Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

A Data-Driven Machine Learning Framework for Behavioral Customer Churn Prediction and Personalized Marketing Targeting

Soumyajit Mukherjee

Sr Manager, Data, Analytics and AI Engineering

Corresponding Author: Soumyajit Mukherjee**, E-mail:** soumyajit.mukherjeefoo@gmail.com

ABSTRACT

Customer churn is a major challenge for e-commerce businesses because losing customers can affect long-term revenue and customer relationships. Businesses may increase customer retention and create more successful marketing campaigns by identifying customers who are likely to churn. However, variations in customer purchasing behavior and class imbalance can make accurate churn identification difficult. This study uses the Brazilian E-Commerce Public Dataset from Olist to provide an ML-based method for customer churn prediction and segmentation. The proposed approach focuses on identifying meaningful customer behavioral patterns and distinguishing between churn and non-churn customers. A Random Forest classifier is trained using an 80:20 stratified train-test split and evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The proposed model achieves 99.95% accuracy, 99.93% precision, 98.98% recall, 98.96% F1-score, and an AUC of 0.9998 on the test set. These results demonstrate the model's ability to accurately identify customer churn while maintaining a low classification error rate. The results show that ML may support data-driven customer segmentation, retention strategies, and targeted marketing in e-commerce settings by offering insightful information about customer behaviour.

KEYWORDS

Customer Segmentation, Customer Behavior, Personalized Marketing, Business Intelligence, Predictive Analytics, Customer Experience, Machine Learning.

ARTICLE INFORMATION

ACCEPTED: 04 October 2026 **PUBLISHED:** 10 November 2026 **DOI:** 10.32996/jcsts.2026.8.10

1. Introduction

In competitive digital markets, customer segmentation and targeting are essential for developing focused marketing strategies. Marketing is a fundamental component of business growth, enabling organizations to understand customer needs, communicate relevant offerings, and build long-term relationships. Segmentation splits consumers into groups based on similar features, buying behaviour, preferences, and interactions, while targeting allows for the formulation of appropriate advertisements for particular groups [1][2]. However, traditional approaches often rely on broad demographic categories and historical purchase information, which may not fully reflect changing customer behavior [3]. Opportunities for data-driven marketing have been created by the expansion of digital commerce, which has produced massive amounts of transactional and customer-related data. Behavioral segmentation uses factors such as purchase frequency, spending patterns, recency, product preferences, and customer interactions to identify meaningful customer groups. Such analysis can support relevant promotional strategies, improve customer engagement, and enable more efficient allocation of marketing resources [4]. However, standard approaches may encounter obstacles when processing large-scale and complex consumer data [5].

Artificial Intelligence (AI) and ML provide advanced capabilities for analyzing complex customer datasets and discovering behavioural patterns [6]. AI-based approaches can support dynamic segmentation, customer profiling, personalized content, campaign planning, and prediction of customer preferences [7][8]. These techniques enable organizations to move beyond static segmentation and utilize data-driven insights for developing more responsive and personalized marketing strategies [9]. This study presents an ML-based approach for behavioral customer churn prediction and personalized marketing targeting using the Brazilian E-Commerce Public Dataset by Olist. This research offers several key contributions as follows:

The study provides an effective data-driven solution for identifying customer churn patterns in e-commerce.

It highlights the importance of customer purchasing and behavioral characteristics in understanding churn.

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

It demonstrates highly accurate discrimination between churn and non-churn customers.

It offers useful insights for customer retention, segmentation, and personalized marketing decisions.

The novelty of this study lies not in the development of a new machine learning algorithm, but in the integration of multidimensional customer behavioural information within an applied framework for churn-risk interpretation and marketing decision support. Unlike approaches that treat churn prediction and customer profiling independently, the proposed framework combines purchase frequency, monetary value, payment behavior, product interaction, and delivery experience to identify behavioral characteristics associated with customer churn. These churn-risk insights can subsequently be translated into actionable customer targeting and personalized marketing strategies. Thus, the framework provides an integrated data-driven perspective linking behavioral analysis, churn prediction, customer segmentation, and marketing action in an e-commerce context.

