

Article

Analyzing the Impact of Feature Selection on Customer Churn Prediction in the Retail E-Commerce Industry

Meryem Chajia , El Habib Nfaoui *  and Soufiyan Ouali 

L3IA Laboratory, Department of Computer Science, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez 30000, Morocco; meryem.chajia@usmba.ac.ma (M.C.); soufiyane.ouali@usmba.ac.ma (S.O.)

* Correspondence: elhabib.nfaoui@usmba.ac.ma

Abstract

Customer churn has become a major challenge in the retail industry, where customer loyalty directly affects business success and sustainability. Despite the significant progress in Artificial Intelligence, especially in prediction tasks, its use in the retail e-commerce domain remains limited and underexplored; this is due to the scarcity and limited quality of available datasets. To address these challenges, this paper proposes a churn prediction approach designed to handle data scarcity while ensuring accurate performance. We experimented with a combination of various feature selection techniques along with several Machine Learning and Deep Learning models to evaluate their performance on a limited tabular dataset. The impact of feature selection on predictive performance was also systematically analyzed. The results demonstrated that feature selection plays an important role in improving model performance by identifying the key features that have the most significance to the classification task. The analysis showed that the L1-based Logistic Regression feature selection method combined with the Extreme Gradient Boosting classifier achieved the best performance, with a Macro F1-score of 95.25%. Based on these results, companies can identify potential churners and implement retention strategies. These findings may provide a useful reference point for future researchers in the retail e-commerce industry.

Keywords: retail e-commerce; churn prediction; feature selection; machine learning; deep learning

1. Introduction

Retail industries are businesses that purchase products from manufacturers and sell them to end customers, with the primary goal of maximizing sales and profitability. Retailers operate in various formats, including physical stores, national chains, international outlets, and online platforms. The latter (i.e., online shopping) has seen rapid growth in recent years. Today, many consumers shop through their smartphones or computers and receive their purchases via home delivery. To meet diverse customer preferences, many businesses have combined physical and online stores. The retail sector has continued to grow year after year, leading to increasingly intense competition in the market. This increased competition gives customers greater flexibility to switch easily from one brand to another, and that is what we call churn.

Customer churn refers to the rate at which customers cease their relation with a business; those customers switch to competitors for many reasons, such as better service quality, more attractive pricing, or better offers. Customer churn is a critical issue in the



Academic Editor: Olufemi A. Omitaomu

Received: 30 April 2026

Revised: 17 June 2026

Accepted: 25 June 2026

Published: 1 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

retail industry, as it directly affects both revenue and customer lifetime value, and this causes the company to lose customers, which negatively impacts sales and revenue, and moreover, it incurs additional costs in acquiring new ones. Customers are among the most valuable assets of any business, and retaining existing customers is generally more cost-effective than acquiring new ones. As a result, in today's highly competitive environment, customer retention has become a key objective for retailers. To retain customers, companies need to understand their needs and respond effectively to shifting market trends, an effort that often comes with significant challenges [1]. Therefore, it is important for businesses to identify potential churners early in order to take proactive measures to prevent them from leaving.

To achieve this, customer data (customer transactions, orders, personal information, and behavior) can be exploited in many ways; Using churn analysis, companies can detect hidden patterns behind the data and then detect churn before it happens, thereby the company can take proactive steps to prevent churners from leaving. Based on that, businesses can develop targeted retention strategies that will then target the right audience, thus preventing customer loss. At present, Machine Learning (ML) and Deep Learning (DL) models are showing very promising results, especially in the context of churn analysis. DL, in particular, can extract complex patterns from both structured and unstructured data by training neural networks on customer datasets. The use of these models can allow companies to make accurate predictions in the context of customer churn, thereby enabling companies to make informed decisions.

The efficiency of the developed solution or model is highly dependent on the quality of the training data or knowledge base. Frequently, the data collected by companies is purely transactional, consisting mainly of customer orders and purchase dates, with each customer identified by an ID. In other cases, the dataset may contain much more detailed information, such as customer demographics, behavioral patterns, product descriptions, and category information. The use of all these features can complicate the modeling process by increasing computational costs and raising the risk of overfitting. As a result, the model performs well on the training data but fails to generalize effectively to new, unseen data [2]. Thus, it is important to select only the features that are relevant to the prediction task and informative with respect to the target variable. This is where feature selection comes into play.

Feature selection helps identify the most important variables, reduces dimension, and can significantly enhance model performance [3] and convergence. There are various feature selection techniques available, each with the potential to improve the accuracy and generalization ability of churn prediction models. Despite its importance, feature selection is often overlooked; many studies, particularly in the churn prediction context, fail to systematically compare the impact of different techniques on model performance. The novelty of this research lies in the systematic evaluation of several feature selection methods for improving churn prediction performance on a limited dataset, which is a significant challenge in the retail e-commerce context. To the best of our knowledge, this is the first study to evaluate and compare all the following feature selection techniques in the context of churn: Chi-square (χ^2), Mutual Information (MI), Minimum Redundancy Maximum Relevance (mRMR), Relief Feature Selection Algorithm (ReliefF), Recursive Feature Elimination with Cross-Validation (RFECV), Boruta, L1-based Logistic Regression, Random Forest-based selection (RF-BS), Permutation Feature Importance-based feature selection (PFI-BS). We focused on the retail e-commerce domain because it is one of the most globalized industries, with a notable impact on economic development [4]. This sector is highly competitive and characterized by rapidly changing customer preferences, evolving market trends, and strong competition among online retailers [5]. In this context,

applying churn prediction models can provide researchers as well as the business owners with actionable insights to accurately predict the behaviors of the customers and detect any intent to churn, thereby, businesses be able to sustain long-term profitability and competitiveness.

The main objective of this study is to compare numerous existing feature selection methods for detecting deep patterns in customer data. In addition, this research aims to identify which types of information should be collected by companies to enhance the effectiveness of churn prediction models, thus enabling companies to improve customer retention and reduce churn rates. The main contributions of this paper are as follows:

- We evaluate the impact of different feature selection methods on the performance of churn prediction models and specifically in the retail e-commerce industry, which is often characterized by limited data.
- We conduct a comprehensive evaluation of nine feature selection techniques applied to both machine learning and deep learning models, and compared model performance with and without feature selection to quantify its impact on predictive performance.
- The study determines the most effective combinations of feature selection techniques and classifiers for customer churn prediction in the retail e-commerce domain, providing valuable insights for researchers and data scientists working with similar datasets or addressing related prediction tasks.
- The study aims to demonstrate that selecting an appropriate feature selection technique can significantly improve the performance of the predictive model.
- The study identifies the key customer attributes that may impact churn, which can help organizations know which type of data they should collect to strengthen their prediction capabilities.

This paper is organized as follows: in Section 2, we give background information and a description of key concepts related to churn prediction; we also present the current state of the art in this domain. In Section 3, we introduce our methodology, while in Section 4, we discuss the experimental results and analysis. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Background and Related Work

In this section, we explain the key terminology used in the field of churn prediction. Moreover, we also examine the current state of the art (SOTA) to highlight the different methods applied in this domain.

2.1. Background

In the customer churn prediction domain, there are three key analytics metrics that are commonly used: customer retention, customer loyalty, and customer churn. These terms are related and sometimes used interchangeably; they represent different aspects of customer behavior and business performance.

Customer retention refers to the process of engaging existing customers to encourage continued purchases of a company's products or services [6]. It aims to reduce customer churn by maintaining strong relationships. Effective customer retention strategies enable companies to keep as many customers as possible. The process of customer retention begins with the customer's first interaction with the company and continues throughout the entire duration of the customer relationship [7]. The Customer Retention Rate Formula is presented in Equation (1).

$$\text{Customer Retention Rate} = 100 \times \frac{\text{NCE} - \text{NEW}}{\text{NCS}} \quad (1)$$

where

NCS = Number of Customers at Start of Time Period,

NCE = Number of Customers at End of Time Period,

NEW = Number of New Customers Acquired During Time Period.

