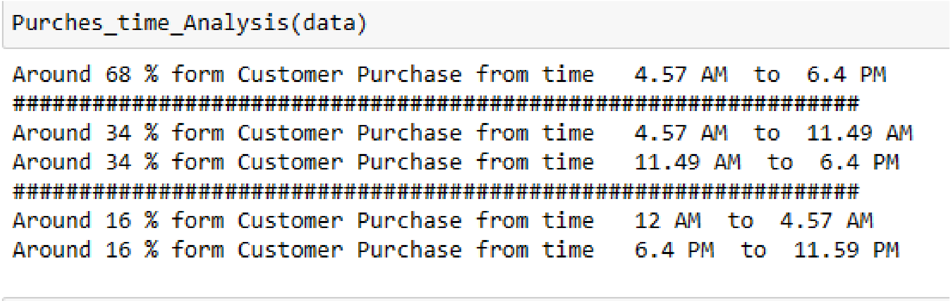


Figure 17 Hour distribution

Based on this observation, we opt for a statistical model designed to work well with such distributions. Finally, we introduce the z-score to describe how a particular value relates to the mean within a group of values. Results indicated that rush times between 4:57 AM and 6:04 PM accounted for approximately 68% of all purchases as shown in figure 18.



```
Purches_time_Analysis(data)

Around 68 % form Customer Purchase from time   4.57 AM  to  6.4 PM
################################################################
Around 34 % form Customer Purchase from time   4.57 AM  to  11.49 AM
Around 34 % form Customer Purchase from time   11.49 AM  to  6.4 PM
################################################################
Around 16 % form Customer Purchase from time   12 AM  to  4.57 AM
Around 16 % form Customer Purchase from time   6.4 PM  to  11.59 PM
```

Figure 18 Purchase time analysis

6- Conclusion

The significance of customer churn and rush time prediction in e-commerce is highlighted as businesses transition to online platforms with a high competition. Early detection of churn and rush times is vital for comprehensive organizational support and achieving customer retention, particularly with this high competition. Businesses should create novel marketing strategies and strategically allocate resources during peak times to facilitate rapid responses to customer needs, thereby enhancing satisfaction and avoiding technical issues, ultimately leading to improved customer retention.

This research successfully achieved its objectives by analyzing the dataset from an E-commerce platform. In the initial task, focusing on predicting customer churn, three classification algorithms—Decision Tree, Random Forest, and XG-Boost—were employed and analyzed. Through comparison of the results, it was found that the XG-Boost algorithm outperformed the others, with an accuracy score of 87.64%, precision of 99.98%, recall of 75.33%, f1-score of 85.92%.and Receiver Operating Characteristic (ROC) of 88%.

These results underscore the complexity and non-linear nature of the factors influencing churn in the E-commerce context. They emphasize the need for advanced and powerful techniques capable of capturing these complexities and also it was concluded that the impact of customer segmentation on churn prediction depends on the dataset and the machine learning model selected. From this, we can infer that powerful algorithms may reduce the need for customer segmentation if they can effectively capture patterns and relationships within the data without relying on segmented customer groups.

In the second task, focusing on predicting rush times, z-score analysis was applied to identify peak purchasing periods. The findings indicated that the timeframe from 4:57 AM to 6:04 PM accounted for approximately 68% of all purchases.

References

[1] Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(2), 459-475.

[2] Bogaert, M., & Delaere, L. (2023). Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art. *Mathematics*, 11, 1137.

[3] Mirkovic, M., et al. (2022). Customer Churn Prediction in B2B Non-Contractual Business Settings Using Invoice Data. *Applied Sciences*, 12(12), 5001.

[4] Pondel, M. et al. (2021). Deep learning for customer churn prediction in ecommerce decision support. In Proceedings of the 24th International Conference on Business Information Systems (BIS 2021).

[5] Alghanam, O. A., et al. (2022). Data Mining Model for Predicting Customer Purchase Behavior in e-Commerce Context. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 13(2), 421-428.

[6] Tang, H. Y., & Suraya, Y. (2023). E-Commerce Customer Churn Prediction for the Marketplace in Malaysia. *Open International Journal of Informatics (OIJI)*, 11(2), 58-66.

[7] Alghazzawi, D. M., et al. (2023). ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review. Sustainability, 15, 7076

[8] Loukili, M., et al. (2023). Machine learning-based recommender system for e-commerce. *International Journal of Artificial Intelligence*, 12(4), 1803-1811

[9] Abas Sunarya, P., et al. (2024). Deciphering Digital Social Dynamics: A Comparative Study of Logistic Regression and Random Forest in Predicting E-Commerce Customer Behavior. Journal of Applied Data Sciences, 5(1), 100-113.

[10] Zhang, X., et al. (2023). A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research. *Journal of Theoretical and Applied Electronic Commerce Research*, 18, 2188–2216.

[11] Rashi, D., et al. (2024). An AI-Based Customer Relationship Management Framework for Business Applications. *International Journal*

of Intelligent Systems and Applications in Engineering (IJISAE), 12(12s), 686–695.