1.1 Paper Structure

The paper is structured as follows: Section II presents a review of existing studies on customer segmentation and personalized marketing. Section III describes the dataset, pre-processing techniques, and proposed methodology. Section IV discusses the experimental outcomes and model comparisons, while Section V summarizes the findings and outlines potential directions for future work.

2. Literature Review

A thorough evaluation of current studies on Behavioural Customer Segmentation was undertaken to identify existing techniques, results, and research gaps that are relevant to this study.

Nigam, Mishra, and Maravi (2026) use the UCI Online Retail dataset, which is preprocessed and transformed into customer-level monthly behavioral features that support static segmentation and dynamic prediction. The experimental results show that Autoencoder+KMeans improves the silhouette score from 0.3770 when using traditional RFM-based K-means to 0.5706, which indicates better cluster segmentation. For dynamic prediction, the tuned XGBoost model acquired an accuracy of 0.7858, outperforming the Markov baseline accuracy of 0.7413 [10]. Khaing et al. (2026) propose a data-driven marketing framework that integrates ML techniques to analyze and predict customer attraction toward a health-oriented beverage brand (Brand X). Five supervised classification algorithms–RF, Gradient Boosted Trees, KNN, Deep Learning, and Logistic Regression–were evaluated using acc, prec, rec, and F1-score metrics. The RF model outperformed all others in terms of accuracy (84.52%) and F1-score (83.9%), suggesting that it is very good at making predictions and is robust to imbalanced marketing data [11].

Garidzira, Matobobo, and Vambe (2025) use the publicly available "Shop Customer Data" dataset from Kaggle. The Silhouette Scores were utilized to evaluate the clusters' quality. Agglomerative Clustering achieved the lowest Silhouette Score (0.1042), while Spectral Clustering performed best (0.1297), closely followed by GMM (0.1283) and KMeans (0.1281) [12]. Tiwari et al. (2025) tested eight different clustering techniques—K-Medoids, DBSCAN, HDBSCAN, BIRCH, GMM, K-Means, Agglomerative Clustering, and Affinity Propagation— to put bank clients into groups depending on how much money they have in their accounts and how often they make transactions. According to their findings, K-Means had the highest Silhouette Score of 0.8606, while BIRCH and Agglomerative Clustering had an outcome of 0.8528. The Calinski-Harabasz Index gave Affinity Propagation a score of 6904.06, but DBSCAN and HDBSCAN produced meaningless clusters and so had missing values in the assessment [13]. Xu (2025) presents research that focuses on applying K-means clustering as an effective method of precision marketing on e-commerce platforms. The proposed model uses K-means clustering to efficiently segment customers into significant groups, thus allowing businesses to launch targeted campaigns with greater relevance. Experimental outcomes show significant improvement over legacy approaches, with an accuracy of 98.2%, an F1-score of 97.7%, and improved segmentation quality. Comparative evaluation shows that K-means greatly surpasses legacy models in terms of precision and recall [14].

Chakraborty et al. (2023) analyze historical sales and customer interaction data of the business to derive the right attributes. Their study used K-Means clustering, an unsupervised machine learning algorithm, to obtain five separable clusters with a silhouette score of 0.57. For product recommendation, a voting ensemble-based recommendation system has been utilized to suggest up to the top 10 products for each customer. The ensemble method consists of four techniques: user-based collaborative filtering, the LightFM algorithm, the KNN With Means algorithm, and a neural network [15].

Despite substantial progress in customer segmentation, existing studies mainly focus on either static clustering or individual prediction and recommendation tasks. Most approaches rely on conventional clustering algorithms with limited integration of behavioral segmentation, customer prediction, and personalized marketing targeting within a unified framework. In addition, several studies use small or domain-specific datasets, limiting the generalizability of their findings. A comparative summary of the existing studies, datasets, methods, and key findings is presented in Table I. Therefore, there is a need for a data-driven AI framework that integrates customer behavioral features, robust segmentation, and predictive targeting to support personalized marketing decisions in e-commerce environments.