Customer loyalty refers to customers who continue to buy from a particular brand, company, or business, due to a strong emotional connection with that brand. Loyal customers are not only more likely to make repeat purchases, but they also tend to promote the brand through word-of-mouth, show interest in additional products or services offered by the same company, and be more likely to provide feedback or reviews. They are generally less inclined to look for the same product or service elsewhere and are less influenced by competitors' marketing efforts [8]. Customer retention and customer loyalty are linked; customer retention metrics can be used to measure customer loyalty, but they do not mean the same thing. Retention measures a business's ability to keep customers over time, while loyalty reflects a customer's willingness to continue engaging with the brand's products or service [9].

Customer churn refers to the proportion of customers who stop using a company's product or service during a specific period. It is typically calculated by dividing the number of customers lost during that time frame by the number of customers at the beginning of the period [10]. The Customer Churn Rate Formula is presented in Equation (2).

$$\text{Churn Rate} = 100 \times \frac{\text{NCS} - \text{NCE}}{\text{NCS}} \quad (2)$$

With NCS, NCE, NEW defined previously.

Customer churn in retail e-commerce refers to the percentage of customers who stop purchasing from an online retailer within a given period. The common challenge for businesses is that customers switch to competitors due to changing preferences or overall shopping experience. Effective churn management can help maintain sustainable growth and minimize customer acquisition costs. Retailers must acquire new customers at a rate that exceeds the loss of existing ones. Customer loyalty often provides the foundation for successful brands, while neglecting retention efforts can limit long-term performance. Moreover, returning customers not only enhance profitability but also serve as a key indicator of a brand's relevance and value to its audience.

2.2. Related Work

Various studies have been conducted to detect churn and especially in the retail domain. Section 2.2 provides an overview of existing studies; highlighting their methodologies, datasets, and main findings.

In [11], researchers conducted a study using a real-world dataset extracted from the website of the Turkish company Colin's. The dataset used in this study includes customer behavior information, such as product types purchased, shopping frequency, and preferred shopping channels (e.g., online, physical store). The authors evaluated several ML algorithms, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR), and Naive Bayes (NB), to identify the most effective model. To balance the dataset, the study used the SMOTE oversampling method [12] as a data augmentation technique; in addition to that, they employed Recency, Frequency, and Monetary (RFM) scores, which enhanced the model's performance significantly. RFM scores are numerical indicators used in marketing analysis to evaluate and segment customers based on their purchase history. Recency refers to how recently a customer made a purchase, Frequency represents the purchase frequency of the customer, and Monetary indicates the total spending the customer made during a specific time frame. Additionally, the

authors suggested adding a user clustering feature based on the RFM score. The results achieved show that XGBoost outperformed the other models, achieving an accuracy of 73%. Similarly, in [13], the authors employed RFM scores as a feature engineering technique to enhance churn prediction analysis and model development; they used a real-world dataset from the European fashion retail industry. This dataset included only customer information, specifically, customer ID, purchase date, and transaction revenue. Despite the absence of demographic or product-level data, the study demonstrated that purchase history alone can offer valuable business insights. The study combined traditional RFM analysis with predictive modeling, enabling the discovery of meaningful patterns: over 50% of customers churn each year, and this highlights a low level of customer loyalty. For churn prediction, both RF and Gradient Boosting (GB) models were applied, achieving comparable performance, with an accuracy of 65.1% for RF and an accuracy of 65% for GB. Another use of the RFM scores is presented in [14], where the authors used this feature for users' segmentation. They utilized a dataset containing two years of transaction records from an online retail store based in the United Kingdom. The dataset includes attributes such as transaction datetime, product description and quantity, unit price, and customer ID. To identify customers most likely to churn, they used the K-means clustering based on RFM scores. Within the same context, the study [15] employed RFM scores for both feature engineering and user segmentation. Based on these scores, customers were segmented into 10 segments. For example, the "Loyal" segment includes customers who have made recent and frequent purchases, though with lower transaction values. The "Need Attention" group represents customers who have made recent transactions with decreasing frequency and value. The "At Risk" segment refers to customers at risk of churn; they are the customers who have not made transactions recently but have a history of low-value and frequent transactions. Certain segments were then labeled as churn and others as "non-churn", and this classification field was subsequently integrated into the main dataset to serve as the target variable. The researchers used an Indonesian fashion e-commerce market dataset that contains demographic and behavioral information. For model selection, seven ML algorithms were evaluated: LR, k-Nearest Neighbors (KNN), RF, Adaptive Boosting (AdaBoost), XGBoost, Decision Tree (DT), and Extra Trees. As a result, the RF algorithm achieved the best accuracy of 99.00%, which makes it the most appropriate model for churn prediction in this context. These studies confirm that RFM-based feature engineering remains a widely adopted and effective strategy for representing customer purchasing behavior in retail churn prediction tasks.

The Brazilian e-commerce retail sector has also attracted research attention. In [16], the authors conducted a comprehensive analysis of customer churn within this context. They combined three distinct datasets: detailed numerical data on customers' orders, textual after-purchase customer reviews, and socio-geo-demographic information such as age, region, and income. The paper integrated feature selection approaches to enhance the convergence of their models. Thus, they applied the Boruta algorithm [17], a feature selection method which automatically identifies the most informative features. For the classifier selection, they compared XGBoost and LR. To interpret the model's prediction and assess feature relevance, they applied several Explainable Artificial Intelligence (XAI) techniques [18], including permutation-based variable importance, partial dependence profiles, and bootstrapped AUC scores. The XGBoost model achieved the highest test AUC of 65.05%. The study demonstrated that customer churn is influenced by several factors: payment value of the first order, shipping cost, product categories, demographic context, geographical location, and number of items purchased; for example, customers purchasing only one or two items are more likely to return for another purchase compared to customers who bought three or more items. Notably, the authors showed that churn

prediction is feasible using only data from the first transaction; in their dataset, 96.57% of all orders corresponded to customers purchasing for the first time, while only 0.05% of orders came from customers who made five or more purchases. Moreover, only 3.18% of first-time buyers make a second order, whereas the probability of progressing from a second to a third purchase increased to 8.56%. This indicates that encouraging customers to make a second purchase is the most challenging stage of retention, while subsequent purchases become progressively easier. Therefore, the study focused exclusively on first orders to better understand initial customer behavior and churn dynamics. The authors highlighted the value of integrating multiple data sources. The main source was store transaction data from the Brazilian company Olist, which was enriched with a geo-demographic dataset obtained from the Brazilian Statistical Office. This integration helped address the challenge of predicting churn in the absence of a long purchase history. However, the study has some limitations. First, only two ML algorithms were tested. Second, the data covers transactions from 2016 to 2018, before the COVID-19 pandemic, which suggests the need for comparison with pandemic and post-pandemic periods to evaluate the generalizability of the identified churn factors. Third, the use of e-commerce data (as opposed to data from brick-and-mortar retail or subscription-based businesses where customers are contractually bound) may limit the applicability of findings to other retail contexts. Consistently, XGBoost achieved superior performance in the study [19]. The dataset used contains transaction records from an online retail gift store in the United Kingdom, and it includes detailed information on customer purchases. To predict customer churn, the study evaluated three machine learning algorithms: RF, Support Vector Machines (SVM), and XGBoost. As a result, XGBoost achieved the highest accuracy, reaching 71%, and this highlights its effectiveness for churn prediction in this retail context. All these results confirm the effectiveness of tree-based ensemble models for churn prediction in retail domain.

Other studies have investigated the performance of DL models for churn prediction in order to capture temporal purchase behavior. For instance, researchers in [20] proposed a deep survival framework to predict customer churn in retail companies operating in non-contractual settings. The model was built using customer transactional data from a large UK-based retailer, which includes features such as Customer ID, Order Number, and Purchase Date. Notably, the approach relied only on transactional information, consciously excluding customer-level attributes like age or gender. To enrich the analysis, a synthetic dataset was generated by calculating the time intervals between consecutive purchases. In the classification stage, the authors employed a recurrent neural network (RNN), specifically a Long Short-Term Memory (LSTM) architecture. This model can capture temporal purchase patterns. The model was evaluated on both real-world and synthetic datasets; on the real retail data, it achieved an accuracy of 65.4% and on the synthetic data, it reached 56%. However, the authors noted some limitations; for example, they assumed that the time between purchases follows an exponential distribution, which may not adequately represent the behavior of highly loyal customers who buy repeatedly but irregularly. Furthermore, the approach is better suited to retail environments with relatively frequent customer transactions. Similarly, in study [21], the LSTM model also demonstrated strong performance. The paper used transactional data from a French retail company; this data contains records of customers with loyalty cards. They compared four models: Linear Regression, Multi-Layer Perceptron (MLP), XGBoost, and LSTM to determine the best performer in terms of precision. Among these, the LSTM model achieved the highest precision, with a performance of 73.70%.