[12] Sharma, M. K., et al. (2023). Machine Learning Based Customer Churn Prediction Using Improved Feature Selection Techniques. *International Research Journal of Engineering & Applied Sciences*, 11(4), 26-36

[13] Tran, H., et al. (2023). Customer churn prediction in the banking sector using machine learning-based classification models. *Interdisciplinary Journal of Information, Knowledge, and Management*, 18, 87-105.

[14] Loukili, M., et al. (2022). Supervised Learning Algorithms for Predicting Customer Churn with Hyperparameter Optimization. *International Journal of Advance Soft Computer Application*, 14(3), 50-63.

[15] Al-Shatnwai, A. M., & Faris, M. (2020). Predicting customer retention using XGBoost and balancing methods. *International Journal of Advanced Computer Science and Applications*, 11(7), 704-712.

[16] Lalwani, P., Mishra, M. K., Chadha, J. S., et al. (2022). Customer churn prediction system: a machine learning approach. *Computing*, 104, 271–294.

[17] Calli, L., & Kasim, S. (2022). Using Machine Learning Algorithms to Analyze Customer Churn in Software as a Service (SaaS) Industry. *Academic Platform Journal of Engineering and Smart Systems*, 10(3), 115-123.

[18] Shetu, S. F., et al. (2021). Predicting Satisfaction of Online Banking System in Bangladesh by Machine Learning. In International Conference on Artificial Intelligence and Computer Science Technology (ICAICST-2021).

[19] Hamdan, I. Z. P., & Othman, M. (2022). Predicting customer loyalty using machine learning for the hotel industry. *Journal of Soft Computing and Data Mining*, 3(2), 31-42.

[20] Sun, Y., et al. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Journal of Heliyon*, 9(2), e13384.

[21] Necula, S. C. (2023). Exploring the impact of time spent reading product information on E-commerce websites: A machine learning approach to analyze consumer behavior. *Behavioral Sciences*, 13, 439.

[22] Xiahou, X., & Harada, Y. (2022). Customer Churn Prediction Using Ada Boost Classifier and BP Neural Network Techniques in the E-commerce Industry. *American Journal of Industrial and Business Management*, 12, 277-293.

[23] Wong, & Booma Poolan Marikannan. (2020). Optimizing E-commerce Customer Satisfaction with Machine Learning. *Journal of Physics: Conference Series, 1712, 012044*.

[24] Wu, Z., Jing, L., Wu, B. *et al* (2022*).* A PCA-AdaBoost model for E-commerce customer churn prediction. *Annals of Operations Research*

[25] Ullah, I., et al. (2019). A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE Access*, 7, 60134-60149

[26] Routray, S. K. (2021). Marketing strategy through machine learning techniques: A case study at Telecom Industry. *International Journal of Innovation Engineering and Science Research*, 5(3), 21-30.

[27] Kimura, T. (2022). Customer churn prediction with hybrid resampling and ensemble learning. *Journal of Management Information and Decision Sciences*, 25(1), 1-23.

[28] Patel, N., Trivedi, S.(2020). Leveraging Predictive Modeling, Machine Learning Personalization And NLP Customer Support and Chatbots to Increase Customer Loyalty.*empirical quests for management essences*,3(3),1-24

[29] Abbassy, M. M. (2023). Using Machine Learning Technique for Analytical Customer Loyalty. *International Journal of Computer Science and Network Security (IJCSNS)*, 23(8), 190-198.

[30] Song, P., & Liu, Y. (2020). An XG Boost Algorithm for predicting purchasing behavior on E-commerce platforms. *Technical Gazette*, 27(5), 1467-1471.

[31] Latheef, J., & Vineetha, S. (2021). Exploring Data Visualization to Analyze and Predict Customer Loyalty in Banking Sector with Ensemble Learning. *International Journal of Innovative Research in Applied Sciences and Engineering,* 4(9), 891-904

[32] Wahul, R. M., Kale, A. P., & Kota, P. N. (2023). An Ensemble Learning Approach to Enhance Customer Churn Prediction in Telecom Industry. *International Journal of Intelligent Systems and Applications in Engineering (IJISAE),* 11(9s), 258–266.

[33] Zhou, Y., et al. (2023). Early warning of telecom enterprise customer churn based on ensemble learning. *PLoS ONE*, 18(10), e0292466.

[34] Ayyapureddi, S. S. R., & Manhar, A. (2023). E-Business Churn Prediction Model Using Machine Learning. *International Journal of Scientific Research in Computer Science, Engineering, and Information Technology*, 9(4), 15-23.

[35] AlKhayrat, M., et al. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, 7(9), 2-23.

[36] Zhang, et al. (2022). Research and Application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, 18(6), 2-10.

[37] Liao,j., et al. (2022). Multi-behavior RFM model based on improved SOM neural network algorithm for customer segmentation. *IEEE Access*, 10, 122501-122512

[38] Basha, A. M. M., et al. (2020). Machine Learning – Structural Equation Modeling Algorithms: The Moderating Role of Loyalty on Customer Retention towards Online Shopping. *International Journal of Emerging Trends in Engineering Research*, 8(5), 1578-1585.

[39] Dong, Y., et al. (2022). Integrated machine learning approaches for e-commerce customer behavior prediction. *Advances in Economics, Business and Management Research*, 211, 1008-1015.

[40] Wu, S., et al. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. IEEE ACCESS,9,62118-6213.