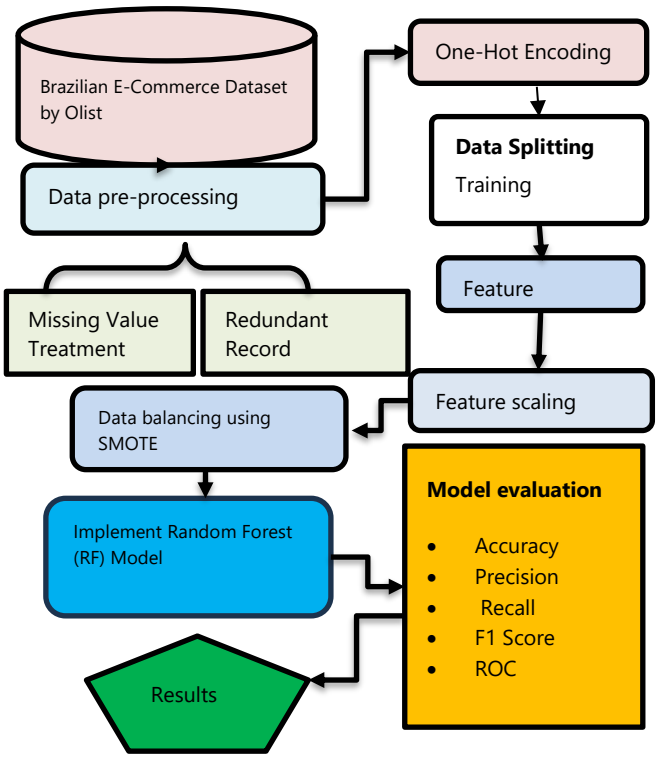
Recent Machine Learning Studies on Customer Behavior and Marketing

Author	Approach	Dataset	Key Findings	Limitations & Future Work
Nigam, Mishra & Maravi (2026)	AI-driven dynamic customer behavioral intelligence framework using an Autoencoder, K-Means, Markov Transition Analysis, and XGBoost for next-segment prediction.	UCI Online Retail Dataset; customer-level monthly behavioral features	Autoencoder + K-Means improved Silhouette Score from 0.3770 to 0.5706. Tuned XGBoost achieved 78.58% accuracy, outperforming the Markov baseline of 74.13%.	Further validation on larger and more diverse retail datasets is needed; future work can improve real-time personalization and deployment scalability.
Khaing et al. (2026)	Data-driven marketing framework using supervised ML to predict customer attraction toward a health-oriented beverage brand.	Survey dataset; 219 records after SMOTE balancing	Random Forest achieved the best performance with 84.52% accuracy and 83.9% F1-score.	Small survey dataset and brand-specific context limit generalization; future work can use more diverse datasets.
Garidzira, Matobobo & Vambe (2025)	Comparative customer segmentation using K-Means, GMM, Spectral, and Agglomerative Clustering.	Kaggle Shop Customer Data, 1,965 customers	Spectral Clustering obtained the highest Silhouette Score of 0.1297, followed by GMM (0.1283) and K-Means (0.1281).	Cluster separation was relatively moderate; future work can incorporate richer behavioral and transaction-level features.
Tiwari et al. (2025)	Comparative evaluation of eight clustering techniques for grouping bank clients according to financial and transaction behavior.	Bank customer/transaction dataset	K-Means obtained the highest Silhouette Score of 0.8606; BIRCH and Agglomerative Clustering achieved 0.8528.	Some density-based methods produced meaningless clusters; future research can incorporate additional behavioral variables and advanced hybrid approaches.
Xu (2025)	K-Means-based precision marketing framework for customer segmentation	E-commerce customer data	Reported 98.2% accuracy and 97.7% F1-score, with improved segmentation and marketing performance	Further testing across different e-commerce environments and larger datasets is required.
Chakraborty et al. (2023)	combining K-Means clustering with an ensemble recommendation system for personalized product suggestions.	Historical sales and customer interaction data	K-Means generated five clusters with a 0.57 Silhouette Score. The ensemble recommender provided up to 10 product recommendations per customer.	Future work can integrate real-time behavioral data and more advanced deep-learning recommendation techniques.