In this study, RFM-based feature engineering was not applied because the available dataset does not contain transactional information required to compute recency, frequency, and monetary indicators. Similarly, customer segmentation techniques commonly used in

retail analytics often rely on transactional behavioral features such as RFM scores, which were not available in this dataset. Therefore, alternative feature selection and modeling strategies were adopted to better address the specific characteristics of the data.

Building on the current state of the art, it can be observed that feature selection—one of the key techniques for robust model development—has received limited attention in the existing literature on customer churn prediction using publicly available datasets. Most previous studies do not thoroughly investigate the impact of selecting the most relevant features on model performance. Feature selection plays a crucial role in data preprocessing by identifying the most informative predictors while eliminating irrelevant or redundant variables, thereby improving model robustness and interpretability. Moreover, the few studies that incorporate feature selection typically rely on a single selection strategy and do not provide a comprehensive evaluation of multiple feature selection techniques. As a result, the specific contribution of feature selection to improving churn prediction performance remains insufficiently explored in this context.

To fill these gaps, the present study provides a comparative evaluation of several feature selection techniques combined with machine learning and deep learning models to identify the most effective predictors for customer churn prediction in the retail industry.

Table 1 represents a summary of Datasets, Methods, and Performance Metrics in the State of the Art.

Table 1. Summary of State-of-the-Art Studies.

Reference	Domain	Data Sources	Models/Methods Used	Feature Selection	Best Model	Reported Performance (Metrics Vary Across Studies)
[15]	Retail Industry	Demographic, behavioral information	LR, KNN, RF, AdaBoost, XGBoost, DT, Extra Trees	None	RF	Accuracy: 99.00%
[11]		Customer behavior information	RF, XGBoost, LR, NB	None	XGBoost	Accuracy: 73%
[20]		Transactional data	LSTM	None	LSTM	Accuracy: 65.4%
[21]		Transactional data	Linear Regression, MLP, XGBoost, LSTM	None	LSTM	Precision: 73.70%
[16]		Customers' orders, reviews, socio-geodemographic data	XGBoost, LR	Boruta algorithm	XGBoost	AUC: 65.05%
[13]		Transactional data	RF, GB	None	GB	Accuracy: 65%
[14]		Transactional data	Customer segmentation (RFM + K-Means)	None	N/A	N/A
[19]		Transactional data	RF, SVM, XGBoost	None	XGBoost	Accuracy: 71%

Abbreviations used in the table: Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), Adaptive Boosting (AdaBoost), Decision Tree (DT), Extreme Gradient Boosting (XGBoost), Naïve Bayes (NB), Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Gradient Boosting (GB), Recency-Frequency-Monetary (RFM), Support Vector Machine (SVM), Area Under the Curve (AUC).

3. Methodology

Our study aims to investigate the reasons driving customer churn and to identify the churners. To this end, we propose a structured and adaptive sequence of steps. As illustrated in Figure 1, we begin by collecting a relevant dataset, followed by data cleaning and train–test splitting. Encoding and scaling are applied to the training set and then applied to the test set. Random oversampling is performed only on the training data to address class imbalance. Prior to this step, multiple feature selection techniques are applied on the original (non-oversampled) training set to obtain reduced feature subsets used for training machine learning and deep learning models. Model performance is finally evaluated on the unseen test set. In this section, we present a detailed description of each step, and we also highlight challenges encountered by the proposed solutions, which may support future work in churn prediction research.

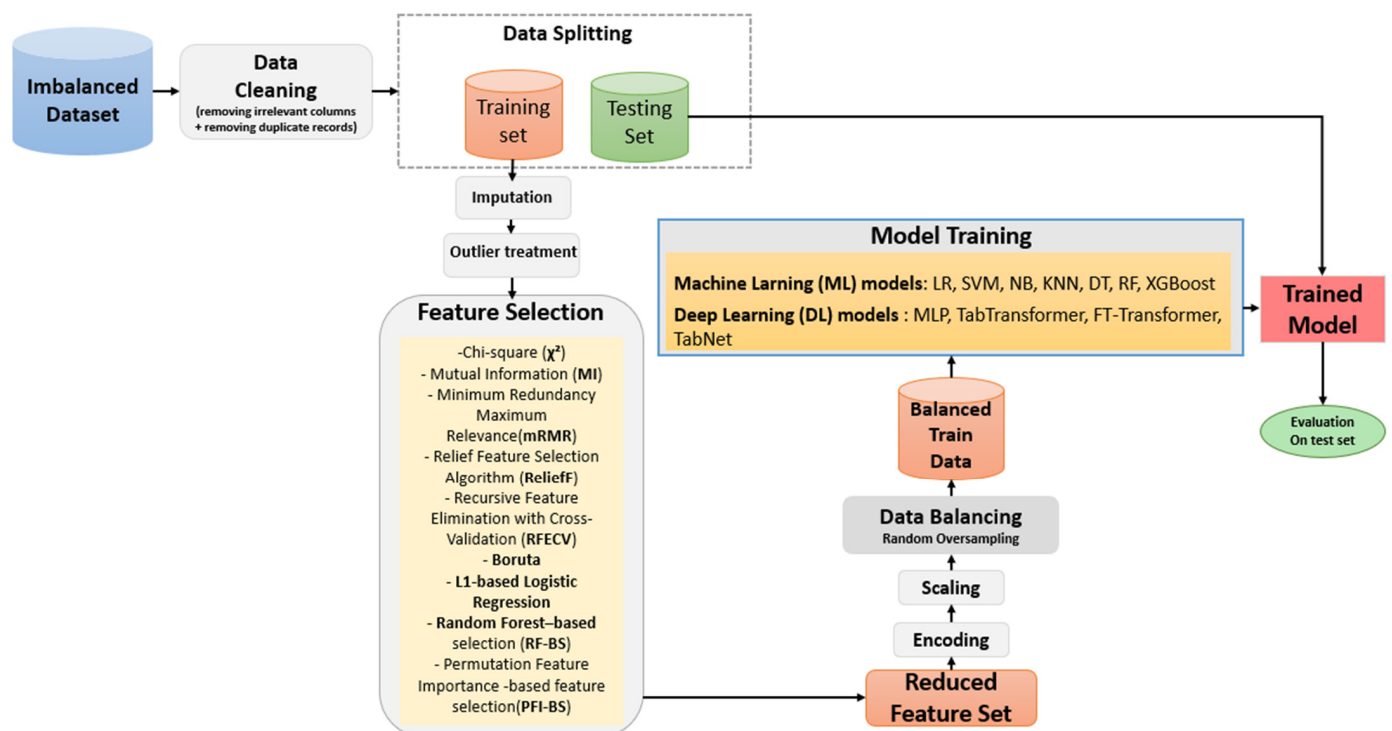


Figure 1. Churn prediction model building methodology.

3.1. Data Description

The dataset used in this study [22] is a publicly available dataset originally derived from customer interactions on a retail e-commerce platform. It includes demographic and behavioral attributes commonly used for churn prediction, which capture factors influencing whether a customer will continue their engagement with the company or not. The dataset consists of 5630 entries, with 948 churned and 4682 non-churned customers. It contains 19 features and one target variable. Among the 19 features, there is a unique identifier (ID) and 18 independent variables that describe each customer. Of these independent variables, 13 (72%) are numerical and 5 (28%) are categorical. The target variable is binary, with 0 indicating non-churn and 1 indicating churn. The data offers various factors that may lead to customer churn. We present in Table 2 the description of different variables. The primary objective of this dataset is to understand customer behavior and predict churn.

Table 2. Dataset variable Description.

