Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 4 (2026) 971 – 983 e-ISSN: 2580-0760

Enhancing E-Commerce Churn Prediction Accuracy Through the Combination of Gradient Boosting and Feature Selection

Yogi Prasetyo Hernoto[1], Dandy Pramana Hostiadi[2], Putu Desiana Wulaning Ayu[3]
[1]Magister Program, Department of Magister Information System, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia
[2,3]Department of Magister Information System, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia
*232011022@stikom-bali.ac.id

Abstract

As the number of e-commerce users continues to expand, analyzing user activity has become essential for improving service quality, since customers are a key factor in strengthening competitiveness and business performance in this industry. Churn, in which customers discontinue service due to dissatisfaction, requires accurate detection to support effective retention strategies. To improve churn prediction accuracy, this study explores the optimal combination of feature selection methods and machine learning classifiers. ANOVA F-Test, Information Gain, Chi-Square, Recursive Feature Elimination (RFE), Variance Threshold, and Least Absolute Shrinkage and Selection Operator (LASSO) are the six feature selection techniques that are assessed in this study. These are combined with three classification algorithms: Gradient Boosting (GB), Random Forest (RF), and Logistic Regression (LR). Gradient Boosting with either Information Gain or LASSO yields the best results, achieving approximately 98% accuracy, 97% precision, 92% recall, 96% F1-score, and 99% ROC-AUC, based on testing on a publicly available Kaggle dataset. This approach outperforms prior studies, providing a robust framework for e-commerce developers to anticipate customer loss. By identifying key predictors of churn, the proposed model offers actionable insights to enhance service and enable proactive customer management.

Keywords: churn; e-commerce; gradient boosting; logistic regression; random forest

How to Cite: Y. P. Hernoto, Dandy Pramana Hostiadi, and Putu Desiana Wulaning Ayu, "Enhancing E-Commerce Churn Prediction Accuracy Through the Combination of Gradient Boosting and Feature Selection ", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 4, pp. 971 -, Jul. 2026.
Permalink/DOI: https://doi.org/10.29207/resti.v10i4.7076

Received: August 13, 2025
Accepted: April 4, 2026
Available Online: July 31, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The e-commerce industry continues to expand rapidly, prompting companies to focus on customer retention to stay competitive [1]. Maintaining customer loyalty is crucial, as it significantly enhances competitiveness and company performance [2]. Churn refers to users who discontinue a service, often because of dissatisfaction with product quality, delivery speed, or support. Identifying early signs of churn enables proactive intervention, such as personalized offers or service upgrades [3]. Churn analysis helps businesses develop strategies to retain customers amid intense competition. Given that global e-commerce retail sales are projected to reach $7.39 trillion by 2025, services such as fast delivery and after-sales support are vital to ensuring customer satisfaction [4].

Previous research has introduced churn detection models using machine learning approaches. Logistic Regression is a statistical technique used to model the probability of an event occurring by fitting a logistic curve to the data. It is particularly appropriate for two possible states, churn or not churn, based on categorical or continuous predictor variables [5]. Applied to churn analysis, it estimates a customer's likelihood of leaving based on factors such as satisfaction level and subscription length.

Random Forest Regression is widely recognized as an ensemble learning method that builds numerous decision trees on randomly selected subsets of the training data and combines their outputs, thereby improving accuracy and limiting overfitting [6-8]. Gradient Boosting Classifier builds predictive models incrementally, iteratively minimizing the residual errors

of prior estimators. In churn analysis, this model can provide more accurate predictions through iterative learning [9], [10].

High churn rates can significantly impact business continuity, especially in e-commerce industries that depend on digital customer engagement [11-14]. A high churn rate indicates poor service and potentially large financial losses. Therefore, proper churn analysis is needed to support optimal service development. Customer churn can result from various factors, such as misaligned pricing relative to service quality, service inconveniences, core service failures, poor complaint handling, competitive pressure, and ethical issues. Recent empirical research finds that churn is usually the outcome of a series of negative experiences, declining satisfaction, mounting dissatisfaction, and repeated service shortfalls rather than any single incident [15].

This study aims to explore user behaviors that impact churn on e-commerce platforms. It presents a novel method for churn prediction that combines feature selection optimization with classification-based machine learning. Achieving optimal churn identification requires improving performance across assessment metrics, including precision, recall, ROC-AUC, and classification accuracy. This study offers a fresh perspective by exploring the optimal combination of feature selection techniques and machine learning models to maximize churn prediction accuracy. It evaluates three classification algorithms Gradient Boosting Classifier, Random Forest Classifier, and Logistic Regression paired with six feature selection methods: ANOVA F-Test, Information Gain, Chi-Square, Recursive Feature Elimination (RFE), Variance Threshold, and LASSO Regression. Through this data-driven approach, the study aims not only to enhance technical precision but also to provide comprehensive insights into the dynamics of consumer behavior and their implications for business sustainability.

Additionally, this study closes a gap that has seldom been addressed in earlier research. Previous studies have analyzed model interpretability using SHAP and LIME and documented the good predictive performance of Random Forest and XGBoost using SMOTE, but they did not specifically address feature selection [16]. Without assessing feature selection, a prior study examined the effect of outliers on CatBoost performance [17]. Feature selection was not incorporated into the experimental framework in studies on stacking ensemble models [18]. Thus, this study highlights the significance of feature selection in enhancing predicted accuracy while also expanding the scope of model evaluation.

A study combining K-Means clustering and SVMs for customer segmentation improves churn prediction [19]. The role of behavioral and spatial factors in driving churn was further emphasized [20]. Furthermore, machine learning has demonstrated its broad applicability by successfully capturing intricate churn patterns in industries such as finance and e-commerce [21].

To address data imbalance, previous research developed an intelligent customer churn management system using SMOTE. Various machine learning techniques were employed, including logistic regression, decision trees, Random Forest, SVM, and k-NN [22]. The Random Forest model demonstrated the best performance, achieving an accuracy of 87% on a dataset of 10,000 records with 14 attributes. However, that study focused primarily on comparing classification models and did not investigate the role of feature selection, which has significant potential to further optimize model performance.

Previous comparative studies have examined various tree-based classification models for predicting customer churn in the telecommunications industry. Analysis results indicate that tree-based and ensemble techniques can outperform conventional algorithms such as SVM, KNN, Naïve Bayes, and MLP. These evaluations involved selected tree algorithms, random forests, and hybrid tree models such as SysFor, LMT, FT, and ADTree. Furthermore, the findings demonstrate that the SMOTE method improves model precision and performance in addressing class imbalance [23].

Research on customer churn in the banking industry was conducted using a combination of machine learning and RShiny interactive visualization. The study results reveal that factors such as gender, the number of products used, and account balance significantly influence churn risk. Among the algorithms tested, Random Forest and XGBoost demonstrated the best performance, with Random Forest achieving the highest accuracy at 78.3% [24]. These findings underscore the importance of data-driven retention strategies and recommend using more in-depth features and localized customer segmentation to enhance model precision.

Even though a variety of machine learning models have been used to forecast customer attrition, feature selection continues to significantly impact model performance. Several studies applied only one or a limited set of feature selection techniques, which may not fully capture the most relevant attributes. To determine the best churn prediction model, this study assesses several feature selection techniques alongside several classifiers.

Statistical testing (ANOVA and Chi-Square), information-based measurement (Information Gain), iterative model-based selection (RFE), variance filtering (Variance Threshold), and regularization techniques (LASSO) are just a few of the methods that use distinct mechanisms to assess feature relevance. By selecting the most informative properties, feature selection helps reduce data dimensionality, eliminate unnecessary features, and enhance the performance of classification models [25], [26]. While model-based feature selection methods, such as RFE, have been extensively researched within machine learning

optimization frameworks [27], the idea of information-based feature assessment has its roots in entropy theory [28].

However, most previous studies typically apply only one or two feature selection methods when developing churn prediction models, resulting in limited comparative evaluation across different techniques. In addition, systematic experimentation combining multiple feature selection methods with several classification algorithms remains limited, particularly for e-commerce churn datasets. To determine the best and most reliable model for customer churn prediction, this study thoroughly compares a range of feature selection strategies with several classification algorithms [25], [26].

2. Methods

A fundamental component of Customer Relationship Management (CRM) research has long been the prediction of customer churn. To determine the causes of customer attrition, foundational research focused on statistical techniques such as survival analysis and logistic regression. As the e-commerce industry moved toward "Big Data," predictive modeling replaced basic descriptive statistics as the focus. Churn is frequently caused by measurable behavioral indicators, such as transaction frequency, tenure, and customer service encounters, according to foundational studies in this field. By characterizing churn as a binary classification problem, these works established the foundation for the current machine learning methodology.