3. Research Methodology

The proposed methodology uses the Olist Brazilian E-Commerce dataset for customer churn prediction. The data are preprocessed by handling missing values, removing duplicates and redundant attributes, and encoding categorical features using frequency and one-hot encoding. The dataset is divided into stratified training and testing sets using an 80:20 ratio. Feature engineering and Random Forest-based feature selection are applied, followed by Min–Max normalization and SMOTE-based balancing on the

training data. A Random Forest classifier is then trained using optimized parameters, and its performance is evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. Fig. 1 shows the proposed flowchart.

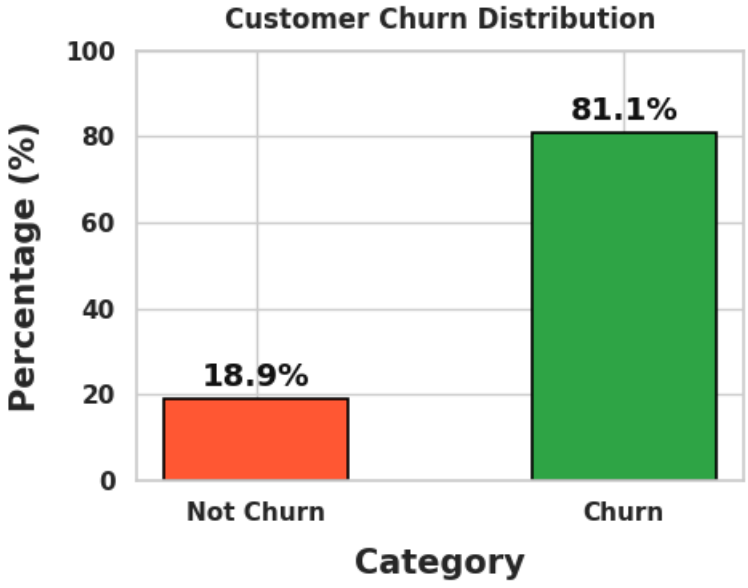


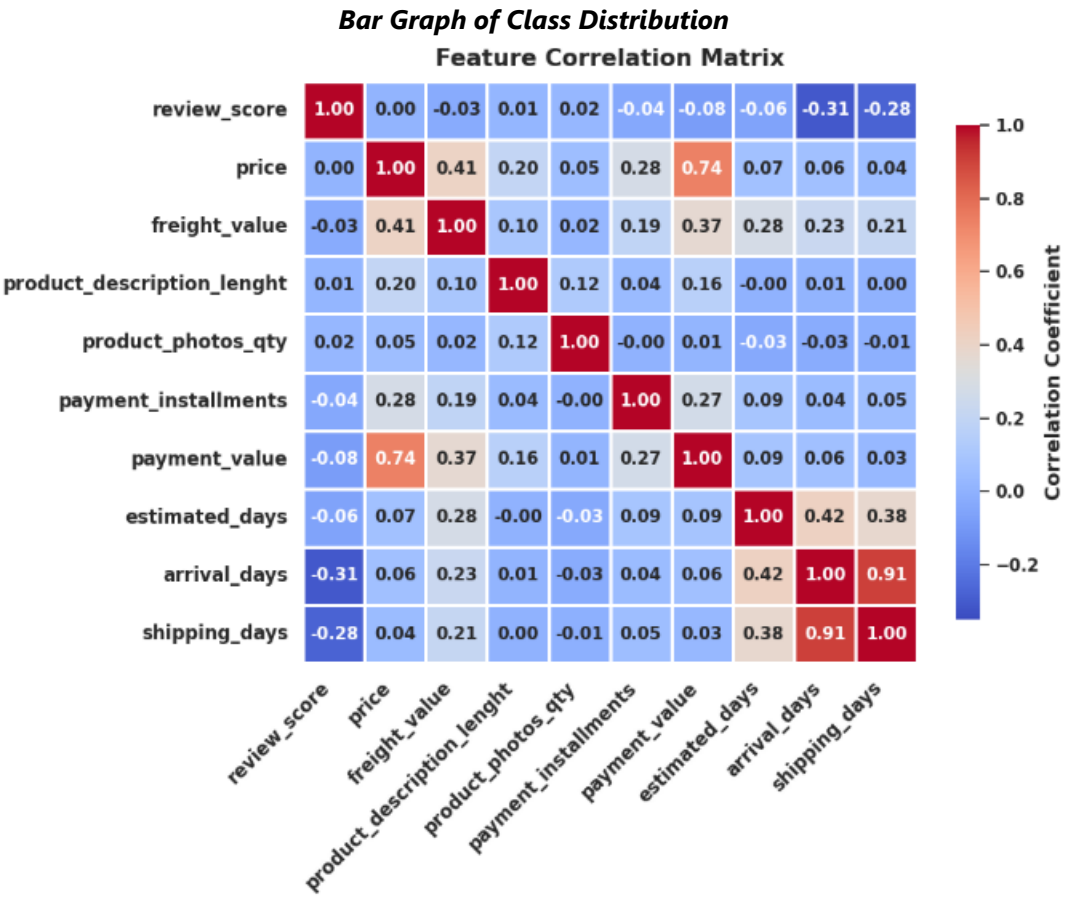
3.1 Proposed Flowchart for Intelligent Customer Segmentation