Columns	Description
Customer ID	Unique identifier for each customer.
Gender	Gender of the customer (Male/Female).
Marital Status	Indicates whether the customer is single, married, or divorced.
Complaint	Total number of complaints filed by the customer.
Order Count (1 month)	Number of orders placed by the customer in the last month.
Preferred Log-In Device	The type of device used by the customer to connect to access the e-commerce platform for making purchases (e.g., mobile phone, computer).
Preferred Payment Method	The payment method preferred by the customer (e.g., Credit Card),
Cashback Amount	The monetary amount refunded or credited back to the customer after making a purchase, usually as part of a promotional offer or loyalty reward.
Warehouse To Home	The distance between the warehouse and the customer's home.
Tenure	The length of time a customer has been associated with the organization.
City Tier	It refers to a classification system that ranks cities based on their economic development, population size, and infrastructure.
Hour Spend on App	The number of hours a customer spends using the retailer's mobile app or website over a certain period.
Number Of Device Registered	The number of unique devices (e.g., smartphones, tablets, computers) a customer has used to log into their account or access the retailer's platform (app or website).
Preferred Order Category	Refers to the type of products a customer most frequently purchases or prefers to order.
Satisfaction Score	A numerical rating that represents how satisfied a customer is with their overall experience.
Number Of Address	The total number of deliveries addresses a customer has saved or used in their account.
Order Amount Hike From Last Year	The increase in the total amount spent by a customer on orders this year compared to the previous year.
Coupon Used	The number of promotional coupon or discount code the customer has applied during their purchase.
Day Since Last Order	The number of days that have passed since a customer placed their most recent order.
Churn	The target variable; it is binary, it indicates whether a customer has churned or not.

The dataset is characterized by a small size and class imbalance (Figure 2) and has a weak dependency between most features and the target label. The Mutual Information (Table 3) between each feature and the target variable is low (values < 0.13). This indicates that, individually, these features share a limited amount of information with the target, and none of them appears to be a strong predictor on its own. However, low Mutual Information values do not imply the absence of useful feature interactions. Features with weak individual relationships to the target may still contribute significantly when combined with other variables. This highlights the complexity of the churn prediction task. Therefore, the goal is not to select features randomly, but to identify those which, even if they have limited predictive power individually, contribute to the model's performance when combined.

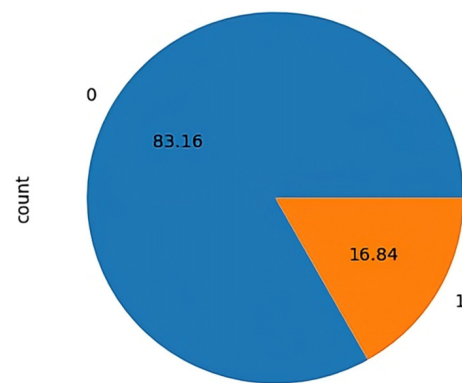


Figure 2. Proportion of Churners (1) and non-Churners (0) in the original Dataset.

Table 3. MI of the features with the target.

Feature	Mutual Information
Tenure	0.125492
CashbackAmount	0.102032
Complain	0.026001
DaySinceLastOrder	0.017684
PreferedOrderCat	0.017345
MaritalStatus	0.016339
OrderCount	0.012283
NumberOfDeviceRegistered	0.010543
NumberOfAddress	0.006497
HourSpendOnApp	0.004846
PreferredPaymentMode	0.004076
OrderAmountHikeFromlastYear	0.003237
PreferredLoginDevice	0.002203
CouponUsed	0.001322
SatisfactionScore	0.001130
WarehouseToHome	0.001129
Gender	0.000467
CityTier	0.000000

3.2. Data Preprocessing

We applied preprocessing steps to the dataset to prepare high-quality input for modeling. First, we removed irrelevant columns; specifically, the customer ID column was excluded; which serves only as a unique identifier and does not contribute to predictive modeling. Following this, we identified and removed duplicate records to prevent redundancy and ensure an unbiased model. We divided the cleaned dataset into two sets: 80% for training and 20% for testing. All preprocessing operations that required parameter estimation, including missing-value imputation, outlier handling, encoding, and feature scaling, were fitted using only the training set and subsequently applied to the test set. This procedure was adopted to prevent information leakage from the test set into the training process. The percentage of missing values in the columns containing missing data did not exceed 5%; therefore, no columns were dropped. Instead, we adopted the

following approach: for categorical features, missing values were imputed using the mode, while for numerical features, the KNN imputer was applied. As presented in Figure 3, we observed the presence of outliers, which can negatively impact model performance, particularly for sensitive models such as SVM and Neural Networks (NN). These models are highly influenced by extreme values, which may distort the learning process and lead to overfitting or reduced generalization capabilities. Therefore, to manage outliers, we employed a quantile-based method [23], which detects and removes extreme values that could negatively impact the model's performance. The split ratio was first influenced by the fact that this ratio is frequently used in machine learning research; moreover, it provides enough data for training the model and also keeps a clean portion of the data to evaluate its performance [24], so that the evaluation better reflects how the model performs in real-world situations. To preserve the balance between classes, the split was done using a stratified approach based on the target label, to ensure that the training and testing sets kept the same class distribution. For categorical features, we applied one hot encoding to convert them into a numerical format suitable for ML/DL algorithms, and numerical features were standardized to ensure comparable scales and to help models like LR, SVM, MLP, TabTransformer, FT-Transformer and TabNet to converge faster while avoiding unstable gradients. Given that the dataset was imbalanced, particularly in the churn label, we applied random oversampling techniques [25] only to the training data to ensure both classes were adequately represented during training. These preprocessing steps collectively ensured that the dataset was properly prepared for effective modeling.

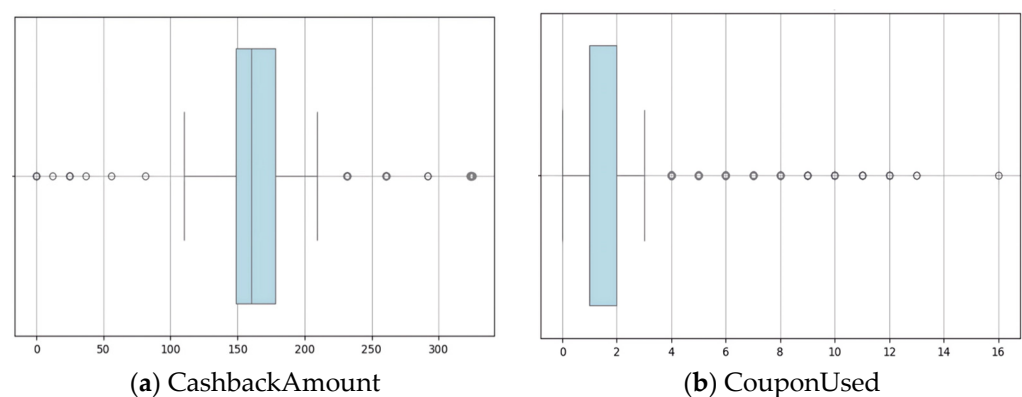


Figure 3. Detection of outliers in the dataset using an IQR-based quantile method. The box represents the interquartile range (IQR), the whiskers indicate the range of non-outlier values within $1.5 \times \text{IQR}$, and the circles denote the detected outliers.

3.3. Feature Selection

Feature selection was performed after dataset preprocessing and applied only to the training set to avoid data leakage. It was conducted on the original training set before any oversampling procedure. This ensured that feature relevance was assessed using only real observations and not influenced by duplicated samples generated during oversampling. In this study, we applied several categories of feature selection methods. For filter-based methods, we used χ^2 [26], MI [27], mRMR [28], and ReliefF [29]. For wrapper-based methods, we applied RFECV [30] and Boruta [31]. For embedded methods, we used L1-based Logistic Regression [32] and RF-BS [33]. In addition, PFI-BS [34] was applied as a model-agnostic method. These diverse approaches ensured the selection of the most relevant features for building accurate and interpretable churn prediction models.

These feature selection methods capture different types of relationships between features and the target variable, including linear and non-linear relationships as well as statistical dependencies. Some methods evaluate univariate relationships, meaning the relationship between each feature and the target variable independently, while others consider multivariate relationships, i.e., they take into account both the relationship between features and the target and the interactions among features themselves. Since feature selection methods rely on different mathematical assumptions, each method was applied using the preprocessing strategy most appropriate to its mechanism. For this reason, categorical variables were encoded using either OneHotEncoder or OrdinalEncoder, while numerical features were discretized or standardized when necessary. The optimal number of selected features was determined either automatically by the method itself or through cross-validated tuning, depending on the feature selection technique. Table 4 summarizes the feature selection methods and their corresponding configurations.