The modeling stages of this research are illustrated in Figures 1 through 4. Figure 1 presents the big picture of the system architecture, which divides the workflow into four main blocks: data cleaning, feature selection, training, and evaluation. Details regarding data preprocessing are then explained in Figure 2. This step is crucial, as it involves handling missing values, converting categorical variables into numerical formats, and normalizing feature scales to ensure the data fed into the model is truly clean and unbiased.

Moving on to the feature selection stage shown in Figure 3, the focus is on eliminating noise. Here, we test various statistical and algorithmic approaches to filter out variables that merely add computational overhead, retaining only those with strong predictive power.

Once the best features have been selected, the process proceeds to model training and evaluation, as illustrated in Figure 4. In this final phase, several classification algorithms are trained using the cleaned data. Their performance is then assessed using multiple evaluation metrics—aiming not just for high accuracy figures on paper, but to ensure the model is genuinely stable and resistant to overfitting when encountering new data.

2.1 Data Collection and Data Exploration

The data used in this study is a publicly available dataset on the transaction behavior of e-commerce customers. Sourced from Kaggle, the dataset was uploaded by Ankit Verma in 2021 (E-Commerce Customer Churn Analysis and Prediction). Using this open dataset enables model testing on real transaction data that reflect the dynamics of customer churn in the industry.

There are 20 attribute columns and 5,630 rows in the dataset. Of these, 15 are numerical attributes, and 5 are categorical. Data exploration was conducted to understand the dataset's characteristics, including data distribution, emerging patterns, and potential missing values or anomalies that could affect subsequent analysis. As shown in Table 1, the dataset comprises numerical and categorical variables describing customer profiles and behavioral attributes relevant to churn prediction.

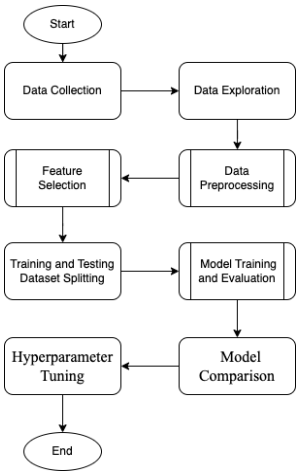


Figure 1. Modeling Pipeline

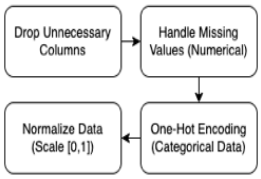


Figure 2. Data Preprocessing Procedure

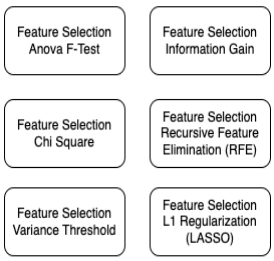


Figure 3. Feature Selection

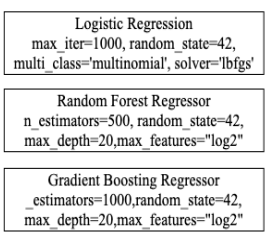


Figure 4. Model Training and Evaluation

Table 1. Description of dataset Variables

Features	Value	Description
CustomerID	50001… 55630	Unique customer identification (Numeric)
Churn	0 and 1	Target variable in this study (Binary)
Tenure	0 … 61	Duration of customer membership (Numeric) Contains 264 missing values
Preferred LoginDevice	Mobile Phone, Computer, Phone	Device most frequently used by the customer to log in (Categorical)
CityTier	1…3	The classification of the city of residence is used to map the distribution of customers by region. (Numeric)
WarehouseToHome	5…127	This variable measures the travel distance between the shipping warehouse location and the customer's residential area (Numeric) Contains 251 missing values
PreferredPayment Mode	Debit Card, Credit Card, E wallet, UPI, COD CC, Cash on Delivery	Primary payment method (Categorical)
Gender	Female, Male	Customer's gender (Categorical)
HourSpendOnApp	0 … 5	Average time (in hours) spent by the customer on the app (Numeric) Contains 255 missing values
NumberOfDevice Registered	1…6	This variable records the total number of devices linked to customers' e-commerce accounts. (Numeric)
PreferedOrderCat	Laptop & Accessories, Mobile Phone, Fashion, Mobile Grocery, Others	Customer preference patterns are measured based on the product categories with the highest order frequency over the past 30 days. (Categorical)
SatisfactionScore	1 … 5	The service's customer satisfaction rating (Numeric)
MaritalStatus	Married, Single, Divorced	Customer's marital status (Categorical)
NumberOfAddress	1 … 22	The quantity of addresses associated with the customer's name (Numeric)
Complain	0 … 1	Complaint indicator submitted in the last month (Binary)
OrderAmountHike FromlastYear	11… 26	Represents the numeric percentage growth in order volume relative to the prior year. (Numeric) Contains 265 missing values
CouponUsed	0 … 16	The quantity of vouchers that the client used throughout the previous month (Numeric)
OrderCount	1 … 16	Total orders placed by customers in a given month (Numeric)
DaySinceLastOrder	0 ... 46	Days that have passed since the client's last purchase (Numeric) Contains 307 missing values
CashbackAmount	0,0 … 324,99	The customer's average cashback amount throughout the previous month (Numeric)

2.2 Preprocessing Data

Setting the target variable (Churn) and using the other attributes as independent variables was the first step in the preprocessing stage. Since the CustomerID column solely functions as a unique identifier, it was eliminated. The features Tenure, WarehouseToHome, HourSpendOnApp, DaySinceLastOrder, and OrderAmountHikeFromLastYear all had missing values.

A K-Nearest Neighbors algorithm k=5 was used to impute missing numerical entries. Recent evaluations have shown that KNN imputation maintains distributional properties and yields lower error rates in tabular datasets, outperforming mean substitution in preserving the underlying data distribution [29], [15]. One-hot encoding of categorical variables converts text categories into binary vectors (0 and 1), which remains one of the best ways to preserve category information [30]. Normalization of numerical features is performed using min-max scaling to ensure all variables fall within a uniform range of [0, 1]. In addition to preventing large-scale features from dominating the calculations, this technique has been shown to make the model training process more stable and accelerate convergence [31].

2.3 Feature Selection

Feature selection is an essential preprocessing step for high-dimensional datasets, such as those encountered in e-commerce, to remove unnecessary or redundant variables. Filter, wrapper, and embedded methods are the three general categories into which methodologies are divided. Because they assess features based on their inherent statistical characteristics, irrespective of any classifier, filter methods (such as information gain, chi-square, and ANOVA) are computationally efficient. Wrapper techniques, such as recursive feature elimination, assess feature subsets by utilizing a particular algorithm as a black box to identify the optimal combination. Lastly, feature selection is performed during model training using embedded techniques such as LASSO. Because redundant features can introduce noise and lead to "the curse of dimensionality," which adversely affects model generalization, choosing the right methodology is crucial. Through feature selection, we identify the most significant data characteristics for predicting potential customer churn. This approach aims to enhance prediction accuracy while maintaining the efficiency of the modeling process. By avoiding needless complexity that could lead to overfitting, this phase improves model performance and accelerates training. To increase the model's effectiveness, precision, and interpretability, only crucial features are kept [32]. ANOVA F-Test,

Information Gain, Chi-Square, Recursive Feature Elimination (RFE), Variance Threshold, and LASSO Regression were the six feature selection techniques used. Whether based on linear correlations, information entropy, category significance, or coefficients from an embedded model, each approach offers a unique viewpoint on feature relevance. A statistical technique called the ANOVA F-Test assesses the linear dependency between the binary target and numerical features. Higher F-score features exhibit a stronger correlation with the target's variability [33]. Because it does not employ a prediction model and assesses attributes independently of the target variable, this approach is classified as a filter method. Equation 1 illustrates how the ten features with the greatest F-scores are chosen for modeling.

$$MS = \frac{sS}{dF} \quad (1)$$

The variable descriptions in Equation 1 include MS for Mean Square, sS for Total Sum of Squares, and dF, which indicates the degrees of freedom in the statistical calculation. Information Gain quantifies the reduction in entropy attributable to each feature and is especially effective for selecting informative, nonlinearly related features, whereas ANOVA only captures linear associations [34]. It is calculated as shown in Equation 2.

$$IG(D,A) = Entropy(D)\Sigma_{v \in Values(A)} \frac{|D_v|}{|D|}.Entropy(D_v) \quad (2)$$

In Equation 2, D = Full Dataset; A = Attribute/feature under evaluation; Values(A) = All unique values of feature A; D_v= Subset of D where A = v; $|D_v|$ = Number of instances in D_v; $|D|$ = Total number of instances; Entropy(D) = Entropy of dataset D; Entropy (D_v) = Entropy of subset D_v. Entropy (D) itself is computed using Equation 3.