The next section lays out the proposed framework's stages in great detail:

3.2 Data Gathering and Analysis

The data for this study are accessible on Kaggle, provided by the Brazilian e-commerce platform Olist Shops. The dataset includes 25 features, covering transactional information such as order details, payment methods, delivery performance, customer feedback, and product attributes, all directly gathered from Olist's online platform. The customer churn distribution indicates a highly imbalanced dataset, where 81.1% of customers belong to the churn class and only 18.9% belong to the non-churn class, as presented in Fig. 2. This imbalance suggests that a significantly larger proportion of customers are leaving the service, highlighting the importance of effective churn prediction and class-balancing techniques during model development.





3.3 Correlation Matrix of Features

Fig. 3 illustrates the relationships among the variables utilized in the dataset. arrival_days and shipping_days have a strong positive correlation (0.91), while payment_value is strongly correlated with freight_value (0.74). Review_score is negatively correlated with arrival_days (-0.31) and shipping_days (-0.28), indicating that longer delivery times tend to reduce customer satisfaction. Most other feature pairs exhibit weak correlations, suggesting low multicollinearity among the variables.

The churn label was defined independently of the predictor variables utilized for model training. The target variable was constructed using the predefined churn criterion, while the predictor variables were derived from customer purchasing and behavioral characteristics. Thus, the information utilized to determine the churn label was not included as an input feature, minimizing the risk of target leakage.

3.4 Data Pre-processing

The Brazilian E-Commerce (Olist) dataset was pre-processed by handling missing values, removing duplicate records, and reducing redundant information. Rows containing Missing Values were removed to maintain data integrity with minimal distortion, while the original datetime columns were excluded to avoid redundancy. High-cardinality categorical variables were transformed using frequency encoding to preserve their distributional information, whereas the remaining Categorical Variables were converted using one-hot encoding for effective model processing. The training data alone were used to fit categorical encoding and Min–Max scaling, which were then applied to the test data. As a result, pre-processing did not utilise any data from the test set.