Table 4. Overview of the feature selection methods used in this study.

Method	Category	Type of Relationship Captured	Preprocessing	Model/Criterion	Hyperparameter Selection
χ^2	Filter-based	Univariate statistical dependence between each feature and the target variable	Numerical variables discretized using KBinsDiscretizer (5 bins); categorical variables encoded using OneHotEncoder	Chi-square statistic	Optimal number of selected features determined using GridSearchCV (5-fold stratified CV, Macro F1-score)
MI		Univariate statistical dependencies between feature and target	Numerical variables standardized using StandardScaler; categorical variables encoded using OneHotEncoder	mutual_info_classif	
mRMR		Multivariate dependencies between feature and target as well as among features themselves, enabling the selection of relevant predictors while reducing feature redundancy.	Categorical variables encoded using OrdinalEncoder	Mutual information relevance–redundancy trade-off	
ReliefF		Multivariate feature interactions and local neighborhood dependencies	Categorical variables encoded using OrdinalEncoder	ReliefF (k = 10 nearest neighbors)	
RFECV	Wrapper-based	Multivariate linear relationships between feature and target	Numerical variables standardized using StandardScaler; categorical variables encoded using OneHotEncoder	Logistic Regression	Iterative feature relevance testing using BorutaPy Automatic selection of the optimal number of selected features
Boruta		Multivariate non-linear relations and interactions via Random Forest	Categorical variables encoded using OrdinalEncoder	Random Forest (n_estimators = 500, class_weight = balanced)	

Table 4. *Cont.*

Method	Category	Type of Relationship Captured	Preprocessing	Model/Criterion	Hyperparameter Selection
L1-based Logistic Regression	Embedded	Multivariate linear sparse relationships	Numerical variables standardized using StandardScaler; categorical variables encoded using OneHotEncoder	Logistic Regression (L1 penalty)	Regularization parameter C optimized using GridSearchCV with stratified 5-fold cross-validation and Macro F1-score (feature selection estimator: C = 0.1; classifier: C = 1). Features were selected automatically by retaining variables with non-zero coefficients from the L1-regularized logistic regression model.
RF-BS		Multivariate non-linear relations and feature interactions	Categorical variables encoded using OneHotEncoder	Random Forest (n_estimators = 300, class_weight = balanced)	Features selected using median importance threshold
PFI-BS	Model-agnostic	Multivariate non-linear relations and feature interactions	Categorical variables encoded using OneHotEncoder	Random Forest (n_estimators = 300, n_repeats = 20)	Features retained when (importance – standard deviation) > 0

For the L1-based Logistic Regression, we independently optimized the regularization parameters of the feature selection estimator and the final classifier using GridSearchCV. We obtained different optimal values, this is due to the fact that the two models satisfy distinct objectives: the feature selection stage aims to induce sparsity and eliminate irrelevant variables through L1 regularization, whereas the final classifier aims to maximize predictive performance on the selected feature subset.

Other feature selection methods were excluded in this study for the following reasons. ExtraTrees, XGBoost, and Gradient Boosting feature importance methods were not considered because they are tree-based feature selection methods and rely on principles similar to the Random Forest importance method already applied. Sequential Feature Selection (SFS) and Exhaustive Feature Search (EFS) were excluded due to their high computational cost. In addition, statistical methods such as the ANOVA F-test were not used as they are primarily designed for numerical variables and are therefore less suitable for mixed-type tabular datasets containing both numerical and categorical features. Information Gain (IG) was not used as it is a specific case of Mutual Information and is commonly used within decision tree-based algorithms. Since tree-based classifiers such as Decision Tree and Random Forest were already included in this study, applying IG as a separate feature selection technique would provide limited additional benefit. Moreover, Mutual Information is more suitable for datasets containing both numerical and categorical variables and can capture both linear and non-linear relationships between features and the target variable. Therefore, MI provides a more general criterion for feature relevance evaluation in churn prediction tasks.

We evaluated multiple combinations of classifiers and feature selection methods. This process allowed us to identify the most effective combination adapted to the specific characteristics of our dataset and to optimize churn prediction performance.

3.4. Model Building and Selection

The dataset used in this study is tabular; which makes the selection of appropriate models and feature selection techniques particularly important to achieve high predictive performance. We considered a variety of ML models, including LR, DT, RF, SVM, NB, KNN, and XGBoost, these models are frequently used in previous studies where they have demonstrated promising results in the literature; moreover, these models are effective in classification tasks with structured data. In parallel, we included DL models, namely MLP, TabTransformer [35], FT-Transformer [36] and TabNet [37] which are capable of capturing complex patterns and feature interactions in the data. We applied Grid Search with 5-fold cross-validation (GridSearchCV) to identify the optimal hyperparameters for each model [38]. During hyperparameter optimization, random oversampling was applied only to the training portion of each cross-validation fold and never to the corresponding validation fold. This nested procedure prevented data leakage and ensured an unbiased estimation of model performance. The hyperparameter optimization process enabled effective hyperparameter tuning and improved the predictive performance. For reproducibility, a random seed of 42 was used in all experiments. The final performance of each model was then evaluated on a separate test set to ensure an unbiased estimation. The tested hyperparameter values and the corresponding optimal configurations are presented in Table 5.

Table 5. Grid search hyperparameter values and selected configurations for each machine learning and deep learning model.

Model	Hyperparameter	Search Values	Selected Value
LR	C	[0.001, 0.01, 0.1, 1, 10, 100]	0.1
	penalty	[l1, l2]	l1
	solver	[liblinear]	liblinear
DT	criterion	[gini, entropy]	entropy
	max_depth	[3, 5, 10]	10
	min_samples_split	[2, 5, 10]	2
	min_samples_leaf	[1, 2, 4]	1
RF	n_estimators	[50, 100, 200]	200
	max_depth	[5, 10, 20]	20
	min_samples_split	[2, 5, 10]	2
	min_samples_leaf	[1, 2, 4]	1
	criterion	[gini, entropy]	gini
SVM	C	[0.1, 1, 10, 100]	100
	kernel	[linear, rbf, poly]	rbf
	gamma	[scale, auto, 0.01, 0.1, 1]	0.1
NB	var_smoothing	np.logspace(−9, 0, 10)	0.1
KNN	n_neighbors	[3, 5, 7, 9]	3
	weights	[uniform, distance]	distance
	metric	[euclidean, manhattan, minkowski]	manhattan

Table 5. *Cont.*

Model	Hyperparameter	Search Values	Selected Value
XGBoost	n_estimators	[100, 200]	n_estimators = 200,
	max_depth	[3, 5, 7]	7
	learning_rate	[0.01, 0.1, 0.2]	0.2
	subsample	[0.8, 1.0]	0.8
	colsample_bytree	[0.8, 1.0]	0.8
	gamma	[0, 0.1, 0.2]	0
MLP	hidden_layer_sizes	[(50,), (100,), (100, 50)]	(100, 50)
	activation	[relu, tanh]	tanh
	solver	[adam]	adam
	alpha	[0.0001, 0.001]	0.001
	learning_rate_init	[0.001, 0.01, 0.1]	0.001
	learning_rate	[constant, adaptive]	adaptive
	max_iter	[200, 500]	500
TabTrans- former	lr	[0.001, 0.0005]	0.0005
	weight_decay	$[1 \times 10^{-5}, 1 \times 10^{-6}]$	1×10^5
	n_blocks	[2, 3, 4]	3
	d_block	[64, 128]	128
	attention_n_heads	[4, 8]	8
	attention_dropout	[0.1, 0.2]	0.1
	ffn_dropout	[0.1, 0.3]	0.1
	residual_dropout	[0.0]	[0.0]
	batch_size	[128, 256]	Default
FT- Transformer	lr	$[1 \times 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}]$	0.0005
	weight_decay	$[1 \times 10^{-5}, 1 \times 10^{-6}]$	0.0
	n_layers	[2, 3, 4]	2
	d_token	[64, 128]	128
	n_heads	[4, 8]	8
	attention_dropout	[0.1, 0.2]	0.1
	ffn_dropout	[0.1, 0.3]	0.1
	residual_dropout	[0.0]	[0.0]
	batch_size	[128, 256]	256
TabNet	n_d	[8, 16]	Default
	n_a	[8, 16]	16
	n_steps	[3, 5]	Default
	gamma	[1.3, 1.5]	Default
	lr	[0.01, 0.02]	Default