$$Entropy(D) = \Sigma_{i=1}^{c} p_i log_2 p_i \quad (3)$$

The variable definitions in Equation 3 include C (number of classes) and p_i (proportion of the i-th class in D). The Chi-Square test is used to select categorical features; the higher the statistical value, the stronger the association with the churn label [35]. The calculation of this Chi-Square statistic is described in Equation 4, where the expected frequency value is derived using Equation 5.

$$x^2 = \Sigma_{i=1}^{r}\Sigma_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (4)$$

$$E_{ij} = \frac{(Row\ Total\ i)\ x\ (Column\ Total\ j)}{Grand\ Total} \quad (5)$$

The variables r and c in Equations 4 and 5 represent the number of rows and columns of categories, while O_{ij} and E_{ij} measure the observed and expected frequency values for cell-i, j. Unlike previous statistical approaches, we also incorporate Recursive Feature Elimination (RFE) a wrapper technique that relies on logistic regression as the estimator to assess the significance of each feature in predicting churn. Technically, RFE eliminates features deemed to contribute less one by one based on their importance ranking. From the total set of variables, we limited the result to the top k=10 features to ensure the model remains efficient without compromising predictive performance. The value of k=10 was determined based on a sensitivity analysis across various parameter ranges. Test results indicate that this value offers the best compromise between model prediction accuracy and computational efficiency. This specific value was chosen to maintain a parsimonious model that reduces computational complexity while retaining the most critical predictors of customer churn. Experimental results indicated that selecting the top 10 features provided an optimal balance, as including additional features led to diminishing returns in accuracy and increased the risk of overfitting [36]. The recursive elimination procedure used for RFE in this study proceeds as follows. The process begins with the complete feature set, as shown in Equation 6.

$$F_{ij} = \{X_1, X_2, \dots, X_p\} \quad (6)$$

In Equation 6, F_0 denotes the initial set of features, while X_1 through X_p denote all p available predictor variables. In each iteration t, the model f_t is trained using the available subset of features, with the primary focus on minimizing the loss function as formulated in Equation 7.

$$F_t = argmin\ L\ (f(X_{F_t}), y)\ f \quad (7)$$

In Equation 7, f_t denotes the model resulting from training at iteration t; L measures the loss function value, while X_F_t and y represent the input features in subset F_t and the target variable, respectively. Based on the results of this iteration, the importance score of each remaining feature is then evaluated using Equation 8.

$$imp_j = |X_p| \text{ (for linear model)} \quad (8)$$

In Equation 8, imp_j denotes the importance score assigned to feature j, obtained from the magnitude of its coefficient in the fitted linear model.
The least important feature in the current subset is then identified using Equation 9.

$$j^* = argmin\ imp_j\ j \in F_t \quad (9)$$

Equation 9 defines j as the index of the feature with the lowest importance score in F_t. The feature with this minimum value is then removed to form the updated subset for the next iteration (Equation 10).

$$F_{t+1} = F_t \backslash \{X_j\} \quad (10)$$

In Equation 10, F_{t+1} denotes the resulting feature subset after removing the identified feature X_j from F_t. This iterative process continues until the number of remaining features equals the target value k, as expressed in Equation 11.

$$|F_t| = k \quad (11)$$

In Equation 11, $|F_t|$ denotes the number of features remaining in the subset at iteration t, and k is the

predetermined number of features to retain. The final selected feature subset is then obtained as shown in Equation 12.

$$F_{selected} = F_k \quad (12)$$

In Equation 12, $F_{selected}$ denotes the final set of selected features, which equals F_k, the feature subset remaining after the k-th iteration.

Variance threshold is an unsupervised technique that evaluates features based on their variance. Features with near-zero variance are removed, as they provide little information [37]. The top ten features with the highest variance are retained, as calculated in Equation 13.

$$Var(X_j) = \frac{1}{n}\Sigma_{i=1}^{n}(x_{ij} - \underline{x_j})^2 \quad (13)$$

In Equation 13, a feature is removed if it has $Var(X_j)<\theta$; $X_j = j$ feature in dataset; x_{ij} = Value of feature in j at sample i; $\underline{x_j}$= Mean of feature X_j: $\underline{x_j} = \frac{1}{n}\Sigma_{i=1}^{n}x_{ij}$; x = Number of samples; $Var(X_j)$ = Variance of feature X_j; θ = Threshold (ex: 0.01).

LASSO Regression is an embedded technique that uses LassoCV to automatically choose the best regularization parameter through cross-validation. LASSO reduces the coefficients of less significant features to zero by adding an L1 penalty to the loss function. According to Equation 14, the top features with the highest absolute coefficients are kept [38].

$$\beta^{\wedge} = \beta arg\ min(\frac{1}{2n}\Sigma_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\Sigma_{j=1}^{p}|\beta_j|) \quad (14)$$

In Equation 14, y: target variable; x_j: feature in j, the coefficient of feature j, the regularization parameter, and the L1-norm to promote sparsity. The final output of feature selection is the top 10 features, which are used for machine learning modeling. This approach reflects the fundamental trade-off between model effectiveness and interpretability. Tables 2, 3, and 4 present the rankings of key features by significance level, ordered from highest to lowest. This ranking is derived from the evaluation of six feature selection methods: Information Gain, Chi-Square, RFE, LASSO, ANOVA F-test, and Variance Threshold. By combining various feature selection approaches, this study presents a comprehensive overview of the most significant attributes for predicting customer churn, examined from both statistical and algorithmic perspectives. It also offers a more thorough understanding of each feature's contribution in terms of linearity, non-linearity, and categorical significance. Additionally, this strategy fills a methodological void that hasn't been thoroughly examined in earlier research. By eliminating extraneous variables, a step that much predictive modeling research avoids, this method improves model efficiency and performance. To enhance model accuracy, recent research trends largely focus on integrating feature selection techniques with ensemble algorithms. Various prior studies have explored strategic combinations to boost detection rates across diverse industries—such as pairing Random Forest with RFE or Support Vector Machines (SVM) with Chi-Square. For example, a number of studies have shown that although individual classifiers are strong, information-theoretic selection techniques like Information Gain greatly enhance their effectiveness.

Table 2. The top features ranked from highest to lowest importance for Information Gain and Chi-Square

No	Information Gain	Chi-Square
1	Tenure	Complaint
2	CashbackAmount	MaritalStatus_Single
3	Complaint	PreferedOrderCat_Mobile Phone
4	DaySinceLastOrder	Tenure
5	MaritalStatus_Married	PreferedOrderCat_Laptop & Accessory
6	PreferedOrderCat_Mobile Phone	PreferedOrderCat_Mobile
7	OrderCount	MaritalStatus_Married
8	WarehouseToHome	PreferedOrderCat_Grocery
9	SatisfactionScore	PreferredPaymentMode_COD
10	CouponUsed	PreferredLoginDevice_Mobile Phone
11	PreferedOrderCat_Mobile	PreferredLoginDevice_Phone
12	NumberOfDeviceRegistered	CityTier
13	MaritalStatus_Single	PreferredPaymentMode_Credit Card
14	PreferedOrderCat_Grocery	PreferedOrderCat_Others
15	PreferredPaymentMode_Cash on Delivery	PreferredPaymentMode_E wallet
16	PreferedOrderCat_Laptop & Accessory	SatisfactionScore
17	PreferredLoginDevice_Mobile Phone	DaySinceLastOrder
18	Gender_Male	CashbackAmount
19	NumberOfAddress	NumberOfDeviceRegistered
20	PreferedOrderCat_Others	PreferredPaymentMode_Debit Card
21	PreferredPaymentMode_E wallet	OrderCount
22	PreferredPaymentMode_COD	Gender_Male
23	PreferredPaymentMode_Credit Card	WarehouseToHome
24	PreferredPaymentMode_Debit Card	NumberOfAddress
25	PreferredPaymentMode_UPI	PreferredPaymentMode_Cash on Deliver
26	HourSpendOnApp	CouponUsed
27	CityTier	PreferredPaymentMode_UPI
28	OrderAmountHikeFromlastYear	OrderAmountHikeFromlastYear
29	PreferredLoginDevice_Phone	HourSpendOnApp

Our study builds upon this trend by investigating the specific combination of information gain and LASSO with the Gradient Boosting Classifier (GBC). While prior work often used simpler linear models, this research leverages GBC's non-linear capabilities to capture more complex customer behaviors, thereby filling a gap in the literature on optimal combinations for high-variance e-commerce datasets.

After completing preprocessing and feature selection, the dataset is split into two subsets: 80% for training and 20% for testing. This separation is performed to preserve the reliability of the test data when evaluating the model's generalization capability, while simultaneously ensuring that the training process utilizes a sufficiently large portion of the data.