3.5 Data Splitting

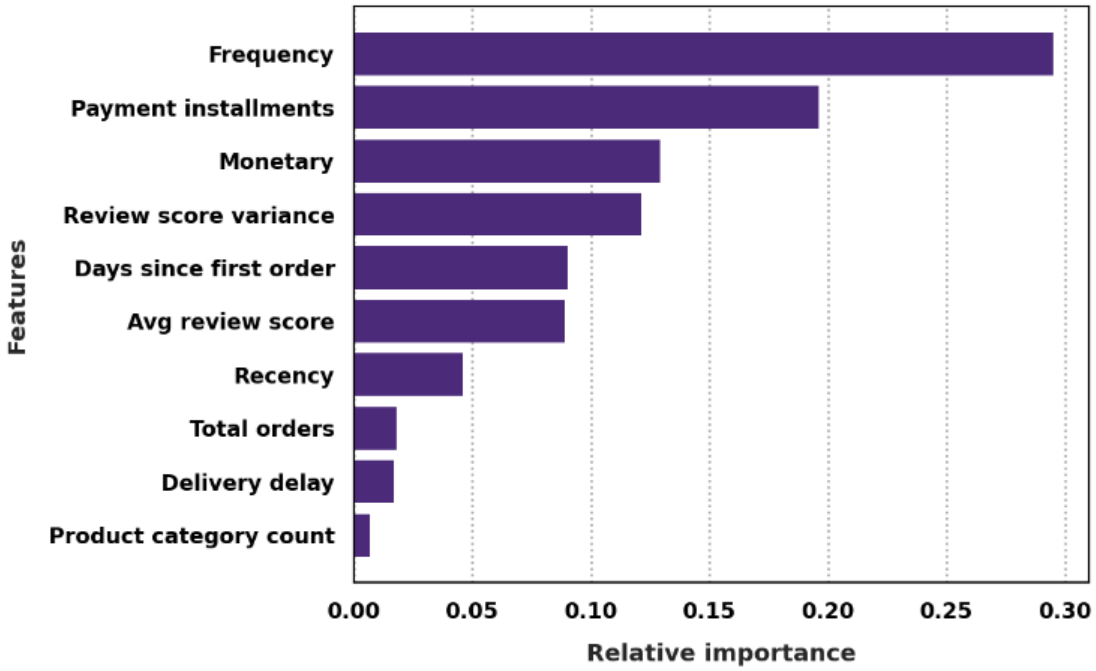
A stratified 80:20 train-test split ensures that both subsets maintain a class distribution proportional to that of the original dataset. This approach provides balanced representation of each class in the training and testing sets, supporting reliable model evaluation.

3.6 Data Leakage Prevention

To minimize the risk of data leakage, the features used for model training were examined to ensure that no feature directly or indirectly encoded the churn target. The target variable was excluded from the input feature set before model training. In addition, customer-level records were separated between the training and testing datasets to prevent the same customer or near-duplicate transactional patterns from appearing in both sets. These steps were applied to ensure that the reported model performance reflects genuine predictive capability rather than information leakage between the training and testing data.

3.7 Feature selection using Random Forest

Feature importance was estimated from the trained Random Forest using impurity-based feature importance. Feature selection was done only using the training set. The selected features were then used for the test set to prevent information leakage. Features with low relative importance were removed, retaining the behavioral variables that contributed most strongly to churn classification. Fig. 4 presents the feature importance of the selected behavioral variables, with Frequency emerging as the most influential feature at approximately 0.30. Payment installments and Monetary value also showed relatively high importance, whereas Product category count and Delivery delay had lower contributions. Overall, the results indicate that customer interaction and purchasing behavior contributed more strongly to churn classification than product and delivery-related variables.