In this research, we used NVIDIA Tesla T4 Graphics Processing Units (GPUs) provided by the Google Colab Pro platform. Given the class imbalance, it would be more accurate to look at the Macro F1-score instead of accuracy to evaluate the model's performance on the test data, as it is the most representative metric for imbalanced datasets; it provides a comprehensive assessment of the model's performance across both classes (Churn and Not Churn). The F1-score is the harmonic mean of precision and recall, balancing between false positives and false negatives. Recall (or sensitivity) measures the proportion of actual positives that are correctly predicted as true positives (TP) (Equation (3)). This metric is particularly important in churn prediction, where missing a churner (high false negatives) can be costly for the company. However, focusing solely on recall may increase false positives, predicting loyal customers as churners can raise retention costs. Precision measures the proportion of predicted positive instances that are actually positive (Equation (4)), which helps in controlling these false positives. By combining both precision and recall, the F1-score (Equation (5)) provides a balanced evaluation of the model's ability to correctly identify churners while minimizing misclassification. Since the dataset is imbalanced, the Macro F1-score was used as the evaluation metric. It is computed as the average of the F1-scores obtained independently for each class and assigns equal importance to both churn and non-churn classes. This makes it particularly suitable for evaluating performance in imbalanced classification problems where both detecting churners and avoiding false churn predictions are important.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

$$\text{F1} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)$$

where:

TP = True Positives, FP = False Positives, FN = False Negatives.

4. Result and Discussion

We first conducted a preliminary experiment to obtain an initial understanding of the task complexity and model performance. To this end, we used the original dataset with only minimal technical preprocessing required for model execution, namely missing-value imputation and categorical encoding, with all 18 features included. We tested only DT and RF models to provide a solid foundation for further analysis and generalization of results. The RF model achieved a Macro F1-score of 88.35%, while the DT model reached 83.04%. These results show that the model's performance was suboptimal, which is due to many challenges related to the dataset. Specifically, the data contained noise, irrelevant and redundant features, and lacked sufficient preprocessing. In addition, the dataset is limited in size, and the dependency between other variables and the churn variable is low, which may limit the models' ability to capture complex patterns. Because of these factors, the models cannot achieve higher predictive performance, which highlights the complexity of this dataset. Therefore, there is a need for advanced preprocessing and feature selection techniques to improve overall performance in subsequent experiments. The features chosen by each technique are presented in Table 6.

Features such as Tenure, Number Of Address, Marital Status, Complaint, Warehouse To Home, Preferred Order Category were selected by most feature selection methods (Table 6), highlighting their strong predictive ability for churn detection. These features are directly related to customer behavior, engagement, and the risk of leaving the company.

Table 6. Features selected by each feature selection method.

Feature Selection Technique/ Features	χ^2	MI	mRMR	Relieff	RFECV	Boruta	L1-Based Logistic Regression	RF-BS	PFI-BS	Selection Frequency
Gender		*		*			*			3
Marital Status	*	*	*	*	*		*	*	*	8
Complaint		*	*	*	*	*	*	*	*	8
Order Count (1 month)	*	*			*			*		4
Preferred Log-In Device	*	*		*	*		*		*	6
Preferred Payment Method	*	*	*	*	*		*		*	7
Cashback Amount	*	*	*			*		*	*	6
Warehouse To Home	*	*	*		*	*	*	*	*	8
Tenure	*	*	*	*	*	*	*	*	*	9
City Tier			*	*	*		*	*		5
Hour Spend on App	*	*	*					*		4
Number Of Device Registered	*	*	*		*		*	*		6
Preferred Order Category	*	*	*	*	*		*	*	*	8
Satisfaction Score	*		*	*	*		*	*	*	7
Number Of Address	*	*	*	*	*	*	*	*	*	9
Order Amount Hike From Last Year								*		1
Coupon Used		*					*	*		3
Day Since Last Order	*	*			*			*		4

* Indicates that the feature was selected by the corresponding feature selection method.

Tenure: This variable reflects the customer's level of loyalty and attachment. The shorter the relationship between the customer and the company, the higher the risk of churn. Customers who have been with the brand for a long time often develop an emotional and behavioral connection, which makes them less likely to leave [39].

Number of Addresses: This variable may indicate the customer's level of engagement and use of the service. A customer with several delivery addresses (e.g., home, work, or family locations) may be more active and involved, which reduces their likelihood of churn.

Marital status: Marital status may be associated with differences in spending ability, purchasing habits, and loyalty, which could influence churn behavior.

Complain (number of complaints): Complaints reflect customer dissatisfaction. A customer who submits complaints frequently may be at high risk of churn, especially if their issues are not solved effectively by customer service. Authors in [40] found that customer complaints and a company's ability to effectively respond to those complaints influence customer retention.

Warehouse To Home: The distance between the warehouse and the customer's location may affect delivery time and logistics efficiency. Longer delivery distances may lead to delayed shipments and increased dissatisfaction, which could increase the likelihood of customer churn.

Preferred Order Category: This feature may reflect customer purchasing preferences and engagement with specific product segments. For example, customers who mainly

purchase highly competitive product categories may be more likely to switch platforms, increasing churn probability.

After the preprocessing steps and model hyperparameter-tuning stage, we evaluated the performance of the different feature selection techniques combined with different classifiers. This evaluation allows us to identify the most effective combination for churn prediction in the context of retail e-commerce. This is crucial for accurately detecting customers at risk of leaving the brand and informing retention strategies to better target this audience. It also provides information about which algorithms perform best with our dataset and feature selection methods, which offers practical guidance for real-world applications. Table 7 provides a summary of the comparison results based on the Macro F1-score.

Table 7. The performance of different combinations of ML and DL models and feature selection methods based on the Macro F1-score of the test set.

Classifiers/ Feature Selection	ML Models							DL Models			Tabnet (Automatic Feature Selection)
	LR	DT	RF	SVM	NB	KNN	XGBoost	MLP	Tab-Trans- former	FT- Transformer	
Before data preprocessing		83.04%	88.35%								
After data preprocessing and without feature selection	70.67%	84.45%	89.05%	86.06%	69.44%	83.23%	93.05%	92.15%	91.28%	81.26%	
χ^2	70.97%	84.55%	90.84%	87.39%	66.40%	83.81%	93.59%	91.14%	85.39%	86.31%	85.95%
MI	72.53%	85.08%	92.98%	87.22%	69.47%	85.25%	94.65%	92.04%	87.56%	85.19%	
mRMR	72.33%	84.57%	92.89%	91.56%	69.30%	86.80%	93.37%	93.29%	90.01%	89.30%	
ReliefF	72.36%	84.49%	91.55%	87.93%	70.42%	89.46%	93.25%	90.09%	87.59%	84.02%	
RFECV	72.39%	85.27%	93.34%	89.85%	69.95%	87.93%	94.48%	92.75%	89.92%	85.34%	
Boruta	70.95%	84.49%	93.14%	77.17%	66.48%	90.10%	94.65%	82.71%	75.88%	75.88%	
L1-based Logistic Regression	72.86%	85.17%	93.39%	87.98%	70.98%	89.36%	95.25%	92.22%	84.59%	88.29%	
RF-BS	71.90%	84.47%	91.12%	86.12%	67.14%	85.67%	93.37%	91.62%	86.85%	88.29%	
PFI-BS	70.93%	84.69%	92.17%	90.54%	70.74%	90.77%	94.28%	91.56%	88.29%	86.81%	

As presented in Table 7, after applying data preprocessing, the results showed a clear improvement in model performance. The RF model improved by 1.41%, while the DT model improved by 0.7%. Several other models also achieved strong performance after preprocessing alone, such as XGBoost with a Macro F1-score of 93.05 and MLP with a Macro F1-score of 92.15. These results highlight the important role of data preprocessing in improving model performance by enabling the algorithms to learn more relevant patterns and reducing prediction errors. For this reason, several feature selection techniques were evaluated to determine whether they could further improve the results.