Table 3. The top features ranked from highest to lowest importance RFE, Lasso

No	RFE	Lasso
1	Tenure	Tenure
2	CityTier	NumberOfAddress
3	WarehouseToHome	CashbackAmount
4	HourSpendOnApp	WarehouseToHome
5	NumberOfDeviceRegistered	DaySinceLastOrder
6	SatisfactionScore	Complaint
7	NumberOfAddress	NumberOfDeviceRegistered
8	Complaint	PreferedOrderCat_Others
9	OrderAmountHikeFromlastYear	PreferedOrderCat_Laptop & Accessory
10	CouponUsed	PreferredPaymentMode_COD
11	OrderCount	SatisfactionScore
12	DaySinceLastOrder	OrderCount
13	CashbackAmount	PreferredPaymentMode_E wallet
14	PreferredLoginDevice_Mobile Phone	PreferredPaymentMode_Cash on Delivery
15	PreferredLoginDevice_Phone	MaritalStatus_Single
16	PreferredPaymentMode_COD	PreferedOrderCat_Grocery
17	PreferredPaymentMode_Cash on Delivery	CityTier
18	PreferredPaymentMode_Credit Card	CouponUsed
19	PreferredPaymentMode_Debit Card	PreferredPaymentMode_Debit Card
20	PreferredPaymentMode_E wallet	PreferredLoginDevice_Phone
21	PreferredPaymentMode_UPI	PreferredLoginDevice_Mobile Phone
22	Gender_Male	PreferredPaymentMode_UPI
23	PreferedOrderCat_Grocery	PreferedOrderCat_Mobile Phone
24	PreferedOrderCat_Laptop & Accessory	PreferredPaymentMode_Credit Card
25	PreferedOrderCat_Mobile	HourSpendOnApp
26	PreferedOrderCat_Mobile Phone	Gender_Male
27	PreferedOrderCat_Others	MaritalStatus_Married
28	MaritalStatus_Married	OrderAmountHikeFromlastYear
29	MaritalStatus_Single	PreferedOrderCat_Mobile

Table 4. The top features ranked from highest to lowest importance: ANOVA F-TEST, Variance Threshold

No	ANOVA F-Test	Variance Threshold
1	Tenure	PreferredLoginDevice_Mobile Phone
2	Complaint	MaritalStatus_Married
3	MaritalStatus_Single	PreferredPaymentMode_Debit Card
4	PreferedOrderCat_Mobile Phone	Gender_Male
5	CashbackAmount	PreferedOrderCat_Laptop & Accessory
6	MaritalStatus_Married	MaritalStatus_Single
7	DaySinceLastOrder	CityTier
8	PreferedOrderCat_Laptop & Accessory	Complaint
9	PreferedOrderCat_Mobile	PreferredPaymentMode_Credit Card
10	PreferredLoginDevice_Mobile Phone	PreferedOrderCat_Mobile Phone
11	NumberOfDeviceRegistered	PreferredLoginDevice_Phone
12	SatisfactionScore	PreferedOrderCat_Mobile
13	PreferedOrderCat_Grocery	SatisfactionScore
14	CityTier	PreferredPaymentMode_E wallet
15	PreferredPaymentMode_COD	PreferredPaymentMode_UPI
16	PreferredLoginDevice_Phone	PreferedOrderCat_Grocery
17	WarehouseToHome	PreferredPaymentMode_COD
18	PreferredPaymentMode_Credit Card	OrderAmountHikeFromlastYear
19	PreferredPaymentMode_E wallet	PreferedOrderCat_Others
20	PreferedOrderCat_Others	NumberOfDeviceRegistered
21	NumberOfAddress	OrderCount
22	OrderCount	PreferredPaymentMode_Cash on Delivery
23	PreferredPaymentMode_Debit Card	CashbackAmount
24	Gender_Male	HourSpendOnApp
25	CouponUsed	Tenure
26	HourSpendOnApp	NumberOfAddress
27	OrderAmountHikeFromlastYear	CouponUsed
28	PreferredPaymentMode_Cash on Delivery	DaySinceLastOrder
29	PreferredPaymentMode_UPI	WarehouseToHome

2.4 Model Training and Evaluation

Three prediction models logistic regression, random forest, and gradient boosting were implemented and validated in the following phase.

Using preprocessed, feature-selected data, the logistic regression model forecasts customer attrition [39]. Equation 15's sigmoid function is used by the logistic regression model to determine the likelihood of client attrition.

$$P(y = 1|X) = \frac{1}{1+e^{-(\beta_o+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}} \quad (15)$$

In Equation 15, $P(y = 1|X)$ = predicted churn probability; β_o = intercept; $\beta_1 \ldots \beta_n$ = coefficients; $x_1 \ldots x_n$ = predictors; e= base of natural log. To optimally boost the model's performance, several hyperparameters were configured in this study. To ensure that the algorithm converges even with huge or complex datasets, the max_iter parameter was set to 1000. To ensure consistency and reproducibility across multiple runs, the random_state was set to 42. Additionally, the multiclass option was set to 'multinomial', enabling multinomial logistic regression, which provides greater stability for multi-class classification.

Technically, a Random Forest constructs multiple independent decision trees simultaneously. Each tree generates its own prediction, and the final decision is determined by a majority vote or by averaging the predictions. To improve generalization and reduce overfitting, modern stochastic variations randomly subsample data and features at each iteration [40]. Equation 16 expresses the overall prediction generated from the ensemble outputs.

$$\hat{y} = \frac{1}{T}\Sigma_{t=1}^{T} f_t(x) \quad (16)$$

In Equation 16, $f_t(x)$ = prediction from the t-th tree, with T = total number of such trees constructed in the ensemble, typically set by the n-estimators parameter. In this study, the model parameters were carefully selected to enhance performance and control complexity. The n_estimators parameter was set to 500, indicating the number of trees trained in the ensemble.

To guarantee consistency and repeatability between runs, the random_state was set to 42. The max_depth was set to 20 to limit the depth of each decision tree and avoid overfitting. Max_depth = 20 consistently produced the best validation performance, leading to its selection based on empirical experiments assessing several depth values. Additionally, only a subset of features equal to the $\log_2$ of the total number of features was considered at each split because the max_features option was set to "log2". This method helps lower correlation within the ensemble and adds diversity among the trees.

Technically, Gradient Boosting constructs a set of decision trees sequentially. Evaluation is performed iteratively, with the model orienting the training of the subsequent tree to minimize the error residuals left by the preceding tree [41]. The model follows these steps:

Step 1: Equation 17 is used to determine the initial constant value β.

$$F0(x) = arg\ arg\ \beta\ \ \Sigma_{i=1}^{N} L(yi, \beta) \quad (17)$$

In Equation 17, $F0(x)$= initial constant prediction; β= value being optimized; L = loss function; yi = actual value; N = number of samples.

Step 2: Calculate the residual value (loss function gradient) for each iteration m = 1 … M as formulated in Equation 18.

$$y_i - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)-F_{m-1(x)}}, i = \{1,2,\ldots,N\} \quad (18)$$

In Equation 18, the bracketed term is the negative gradient (pseudo residual); yi = actual value; $F(x) = F_{m-1(x)}$

Step 3: The initial model is obtained by fitting the sample data using the basic classifier. The model $(x_I; a)$ is fitted once the model's parameters are determined using the least squares method. with parameters that are tuned to reduce error.

Step 4: The optimal step size (model weight) is computed to minimize the overall loss as shown in Equation 19.

$$\beta_m = \alpha\beta\ arg\ arg\ min\ \Sigma_{i=1}^{N} L(y_I, F_{m-1}(x) + \beta h(x_I; a)) \quad (19)$$

In Equation 19, β_m= optimal step size at iteration m; $h(x_I; a)$ = base learner with parameters a.

Step 5: The model is updated as shown in Equation 20.

$$F_{m(x)} = F_{m-1}(x) + \beta_m h(x_I; a) \quad (20)$$

In Equation 20, $F_{m(x)}$ = updated model after iteration m = $F_{m-1}(x) + \beta_m h(x_I; a)$. To maximize the boosting process and ensure reliable results, the settings were established. The number of boosting trees to be built consecutively was specified by setting the n_estimators parameter to 1000. To guarantee consistent, repeatable outcomes across runs, the random_state was set to 42. To control model complexity and avoid overfitting, max_depth was set to 20, limiting the depth of each decision tree. Max_depth = 20 consistently produced the best validation results while maintaining an ideal balance between model complexity and generalization. This value was established by empirical investigation, in which several depth values were assessed. Applying the `max_features="log2"` configuration limits the number of features considered at each decision node to the base-2 logarithm of the total number of features. This restriction aims to reduce variance and promote better model generalization.