3.8 Feature Importance of Customer Behavioral Features

3.9 Min-Max Normalization

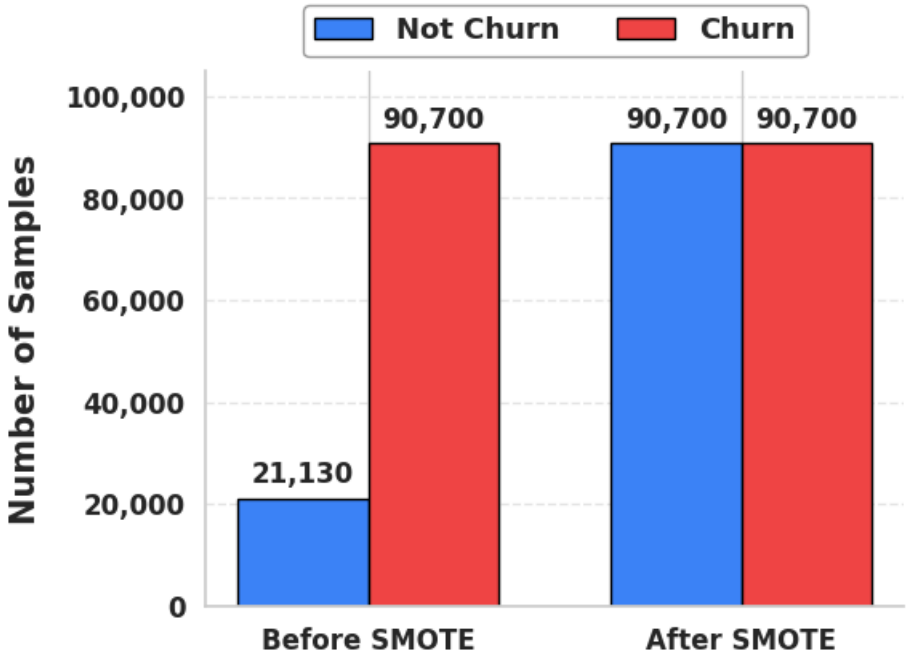
To bring the numerical attributes to a common scale, Min–Max normalization is applied to transform the feature values into the interval [0, 1]. The stability and effectiveness of the classification models are enhanced by decreasing the dominance of characteristics with higher numerical ranges in the learning process via this scaling process. The normalised value of each characteristic is derived using the mathematical formula in Equation (1):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X represents the original value of the feature, X' is the normalized value, X_{min} is the minimum value of the feature, and X_{max} is the maximum value of the same feature.

3.10 Data balancing using SMOTE

Data balancing is an important preprocessing technique used to reduce class imbalance and ensure that ML models receive sufficient representation from each class. In this study, SMOTE is employed to increase the representation of the minority class. SMOTE identifies neighboring samples within the minority class and generates new observations by interpolating between selected samples. SMOTE was applied only to the training data after the train-test split, while the test data remained unchanged for unbiased model evaluation. The training dataset is updated with the produced samples, leading to a more even distribution of classes and facilitating more trustworthy model learning and classification results.



3.11 Class Distribution Before and After SMOTE Balancing

Fig. 5 illustrates the class distribution before and after applying SMOTE. Before SMOTE, the dataset was highly imbalanced, with 21,130 non-churn samples and 90,700 churn samples. After SMOTE, the minority class (non-churn) was oversampled to 90,700 samples, resulting in a balanced dataset with equal representation of both churn and non-churn classes. This balancing process helps reduce model bias toward the majority class and improves the reliability of churn prediction.

3.12 Random Forest (RF) Model

RF is an ML ensemble approach that uses a combination of DTs to enhance classification accuracy while decreasing the likelihood of overfitting. The model uses randomly chosen subsets of training samples and features to create many independent decision trees. Each tree makes an individual forecast, and the final categorization result is chosen by majority vote among all created trees. The RF prediction function can be represented as Equation (2).

$$\hat{Y} = mode\{T_1(X), T_2(X), T_3(X), \dots\dots, T_N(X)\} \quad (2)$$

where $\hat{Y}$ represents the final predicted class, Ti (X) denotes a prediction of the ith decision tree, X represents an input feature vector, and n indicates the total number of trees in the forest. The ensemble learning mechanism of RF enhances model stability, improves generalization capability, and provides effective classification performance for churn prediction. The probability estimation in the RFC is calculated by aggregating the outputs of all individual trees, which can be expressed as Equation (3):

$$P(Y = c|X) = \frac{1}{N}\sum_{i=1}^{N} I(T_i(X) = c) \quad (3)$$

where P(Y=c|X) represents the probability of class c, N is the number of decision trees, and I (·) is an indicator function that returns 1 when the condition is satisfied. This probability-based decision mechanism enables RF to accurately distinguish between churn and non-churn customers.