The experimental results demonstrate that integrating feature selection techniques with classification algorithms has a significant impact on model performance. Except for Naive Bayes and SVM when combined with the Boruta algorithm, all machine learning models showed noticeable improvement after applying feature selection. Among the tested configurations, the combination of the L1-based Logistic Regression feature selection method and the XGBoost classifier achieved the best results, reaching a Macro F1-score of 95.25%.

The L1-based Logistic Regression method proved particularly effective in selecting the most informative features related to the target variable. These selected features helped

isolate the most discriminative variables for churn prediction, and XGBoost leveraged them during its boosting process to improve predictive performance. This selection process also reduced the dimensionality of the dataset, which helps minimize overfitting and improve model convergence. When the feature space becomes cleaner and contains fewer dimensions, XGBoost is better able to learn complex patterns. This could explain the strong performance of the L1-based Logistic Regression and XGBoost combination.

The next five best performances were also achieved by XGBoost with different feature selection methods, which makes it the best classifier in this case study. This is consistent with previous studies, when XGBoost has demonstrated strong performance in churn prediction and tabular datasets. Overall, XGBoost can be considered the most robust and effective classifier in this study.

Although the L1-based feature selection combined with XGBoost achieved the best performance, the improvement over mutual information-based selection is marginal (0.6%). Interestingly, Boruta combined with XGBoost achieved almost the same performance as MI combined with XGBoost, which shows that different feature selection methods can lead to comparable predictive results. Overall, all three approaches perform in a competitive way. In practice, the choice between them may depend on the trade-off between performance gain and operational complexity.

The next best performance was obtained using L1-based Logistic Regression combined with Random Forest (Macro F1-score = 93.39%), followed closely by RFECV combined with Random Forest (Macro F1-score = 93.34%). Another strong result was achieved by the combination of mRMR and MLP (Macro F1-score = 93.29%). Among the tested deep learning approaches, this configuration achieved the best performance and showed competitive results compared to the strongest machine learning models, although it remained slightly below XGBoost. These findings indicate that deep learning models do not systematically outperform classical machine learning models on this tabular dataset.

All deep learning models achieved their lowest Macro F1-scores when combined with the Boruta algorithm. This method selected the smallest number of features (5 out of 18), which reflects the conservative nature of the method, Boruta algorithm retains only variables whose importance is consistently higher than that of randomized shadow features used for comparison. This strong reduction suggests that deep learning models may not perform well when the feature space is heavily constrained. Although this aggressive selection helps remove irrelevant variables, it can also eliminate information that contributes to complex feature interactions. As a result, deep learning approaches, which generally benefit from richer feature representations, experienced a noticeable decrease in predictive performance when trained on the Boruta-selected subset.

TabNet is a deep learning architecture specifically designed for tabular data. It performs built-in automatic feature selection during training using a sequential attention mechanism that focuses on the most informative variables at each decision step. For this reason, it was not combined with external feature selection methods in this study. While a direct comparison could still be informative, the results indicate that several machine learning and deep learning classifier–feature selection combinations achieved higher Macro F1-scores than TabNet. This suggests that model-based feature selection combined with classical classifiers can be highly competitive on this dataset. However, these observations are based on average performance across cross-validation folds, and a statistical significance analysis would be required to confirm whether the observed differences are statistically meaningful.

Our findings demonstrate that there is no single feature selection method that is universally optimal; its effectiveness depends on the classifier used. While some combinations improved performance, others resulted in lower performance. For example, the Naive Bayes classifier showed reduced performance when combined with χ^2 , mRMR, Boruta, and RF-BS. These results suggest that applying a feature selection method does not necessarily guarantee improved performance.

In general, feature selection noticeably improved the performance of several models, sometimes with substantial gains. As shown in Table 8, feature selection improved performance for 9 out of 10 models, which strongly suggests that removing irrelevant or noisy variables helped the classifiers learn better decision boundaries. The strongest improvements were observed for FT-Transformer, KNN, SVM, and RF, indicating that these models particularly benefit from the removal of redundant and irrelevant features. In contrast, the performance of TabTransformer slightly decreased after applying feature selection, suggesting that it benefits from richer feature representations and may lose useful information when some features are removed.

Table 8. Comparison of classifier performance before and after feature selection based on test-set Macro F1-score.

ML/DL Model	Macro F1-Score Before Feature Selection	Best Macro F1-Score After Feature Selection	Difference
FT-Transformer	81.26%	89.30% (with mRMR)	+8.04%
KNN	83.23%	90.77% (with PFI-BS)	+7.54%
RF	89.05%	93.39% (with L1-based Logistic Regression)	+4.34%
SVM	86.06%	91.56% (with mRMR)	+5.5%
XGBoost	93.05%	95.25% (with L1-based Logistic Regression)	+2.2%
LR	70.67%	72.86% (with L1-based Logistic Regression)	+2.19%
NB	69.44%	70.98% (with L1-based Logistic Regression)	+1.54%
MLP	92.15%	93.29% (with mRMR)	+1.14%
DT	84.45%	85.27% (with RFECV)	+0.82%
TabTransformer	91.28%	90.01% (with mRMR)	−1.27%

As shown in Table 8, the L1-based Logistic Regression feature selection method helped four models achieve their best performance, highlighting its effectiveness across multiple classifiers. L1-based logistic regression performs feature selection during model training by shrinking the coefficients of features that do not contribute to the model in a meaningful way to zero through regularization. This helps remove irrelevant or redundant variables, which are common in tabular datasets. As a result, the model retains only the most informative features and can improve classification performance.

FT-Transformer improved by 8.04% with feature selection, whereas TabTransformer decreased by 1.27%. Both models share a Transformer-based architecture but exhibited markedly different behavior. One possible explanation for the different behavior of FT-Transformer and TabTransformer lies in their architectural differences. TabTransformer learns from contextual embeddings generated through self-attention across categorical features, allowing the model to capture relationships among variables. As a result, removing features through feature selection may reduce the amount of contextual information

available to the model. In contrast, FT-Transformer processes all features (categorical and numerical) as tokens within the Transformer architecture. By removing irrelevant or noisy features, feature selection helps create a cleaner input space, allowing the model to focus on the most informative variables and potentially achieve better predictive performance.

Overall, these results highlight the importance of aligning feature selection techniques with classifier characteristics. When feature selection methods are appropriately matched with learning algorithms, they can significantly improve churn prediction performance by focusing on the most informative variables.

5. Conclusions

Churn detection has become a critical goal for companies, especially in the highly competitive retail e-commerce sector. This paper investigates the impact of feature selection on improving model performance. The results indicate that feature selection plays an important role in enhancing predictive performance depending on the classifier and the characteristics of the dataset. In our work, we also developed a robust predictive model to detect customers with a high risk of churn.

Our process began with data cleaning and preprocessing, followed by testing a wide range of feature selection techniques. We then applied various machine learning and deep learning algorithms to identify the best algorithm-feature selection combination. We evaluated model performance using the Macro F1-score and compared the different combinations to identify the most effective approach. The results show that the XGBoost model combined with the L1-based Logistic Regression feature selection technique achieved the highest performance, with a Macro F1-score of 95.25%, surpassing the next-best model by 0.6% and the lowest-performing one by 28.77%. This finding highlights the crucial role of appropriate feature selection in improving predictive performance.

Apart from the features selected by most feature selection methods, the L1-based Logistic Regression, which outperformed all other techniques when combined with XGBoost, highlighted the importance of several additional variables, including Preferred Log-In Device, Preferred Payment Method, City Tier, Number Of Device Registered, Satisfaction Score and Coupon Used. These features may provide additional insights into customer behavior and may contribute to improving prediction results when used. For instance, Preferred Log-In Device, indicates the device through which customers interact with the platform, which may relate to engagement patterns and purchasing behavior. Preferred Payment Method may correlate with purchase frequency and engagement. City Tier may capture socio-economic and geographic differences that could influence purchasing power. Number Of Device Registered may indicate digital engagement, as customers using multiple devices may be more active and engaged with the brand. Satisfaction Score is often considered as an important indicator of customer loyalty. A low satisfaction score may reflect a negative customer experience and could be associated with a higher likelihood of churn, while satisfied customers are more likely to stay with the brand. Finally, Coupon Used reflects customers' responsiveness to promotional incentives. Customers who frequently use coupons may be more price-sensitive and therefore could be more likely to switch brands more when promotional offers are unavailable, making this variable a meaningful predictor of churn behavior. Collectively, these features may contain valuable information that can improve model performance.