2.5 Model Performance Comparison and Result Interpretation

To find the optimal model for forecasting customer attrition, a performance comparison is conducted

following model evaluation. Using classification metrics such as accuracy, Logistic Regression is compared with Random Forest and Gradient Boosting classifiers. To ensure that the chosen model is not only accurate but also dependable and consistent in its performance, the model's stability and ability to generalize to unseen data are also taken into account.

Following churn prediction, further analysis is conducted to identify the key factors driving customers' discontinuation of service use. The most important characteristics found by the Gradient Boosting Classifier are used in conjunction with the Information Gain feature selection approach to further evaluate the causes of churn. The analysis results are visualized using feature-importance graphs that illustrate churn behavior patterns as interpreted by the machine learning models.

Table 5. Best Evaluation Result

Experiment	Model	Feature Selection	Accuracy	Precision	Recall	F1 Score	ROC
1	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
2	Gradient Boosting	Information Gain	0.987567	1	0.926316	0.961749	0.999567
3	Gradient Boosting	Information Gain	0.986679	1	0.921053	0.958904	0.999561
4	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
5	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
6	Gradient Boosting	Information Gain	0.987567	1	0.926316	0.961749	0.999635
7	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
8	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
9	Gradient Boosting	Information Gain	0.987567	0.994382	0.931579	0.961957	0.999623
10	Gradient Boosting	Information Gain	0.986679	1	0.921053	0.958904	0.999786

Based on Table 5, the Gradient Boosting model demonstrated stable performance across all ten test scenarios. The best results were achieved in the ninth experiment using Information Gain as the feature selection method, yielding an accuracy of 0.9876, a recall of 0.9316, and an F1-score of 0.9619. The superiority of Information Gain was evident compared with LASSO, particularly in sensitivity (recall) and classification balance. This confirms that Information Gain is the optimal feature selection technique for churn detection in e-commerce.

Figure 5 shows a much more stable, near-ideal pattern. Precision, ROC AUC, and accuracy exceed 0.95 in nearly all scenarios. Recall indicates better identification of potential churners. The F1 score also increases, showing balanced performance. This overall consistency across repeated experiments suggests a well-trained model with minimal bias, likely achieved through combinations such as gradient boosting with information gain or LASSO, which proved most effective in this study.

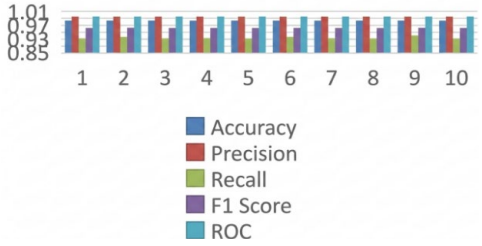


Figure 5. Best Performance Result

Recall consistently yields the lowest results across all tests, as shown in Figure 6, where all evaluation metrics show very low and inconsistent values. This suggests a serious flaw in identifying clients who are likely to leave. The low recall and F1-score values indicate a tendency for the model to be biased toward the majority class (non-churn), even though accuracy and precision remain above 0.70. This limitation in classification performance is also reflected in the ROC AUC scores, which fall within the 0.85 to 0.89 range. Less successful combinations, like logistic regression with the ANOVA F-test, are probably to blame for this performance.

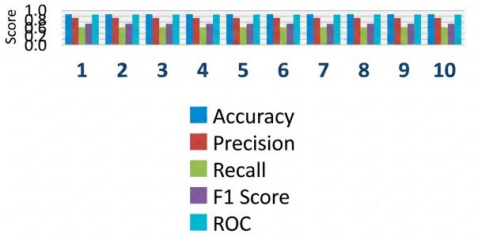


Figure 6. Lowest Performance Result

As summarized in Table 6, all scenarios employing the Logistic Regression algorithm and the ANOVA F-test yielded the lowest yet most stable scores. This model achieved an accuracy of 0.89076, a precision of 0.77686, a recall of 0.494737, and an F1-score of 0.604502. The model is unsuccessful at churn prediction in the e-commerce setting due to its low recall, indicating it struggles to identify clients likely to churn. This implies that using ANOVA and logistic regression together is not the best option for this dataset.

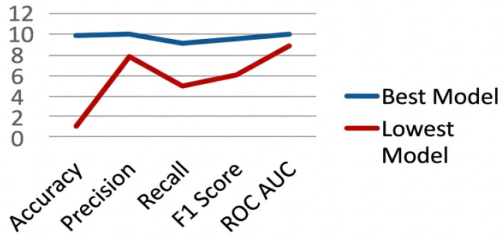


Figure 7. Model Comparison

By utilizing five evaluation metrics—accuracy, precision, recall, F1-score, and ROC AUC—we compare the models with the best and worst performance, as shown in Figure 7. The best model, the

Gradient Boosting Classifier in conjunction with the Information Gain feature selection method, received near-maximum scores across all assessment criteria (approaching 10 on a normalized scale), as shown by the blue line. The model that performed the worst, logistic regression with the ANOVA F-test feature selection approach, is represented by the orange line. It recorded noticeably low values, especially in recall and accuracy. The significant performance difference between models in forecasting e-commerce client attrition is shown in this chart.

Table 6. Lowest Evaluation Result

Experiment	Model	Feature Selection	Accuracy	Precision	Recall	F1 Score	ROC
1	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
2	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
3	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
4	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
5	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
6	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
7	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
8	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
9	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017
10	Logistic Regression	ANOVA F-Test	0.89076	0.77686	0.494737	0.604502	0.888017

Ten experimental runs were conducted to assess each model in Tables 7 and 8 to ensure statistical rigor. With an almost flawless ROC-AUC of 0.9995 ± 0.00007 and an average accuracy of 0.9869 ± 0.0004, the Gradient Boosting Classifier showed exceptional stability. As indicated by the very low standard deviation values across all metrics, the model is able to maintain performance consistency regardless of changes in data partitioning.

Table 7. Performance Stability Analysis of Best Evaluation on Gradient Boosting Classifier

Metrics	Mean	Standard Deviation
Accuracy	0.9869	0.0004
Precision	0.9994	0.0017
Recall	0.9231	0.0035
F1 Score	0.9598	0.0013
ROC	0.9995	0.00007

Table 8. Performance Stability Analysis of the Lowest Evaluation on the Logistic Regression Classifier

Metrics	Mean	Standard Deviation
Accuracy	0.8907	0.0000
Precision	0.7768	0.0000
Recall	0.4947	0.0000
F1 Score	0.6045	0.0000
ROC	0.8880	0.0000

Although the suggested model reaches an impressive accuracy of 98.69%, there are still issues with its generalizability. Because this study was limited to a single dataset, the performance measures are strongly correlated with the dataset's specific data distribution and customer behavior patterns.

From a practical perspective, the Gradient Boosting Classifier's high Recall of 92.31% provides significant value for e-commerce practitioners. Given that customer acquisition costs far exceed retention costs, the model's ability to accurately identify 92 out of 100 customers at risk of churning enables the marketing team to implement proactive, data-driven retention strategies. By prioritizing these high-risk segments, developers can enhance service quality and maintain market competitiveness more effectively than traditional linear models allow.

The best outcome from 10 trial runs using the Gradient Boosting model and Information Gain feature selection is shown in the confusion matrix in Figure 8. This outcome was chosen because it achieved the best evaluation performance across ROC AUC, F1 Score, Accuracy, and Precision. The model demonstrates strong predictive capability and high consistency. This is evidenced by its success in accurately classifying 900 non-churn samples and 97 churn samples, with a minimal error rate, specifically only 36 false positives and 93 false negatives.

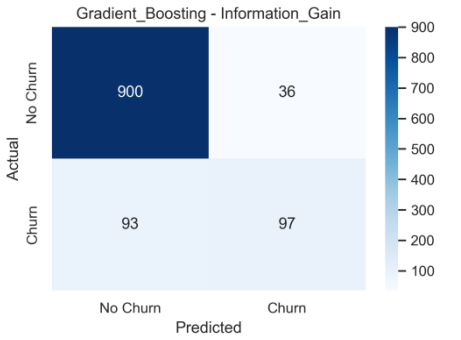


Figure 8. Confusion Matrix Visualization

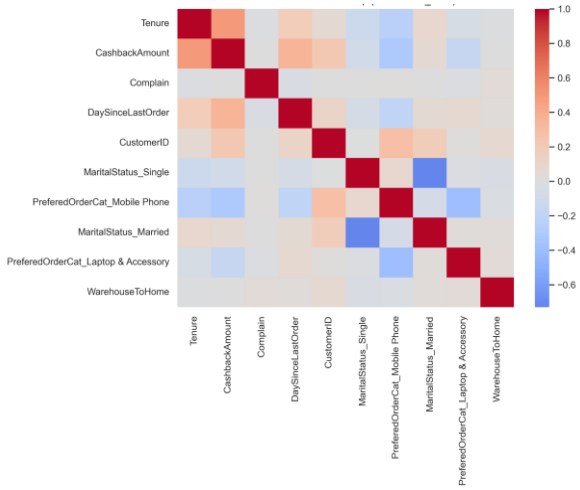


Figure 9. Heatmap Visualization

In Figure 9, the correlation heatmap, the selected features of Tenure, CashbackAmount, DaySinceLastOrder, Complain, MaritalStatus (Single and Married), PreferredOrderCat (Mobile Phone and Laptop & Accessory), and WarehouseToHome, show low to moderate inter-correlations, indicating minimal multicollinearity and great feature independence. The moderate relationship between Tenure and CashbackAmount reflects customer loyalty behavior, while DaySinceLastOrder captures recency effects relevant to churn prediction. Given one-hot encoding, the degree of negative correlation among the marital status categories is expected. Because these traits produced superior prediction performance, they were selected based on information gain. Using this feature subset yields more consistent evaluations for Logistic Regression, Random Forest, and Gradient Boosting—particularly for ensemble models. These findings further reinforce the effectiveness and reliability of its application in classification.

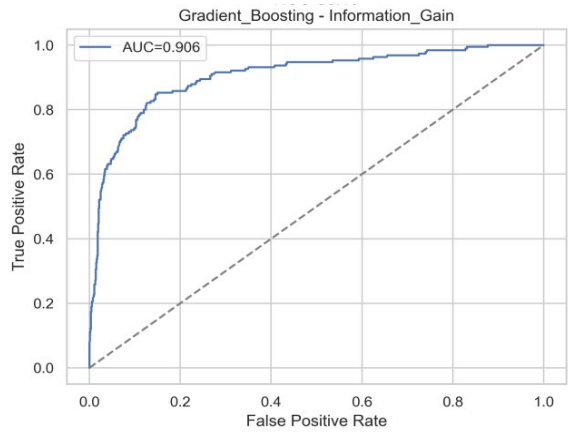


Figure 10. ROC Curve

With an AUC of 0.906, the ROC curve in Figure 10 for the gradient boosting model using Information Gain-selected features shows the best classification performance. This shows that the model has an outstanding discriminative capacity, with a 90.6% accuracy in distinguishing between churn and non-churn consumers. This model successfully achieved a high true positive rate while maintaining a relatively low false positive rate. This is reflected in the curve's position, which arches well above the random classifier's diagonal, particularly in the low false-positive range. The curve's sharp first climb indicates that the algorithm successfully identifies the most important churn cases early on, which is crucial in real-world business situations where early detection enhances retention tactics. Overall, the high AUC indicates that the combination of Gradient Boosting and Information Gain feature selection yields a strong, trustworthy predictive model.

Table 9. Independent t-test results of three algorithms

Rank	Accuracy	Precision	Recall	F1 Score	ROC
LR=>GB	P < 0.001	P < 0.001	P < 0.001	P < 0.001	P < 0.001
GB=>RF	P < 0.001	P < 0.001	P < 0.001	P < 0.001	P < 0.001
RF=>LR	P < 0.001	P < 0.001	P < 0.001	P < 0.001	P < 0.001

To evaluate the significance of performance differences among Logistic Regression, Random Forest, and Gradient Boosting, independent t-tests were conducted on the accuracy, precision, recall, F1-score, and ROC AUC metrics, as presented in Table 9. The findings show that all pairwise comparisons yielded very low p-values ($p < 0.001$), indicating statistically significant differences in algorithm performance. These results verify that the observed differences in categorization outcomes are not the consequence of chance.

Notably, Gradient Boosting demonstrates significantly superior performance, particularly compared to Random Forest on Precision and ROC, highlighting its stronger discriminative capability and predictive reliability. Additionally, the significant differences observed between Logistic Regression and the ensemble-based models suggest that more complex learning approaches are better suited for capturing non-linear patterns in the data. Overall, the statistical evidence supports Gradient Boosting as the most robust algorithm for the classification task under study.

The high importance of features like "Tenure" and "Complaint" suggests that customer dissatisfaction must be addressed immediately. A sudden spike in complaints should trigger an automated "high-risk" flag in the CRM system. Instead of offering generic discounts to the entire customer base, which can erode profit margins, the model allows companies to focus their resources exclusively on the high-risk segment, ensuring a higher Return on Investment (ROI) for retention campaigns.

These findings have several useful implications for decision-making and consumer retention strategies in e-commerce settings. The findings of the experiment show that churn prediction performance is greatly enhanced when feature selection methods are combined with ensemble machine learning. When combined with Information Gain, the Gradient Boosting (GBC) model produced the most reliable and accurate results, with a recall of more than 92% and an ROC AUC of nearly 1. Compared with Random Forest and Logistic Regression, this shows better discriminative power and prediction accuracy. The substantial performance difference between the ensemble-based models and Logistic Regression indicates that boosting techniques are more effective at capturing intricate, non-linear patterns in e-commerce data. GBR is the most reliable method for this classification problem, according to statistical evidence.

From a business perspective, these findings offer several practical implications for e-commerce decision-making. The high importance of features such as 'Tenure' and 'Complaint' indicates that customer dissatisfaction is a primary driver of churn. Consequently, a sudden spike in complaints should trigger an automated 'high-risk' flag in the CRM system, enabling immediate intervention. Furthermore, by achieving high Recall, the model minimizes 'missed'

churners, allowing companies to transition from generic, margin-eroding discounts to targeted retention campaigns. By focusing resources exclusively on the high-risk segment identified by the model, firms can ensure a higher return on Investment (ROI) and maintain healthier profit margins while proactively securing customer loyalty.

In practical terms, these results encourage e-commerce platforms to adopt data-driven retention strategies. Businesses can concentrate retention efforts on high-risk clients through targeted promotions and service enhancements by incorporating the model into customer analytics or CRM systems. Furthermore, feature selection makes the method effective and scalable for practical implementation by lowering data complexity while preserving good predictive performance.

The combination of feature selection and classifier choice has a significant impact on predictive performance in churn analysis. The model's ability to detect churn-prone customers is demonstrated by the results, which show that Gradient Boosting performs exceptionally well after dimensionality reduction, with Information Gain and LASSO contributing considerably to increases in recall and F1. Recall and F1 score, two critical indicators for identifying as many at-risk clients as possible, have shown particularly notable gains.

This work presents a novel visual comparison of model performance before and after feature selection. Business professionals and data scientists can intuitively grasp the significance of the feature selection stage for the efficacy of churn prediction models through quantitative evaluation and visualization of each model-feature selection method combination.

Jingyuan Li’s research, "Customer Churn Prediction using Machine Learning," is also cited for its insightful application of advanced models and techniques. It emphasizes the strengths of Random Forest and XGBoost for accurate predictions, addresses class imbalance using SMOTE, and enhances model explainability with SHAP and LIME [16].

Bagaskara Putra Wibowo and Lili Ayu Wulandhari, in their research titled E-Commerce Customer Churn Prediction Using Machine Learning Approaches, evaluate the performance of XGBoost and CatBoost in predicting churn across three experimental settings: testing on raw data, K-Means-based outlier handling, and outlier elimination using DBSCAN. Their findings revealed that removing outliers tended to reduce performance, with CatBoost consistently outperforming XGBoost. Notably, their model without outlier removal achieved the highest F1 score (96%) on CatBoost [17].

Previous research by Haoran Ren, titled "Machine Learning-Based Prediction of Customer Churn Risk in E-commerce," employed Logistic Regression, Random Forest, and Gradient Boosted Decision Trees to predict customer churn risk. With the highest AUC and KS scores and the highest accuracy, the stacking model outperformed the others. "The study results indicate that Random Forest outperforms other methods, particularly on test data, demonstrating the reliability of ensemble models such as stacking and RF in churn prediction [18]."

According to the performance results, LR performed the worst; between training and testing, the AUC decreased from 0.880 to 0.870. While RF obtained an AUC of 0.937 in testing, GBDT fared better (AUC 0.910 in training, 0.921 in testing). With an accuracy of 0.928 and a ROC AUC of 0.940, the stacking strategy produced the best results, highlighting the predictive power of ensemble-based learning.

The notion that ensemble approaches, especially Gradient Boosting and Random Forest, exhibit superior classification and generalization skills is supported by these comparative results. Furthermore, the best churn prediction performance is possible with stacking strategies that incorporate several models. Because of this, they offer a useful solution for practical use on e-commerce platforms, where interpretability, correctness, and resilience are crucial.