This research makes significant findings that are valuable for both researchers and practitioners in the field of customer churn analytics. First, it provides insights into the systematic evaluation and integration of feature selection techniques with machine learning and deep learning algorithms to enhance predictive performance. Second, it identifies the best algorithm-feature selection combination. Third, it reveals the key features that

significantly impact churn. However, one limitation of this work relates to the static data used. Customer behavior is dynamic by nature: customers change their preferences, which makes purchasing patterns evolve over time, while our model is based on fixed data it may not capture the changes in customer behavior. To keep its reliability and applicability in real-world settings, the model should be updated regularly with recent data that capture emerging trends and shifts in customer behavior.

To strengthen the study, we plan to extend our approach by using and analyzing real-time data to enable early detection of churners and maintain the predictive performance of our model. While the proposed framework achieved strong results on the selected retail churn dataset, evaluating it on additional public churn datasets would provide further evidence of its robustness and generalizability across different data distributions and business contexts. This represents an important direction for future research.

Author Contributions: Conceptualization, M.C. and E.H.N.; methodology, M.C. and E.H.N.; software, M.C. and S.O.; validation, E.H.N.; investigation, M.C. and S.O.; resources, M.C.; writing—original draft preparation, M.C. and S.O.; writing—review and editing, M.C. and E.H.N.; visualization, M.C. and S.O.; supervision, E.H.N.; project administration, E.H.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available on Kaggle: <https://www.kaggle.com/datasets/nastaranzandi/customer-churn-dataset-for-a-retail-industry> (accessed on 10 November 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- De, S.; Prabu, P. Predicting customer churn: A systematic literature review. *J. Discret. Math. Sci. Cryptogr.* **2022**, *25*, 1965–1985. [CrossRef]
- Hawkins, D.M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12. [CrossRef] [PubMed]
- Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv.* **2017**, *50*, 94. [CrossRef]
- Gonda, G.; Görgényi-Hegyes, E.; Nathan, R.J.; Fekete-Farkas, M. Competitive factors of fashion retail sector with special focus on SMEs. *Economies* **2020**, *8*, 95. [CrossRef]
- Chamberlain, B.P.; Cardoso, A.; Liu, C.B.; Pagliari, R.; Deisenroth, M.P. Customer lifetime value prediction using embeddings. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 1753–1762. [CrossRef]
- Today's Daily Eggspert, Customers, Come Back! Learn How to Use Customer Retention to Your Advantage. Available online: <https://www.crazyegg.com/blog/customer-retention/> (accessed on 10 November 2025).
- What is Customer Retention? Available online: <https://www.ibm.com/think/topics/customer-retention> (accessed on 10 November 2025).
- Importance of Customer Loyalty: A Short Guide. Available online: <https://merren.io/the-importance-of-customer-loyalty/> (accessed on 2 January 2026).
- Customer Loyalty vs. Customer Retention: What's the Difference? Available online: <https://www.raven360.com/blog/customer-loyalty-vs-customer-retention> (accessed on 10 November 2025).
- Customer Churn—What It Is and Why It Matters for Every Business [Definition]? Available online: <https://blog.hubspot.com/service/what-is-customer-churn> (accessed on 10 June 2025).
- Karaarslan, H.; Baştuğ, M.; Şen, C.G.; Işık, E.E. A comparative study on customer churn analysis using machine learning and data enrichment techniques. *J. Soft Comput. Decis. Anal.* **2024**, *2*, 225–235. [CrossRef]

12. Gosain, A.; Sardana, S. Handling class imbalance problem using oversampling techniques: A review. In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 13–16 September 2017; pp. 79–85. [\[CrossRef\]](#)
13. Giri, C.; Johansson, U. Data-driven business understanding in the fashion and apparel industry. In Proceedings of the 18th International Conference on Modeling Decisions for Artificial Intelligence, Umeå, Sweden, 27–30 September 2021.
14. Bagul, N.; Berad, P.; Surana, P.; Khachane, C. Retail customer churn analysis using RFM model and k-means clustering. *Int. J. Eng. Res. Technol.* **2021**, *10*, 349–354. [\[CrossRef\]](#)
15. Fadhila, H.L.; Permadi, V.A.; Tahalea, S.P. Optimising the fashion e-commerce journey: A data-driven approach to customer retention. *Knowl. Eng. Data Sci.* **2024**, *7*, 58–70. [\[CrossRef\]](#)
16. Matuszelański, K.; Kopczewska, K. Customer churn in retail e-commerce business: Spatial and machine learning approach. *J. Theor. Appl. Electron. Commer. Res.* **2022**, *17*, 165–198. [\[CrossRef\]](#)
17. Anand, N.; Sehgal, R.; Anand, S.; Kaushik, A. Feature selection on educational data using Boruta algorithm. *Int. J. Comput. Intell. Stud.* **2021**, *10*, 27–35. [\[CrossRef\]](#)
18. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.* **2023**, *55*, 194. [\[CrossRef\]](#)
19. Patil, A.P.; Deepshika, M.P.; Mittal, S.; Shetty, S.; Hiremath, S.S.; Patil, Y.E. Customer churn prediction for retail business. In Proceedings of the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDs), Chennai, India, 1–2 August 2017; pp. 845–851. [\[CrossRef\]](#)
20. Equihua, J.P.; Nordmark, H.; Ali, M.; Lausen, B. Modelling customer churn for the retail industry in a deep learning based sequential framework. *arXiv* **2023**, arXiv:2304.00575. [\[CrossRef\]](#)
21. Agbemadon, K.B.; Couturier, R.; Laiymani, D. Churn detection using machine learning in the retail industry. In Proceedings of the International Conference on Computer, Control and Robotics (ICCCR), Shanghai, China, 8–10 April 2022. [\[CrossRef\]](#)
22. Customer Churn Dataset for a Retail Industry. Available online: <https://www.kaggle.com/datasets/nastaranzandi/customer-churn-dataset-for-a-retail-industry> (accessed on 10 November 2025).
23. Hodge, V.; Austin, J. A survey of outlier detection methodologies. *Artif. Intell. Rev.* **2004**, *22*, 85–126. [\[CrossRef\]](#)
24. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2022.
25. Shelke, M.S.; Deshmukh, P.R.; Shandilya, V.K. A review on imbalanced data handling using undersampling and oversampling technique. *Int. J. Recent Trends Eng. Res.* **2017**, *3*, 444–449. [\[CrossRef\]](#)
26. Liu, H.; Setiono, R. Chi2: Feature Selection and Discretization of Numeric Attributes. In Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, Herndon, VA, USA, 5–8 November 1995; pp. 388–391. [\[CrossRef\]](#)
27. Herman, G.; Zhang, B.; Wang, Y.; Ye, G.; Chen, F. Mutual information-based method for selecting informative feature sets. *Pattern Recognit.* **2013**, *46*, 3315–3327. [\[CrossRef\]](#)
28. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Urbanowicz, R.J.; Meeker, M.; La Cava, W.; Olson, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. *J. Biomed. Inform.* **2018**, *85*, 189–203. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [\[CrossRef\]](#)
31. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [\[CrossRef\]](#)
32. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [\[CrossRef\]](#)
33. Feature Importances with a Forest of Trees. Available online: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html (accessed on 10 June 2026).
34. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Huang, X.; Khetan, A.; Cvitkovic, M.; Karnin, Z. TabTransformer: Tabular data modeling using contextual embeddings. *arXiv* **2020**, arXiv:2012.06678. [\[CrossRef\]](#)
36. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18932–18943.
37. Arik, S.Ö.; Pfister, T. TabNet: Attentive interpretable tabular learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 6679–6687. [\[CrossRef\]](#)
38. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

39. Ascarza, E.; Neslin, S.A.; Netzer, O.; Anderson, Z.; Fader, P.S.; Gupta, S.; Hardie, B.G.S.; Lemmens, A.; Libai, B.; Neal, D.; et al. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Cust. Needs Solut.* **2018**, *5*, 65–81. [\[CrossRef\]](#)
40. Knox, G.; Van Oest, R. Customer complaints and recovery effectiveness: A customer base approach. *J. Mark.* **2014**, *78*, 42–57. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.