The top outcomes from this study's suggested approach, which combines the Gradient Boosting algorithm with Information Gain and LASSO for feature selection, showed remarkably high and reliable performance across all assessment measures. The ROC AUC almost reached perfection, with the maximum value recorded at 0.9997, while the average accuracy reached 0.9867, precision peaked at 1.0, recall ranged from 0.9211 to 0.9316, and F1 score approached 0.96. The effectiveness of the suggested combination is confirmed by these results, which far outperform those from several earlier trials.

Theoretically, better performance can be explained by the inherent compatibility of Gradient Boosting and Information Gain. Information Gain, which closely resembles the internal splitting process of decision trees used in Gradient Boosting, assesses features by their ability to maximize class separability and reduce entropy. Preselecting highly discriminative features using Information Gain enables each boosting phase to focus on the most informative patterns in the data since Gradient Boosting builds trees sequentially, with each iteration concentrating on fixing the mistakes of the preceding one. In high-dimensional e-commerce datasets, where irrelevant features can seriously impair model performance, this synergy between the feature selection criterion and the model's sequential learning process minimizes redundancy, reduces noise in the input space, and ultimately improves both predictive accuracy and generalization capability.

Table 10 illustrates that their top-performing model, XGBoost, achieved an accuracy of 0.96, a recall of

0.88, and an ROC AUC of 0.9872, compared with the work of Jingyuan Li [16], which tested multiple machine learning models on e-commerce data. Despite being comparatively high, these numbers are nevertheless lower than those of the study's suggested Gradient Boosting model, especially for the F1 score (0.89 vs. 0.96) and ROC AUC (0.9872 vs. 0.9997). Additionally, the study's flawless precision score of 1.0 indicates a high capacity to accurately identify churn cases.

Table 10. Performance Comparison based on [16]

Method	Acc	Prec	Recall	F1-Score	ROC AUC
Logistic Regression	0.78	0.95	0.42	0.55	0.8403
Decision Tree	0.92	0.96	0.74	0.78	0.8826
Random Forest	0.95	0.98	0.83	0.86	0.9851
XGBoost	0.96	0.98	0.88	0.89	0.9872
LightGBM	0.93	0.96	0.78	0.8	0.9708

As shown in Table 11, the study by Haoran Ren [18] reported the best performance using Stacking and Random Forest models, achieving an accuracy of 0.928 and an ROC AUC of 0.94. Although high, these results still fall short of the performance achieved by our proposed method. The lower recall and F1-score values of the stacking-based models [18] indicate a limited ability to detect actual churners compared to the Gradient Boosting model used in this study.

Table 11. Performance Comparison based on [18]

Cross-Validated Performance on Training Data			
Models	KS	AUC	accuracy
LR	0.668	0.88	0.89
GBDT	0.701	0.91	0.899
RF	0.766	0.927	0.907
Stacking model	0.744	0.927	0.912
Model Performance on Test Data			
Models	KS	AUC	accuracy
LR	0.64	0.87	0.886
GBDT	0.718	0.921	0.903
RF	0.772	0.937	0.91
Stacking model	0.779	0.94	0.928

Table 12. Performance Comparison based on [17]

First Result			
Method	Highest Train F1-Score (%)	Highest Test F1-Score (%)	
XGBoost	100	95.2	
Catboost	100	96	
Second Result			
Outlier Removed (%)	Highest Train F1-Score (%)	Highest Test F1-Score	Algorithm
100	100	47.9	XGBoost
90	100	70.5	Catboost
80	100	76.8	Catboost
70	100	84	Catboost
60	100	85.2	Catboost
50	100	86.7	Catboost
40	100	87.2	Catboost
30	100	90.7	Catboost
20	100	90	Catboost
10	100	92.6	Catboost
Third Result			
Method	Highest Train F1-Score (%)	Highest Test F1-Score (%)	
XGBoost	100	94	
Catboost	100	95	
Final Result on F1-Score Comparison in three Tests			
Experiment	Method	Highest F1-Score	Algorithm
First	Baseline	96	Catboost
Second	K-Means Clustering	92.6	Catboost
Third	DBSCAN	95	Catboost

Table 13. Best Experimental Results of the Proposed Gradient Boosting with Feature Selection Method

Experiment	Model	Feature Selection	Accuracy	Precision	Recall	F1-Score	ROC
1	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
2	Gradient Boosting	Information Gain	0.987567	1	0.926316	0.961749	0.999567
3	Gradient Boosting	Information Gain	0.986679	1	0.921053	0.958904	0.999561
4	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
5	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
6	Gradient Boosting	Information Gain	0.987567	1	0.926316	0.961749	0.999635
7	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
8	Gradient Boosting	LASSO	0.986679	1	0.921053	0.958904	0.999533
9	Gradient Boosting	Information Gain	0.987567	0.994382	0.931579	0.961957	0.999623
10	Gradient Boosting	Information Gain	0.986679	1	0.921053	0.958904	0.999786

Table 14. Average Performance Results of the Proposed Method across 10 Experiments

Accuracy	Precision	Recall	F1-Score	ROC
9.869454	9.994382	9.231582	9.597783	9.995837

Lastly, Table 12 from the study by Bagaskara Putra Wibowo and Lili Ayu Wulandhari [17] shows that CatBoost achieved the best results, with test F1 scores of 96% in the baseline trial and 92.6% when paired with K-Means clustering. Although these findings are comparable, a thorough comparison is limited because the study does not include a complete set of performance indicators, such as accuracy, precision, and ROC AUC. On the other hand, the suggested model maintains high and consistent performance across all other criteria and achieves an identical or marginally higher F1 score.

As shown by the ROC AUC, this study concludes that the proposed approach outperforms prior studies in classification accuracy, precision-recall balance, and overall model resilience. The results show that churn prediction performance in the e-commerce domain can be significantly improved by selecting a suitable feature selection technique and combining it with a powerful algorithm, such as Gradient Boosting. The performance comparison includes the results of the suggested approach in this study, along with other research on customer churn prediction in e-commerce using different machine learning models, as summarised in Tables 13 and 14.

4. Conclusions

This study set out to compare several machine learning models and feature selection techniques to assess how user behavior affects e-commerce turnover. To improve customer retention and develop churn prediction strategies, selecting the appropriate model and features is essential.

Several important research problems were answered. First, whether patterns of client behavior, including the frequency of visits, the quantity of orders placed, or the length of interactions, can be used as markers of possible attrition. Second, which characteristics or factors have the biggest impact on consumer behavior, either making them discontinue using the service or motivating them to make more frequent purchases? Third, how feature selection methods might enhance the precision and effectiveness of predictive models. Finally, how well machine learning models can forecast churn and increase order volume.

This study confirms that the combination of feature selection technique and classifier choice has a significant impact on model performance in forecasting e-commerce turnover. When paired with Information Gain and LASSO feature selection techniques, the Gradient Boosting Classifier consistently produced the best results. This resulted in a considerable average improvement in Accuracy by 0.9869, Precision by 0.9994, Recall by 0.9232, F1-Score by 0.9598, and ROC AUC by 0.9996.

Logistic regression, on the other hand, performed the worst, particularly when using the ANOVA F-Test, and only marginally improved with alternative feature selection techniques.

The Gradient Boosting model was the best option in this situation since it consistently demonstrated accuracy, stability, and sensitivity to churn detection. Comparisons with earlier research, which also demonstrated that boosting-based models (such as XGBoost, CatBoost, and GBDT) typically perform well in managing the complexity of churn data, corroborate these findings [16–18]. Deep learning techniques may be investigated in future studies to handle more intricate consumer behavior patterns. To improve practical applicability, model success should also consider business-oriented measures such as Customer Lifetime Value (CLV) and Return on Investment (ROI), in addition to machine learning metrics such as AUC and F1-score.

Acknowledgements

The Department of Magister Information System of Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia, and the supervisory lecturers are especially appreciated by the authors for granting access to the research data.

References

[1] M. Orzoł and K. S. Depczyńska, "Development trends in e-commerce sector", *Procedia Computer Science*, vol. 225, pp. 4157–4166, Dec. 2023. doi: https://doi.org/10.1016/j.procs.2023.10.412

[2] S. Baghla and G. Garg, "Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce," *Microprocessors and Microsystems*, vol. 94, Oct. 2022. doi: https://doi.org/10.1016/j.micpro.2022.104680

[3] B. Prabadevi, R. Shalini, B.R. Kavitha, "Customer churning analysis using machine learning algorithms," *International Journal of Intelligent Networks*, vol. 4, pp. 145–154, 2023. doi: https://doi.org/10.1016/j.ijin.2023.05.005

[4] eMarketer, "Global ecommerce forecast 2022," Apr. 11, 2022. [Online]. Available: https://www.emarketer.com/content/global-ecommerce-forecast-2022

[5] D. G. Kleinbaum and M. Klein, Logistic Regression: A Self-Learning Text, 4th ed. Cham, Switzerland: Springer, 2021. doi: https://doi.org/10.1007/978-3-030-89886-4

[6] Obata, S., Cieszewski, C. J., Lowe, R. C., III, & Bettinger, P. "Random Forest Regression Model for Estimation of the Growing Stock Volumes in Georgia, USA, Using Dense Landsat Time Series and FIA Dataset," *Remote Sensing*, vol. 13, no. 2, p. 218, 2021. doi: https://doi.org/10.3390/rs13020218

[7] L. Patelli, M. Cameletti, N. Golini, and R. Ignaccolo, "A path in regression Random Forest looking for spatial dependence: a taxonomy and a systematic review," *arXiv preprint* arXiv:2303.04693, 2023. doi: https://doi.org/10.48550/arXiv.2303.04693

[8] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, 2024. doi: https://doi.org/10.58496/BJML/2024/007

[9] S. H. Samadi, B. Ghobadian, and M. Nosrati, "Prediction of higher heating value of biomass materials based on proximate analysis using gradient boosted regression trees method," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 43, no. 6, pp. 672–681,

2019. doi: https://doi.org/10.1080/15567036.2019.1630521
[10] F.-K. Wang and T. Mamo, "Gradient boosted regression model for the degradation analysis of prismatic cells," *Computers & Industrial Engineering*, vol. 144, p. 106494, 2020. doi: https://doi.org/10.1016/j.cie.2020.106494
[11] K. Rizkiyah, L. Nurmayanti, R. D. N. Macdhy, and A. Yusuf, "Pengaruh Digital Payment terhadap Perilaku Konsumen: Pengguna Platform Digital Payment OVO," *Jurnal Ilmiah Manajemen*, vol. 16, pp. 107–126, Apr. 2021. doi: https://doi.org/10.33369/insight.16.1.107-126
[12] Kim A. Johnston and Anne B. Lane, "The role of time and social churn in impact assessment: An engagement-based model," *Environmental Impact Assessment Review*, vol. 71, pp. 102–109, 2018. doi: https://doi.org/10.1016/j.eiar.2018.04.009
[13] W. H. Khoh, Y. H. Pang, S. Y. Ooi, L.-Y.-K. Wang, and Q. W. Poh, "Predictive churn modeling for sustainable business in the telecommunication industry: Optimized weighted ensemble machine learning," *Sustainability*, vol. 15, 2023. DOI: https://doi.org/10.3390/su15118631
[14] W. Park and H. Ahn, "Not all churn customers are the same: Investigating the effect of customer churn heterogeneity on customer value in the financial sector," *Sustainability*, vol. 14, 2022. doi: https://doi.org/10.3390/su141912328
[15] R. Hugo, B. Belé, C. M. António, and G. R. Ricardo, "Determinants of churn in telecommunication services: A systematic literature review," *Journal of Business Economics*, 2023. doi: https://doi.org/10.1007/s11301-023-00335-7
[16] J. Li, "Customer churn prediction using machine learning: A case study of e-commerce data," *International Journal of Computer Applications*, vol. 186, no. 48, 2024. doi: https://doi.org/10.5120/ijca2024924140
[17] B. P. Wibowo and L. A. Wulandhari, "E-commerce customer churn prediction using machine learning approaches," *Jurnal Penerapan Sistem Informasi*, vol. 5, no. 4, pp. 1597–1606, 2024. doi: https://doi.org/10.30645/kesatria.v5i4.482
[18] H. Ren, "Advances in Economics, Management and Political Sciences", vol. 153, no. 1, pp. 47–52, 2024. doi: https://doi.org/10.54254/2754-1169/2024.19473
[19] Y. H. Xiancheng Xiahou, "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458-475, 6 April 2022. doi: https://doi.org/10.3390/jtaer17020024
[20] M. Kamil and K. Katarzyna, "Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach," *Journal of theoretical and applied electronic commerce research*, vol. 17, no. 1, pp. 165-198, 2022. doi: https://doi.org/10.3390/jtaer17010009
[21] A. L. Renato, C. S. Thiago, and M. T. Benjamin, "Propension to customer churn in a financial institution: a machine learning approach," *Neural Computing and Applications*, vol. 34, no. 1, p. 11751–11768, 2022. doi: https://doi.org/10.1007/s00521-022-07067-x
[22] M. Maduna, A. Telukdarie, I. Munien, U. Okonkwo, and A. Vermeulen, "Smart Customer Churn Management System Using Machine Learning," *Procedia Computer Science*, vol. 237, pp. 552–558, 2024. doi: https://doi.org/10.1016/j.procs.2024.05.139
[23] E. Fatima et al., "Empirical Analysis of Tree-Based Classification Models for Customer Churn Prediction," *Scientific African*, 2023. doi: https://doi.org/10.1016/j.sciaf.2023.e02054
[24] Pahul et al., "Investigating Customer Churn in Banking: A Machine Learning Approach and Visualization App for Data Science and Management," *Data Science and Management*, vol. 7, no. 1, pp. 7–16. 2024. doi: https://doi.org/10.1016/j.dsm.2023.09.002
[25] S. Khalid, T. Khalil, and S. Nasreen, "A Survey of Feature Selection Techniques in Machine Learning," in *Proc. 2014 Science and Information Conf. (SAI)*, London, UK, 2014, pp. 372–378, 2014. doi: https://doi.org/10.1109/SAI.2014.6918213
[26] A. Elisseeff, & I. Guyon, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.
[27] I. Guyon, J. Weston, S. Barnhill, & V. Vapnik,, "Machines, Gene Selection for Cancer Classification using Support Vector," Machine Learning, vol. 46, p. 389–422, 2002. doi: https://doi.org/10.1023/A:1012487302797
[28] C. E. Shannon, "A Mathematical Theory of Communication," The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
[29] S. Alwateer, E. Atlam, M. A. El-Raouf, O. A. Ghoneim, and I. Gad, "Missing data imputation: A comprehensive review," *Journal of Computer and Communications*, vol. 12, no. 11, pp. 53–75, 2024. doi: https://doi.org/10.4236/jcc.2024.1211004
[30] W. Zhu, R. Qiu, and Y. Fu, "Comparative study on the performance of categorical variable encoders in classification and regression tasks," *arXiv preprint*, 2024. doi: https://doi.org/10.48550/arXiv.2401.09682
[31] J. M. H. Pinheiro, S. V. B. de Oliveira, T. H. S. Silva, P. A. R. Saraiva, E. F. de Souza, L. A. Ambrosio, and M. Becker, "The impact of feature scaling in machine learning: Effects on regression and classification tasks," *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2506.08274
[32] F. Chandrashekar and S. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014. doi: https://doi.org/10.1016/j.compeleceng.2013.11.024
[33] S. Abdumalikov and Y. Yusupov, "Performance analysis and improvement of machine learning with various feature selection methods for EEG-based emotion classification," *Applied Sciences*, vol. 14, no. 22, 2024. doi: https://doi.org/10.3390/app142210511
[34] P. Nicholas, et.al, "A review of feature selection methods for machine learning," *Frontiers in Bioinformatics*, vol. 2, 2022. doi: https://doi.org/10.3389/fbinf.2022.927312
[35] A. Sikri, N. P. Singh, and S. Dalal, "Chi-square method of feature selection: Impact of pre-processing of data," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, pp. 241–248, 2023.
[36] A. M. Priyatno and T. Widyaningtyas, "A systematic literature review: Recursive feature elimination algorithms," *Jurnal Ilmu Pengetahuan dan Teknologi Komputer*, vol. 9, no. 2, pp. 196–207, 2024. doi: https://doi.org/10.33480/jitk.v9i2.5015
[37] A. Gupta, "Feature selection techniques in machine learning," *Analytics Vidhya*, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/
[38] M. Parzinger, L. Huber, F. Sauter, and U. Schmid, "Comparison of different training data sets from simulation and experimental measurement with artificial users foroccupancy detection – Using machine learning methods Random Forest and LASSO," *Building and Environment*, vol. 223, 2022. doi: https://doi.org/10.1016/j.buildenv.2022.109313
[39] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: Wiley, 2013. doi: https://doi.org/10.1002/9781118548387
[40] F. Piotr and Z. Adam, "Benchmarking state-of-the-art gradient boosting algorithms for classification," *arXiv preprint* arXiv:2305.17094, 2023. doi: https://doi.org/10.48550/arXiv.2305.17094
[41] W. R. Lydia, "Enhancing the performance of gradient boosting trees on regression problems," *Journal of Big Data*, 2025. doi: https://doi.org/10.1186/s40537-025-01071-3