The proposed Random Forest model is configured with 100 decision trees (n_estimators = 100), a maximum tree depth of 20 (max_depth = 20), and a minimum of 2 samples required for splitting (min_samples_split = 2). The minimum number of samples per leaf is set to 1 (min_samples_leaf = 1), with Gini impurity (criterion = "gini") used as the splitting criterion. The model employed sqrt feature selection (max_features = "sqrt") and a random state of 42 (random_state = 42) to ensure reproducible results.

3.13 Evaluation Metrics

The effectiveness of the proposed model is evaluated using several standard performance metrics derived from the confusion matrix. The matrix summarizes correct and incorrect predictions through True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). The equations for calculating accuracy, precision, recall, and F1-score are presented in Equations (4)-(7), respectively.

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+Fp+TN+FN}} \quad (4)$$

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (5)$$

$$Recall = \frac{\text{TP}}{TP+FN} \quad (6)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

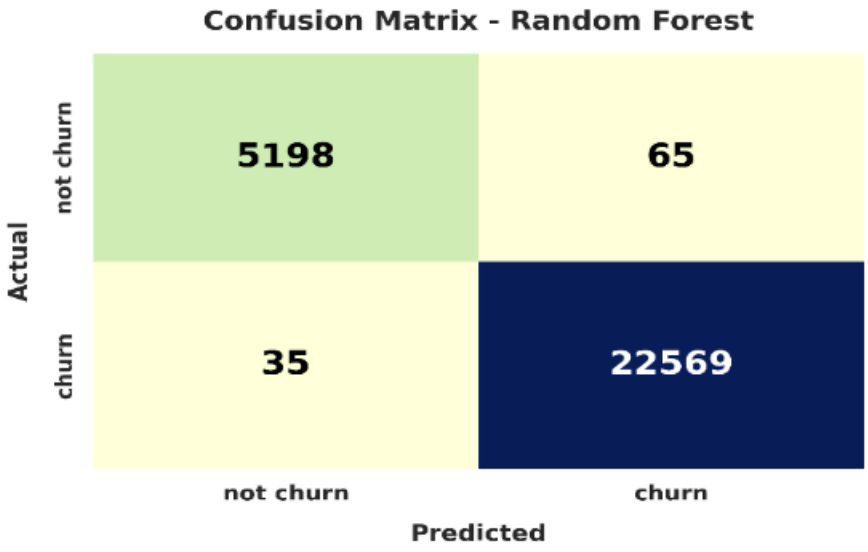
The percentage of input samples that were properly categorized is known as accuracy. Precision is the ratio of the number of positive cases that were accurately predicted to the total number of positive instances. The percentage of true positives that the model properly identifies is called recall. F1-score provides the harmonic mean of precision and recall, balancing both measures on a scale from 0 to 1. The ROC curve illustrates the trade-off between the TPR and FPR across different classification thresholds, where TPR corresponds to sensitivity or recall and FPR equals 1-specificity.

4. Results and Discussion

All experiments were performed on a high-performance system with an Intel Core i7-12700K CPU (12 cores, 3.6 GHz), 32 GB of DDR4 RAM, an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM), and a 1 TB NVMe SSD. Python 3.10, SQL, scikit-learn, Pandas, Matplotlib, Power BI Desktop, and Microsoft SQL Server were used on Windows 11 Pro for data processing, machine learning, segmentation, churn prediction, and interactive dashboard development. The Brazilian E-Commerce (Olist) dataset is utilized to train the suggested Random Forest model, which is then assessed using acc, prec, rec, and F1-score, as shown in Table II. The model achieved 99.95% accuracy, 99.93% precision, 98.98% recall, and 98.96% F1-score, indicating highly consistent classification performance. Overall, these results demonstrate the model's ability to effectively distinguish customer segments and support data-driven marketing analysis.

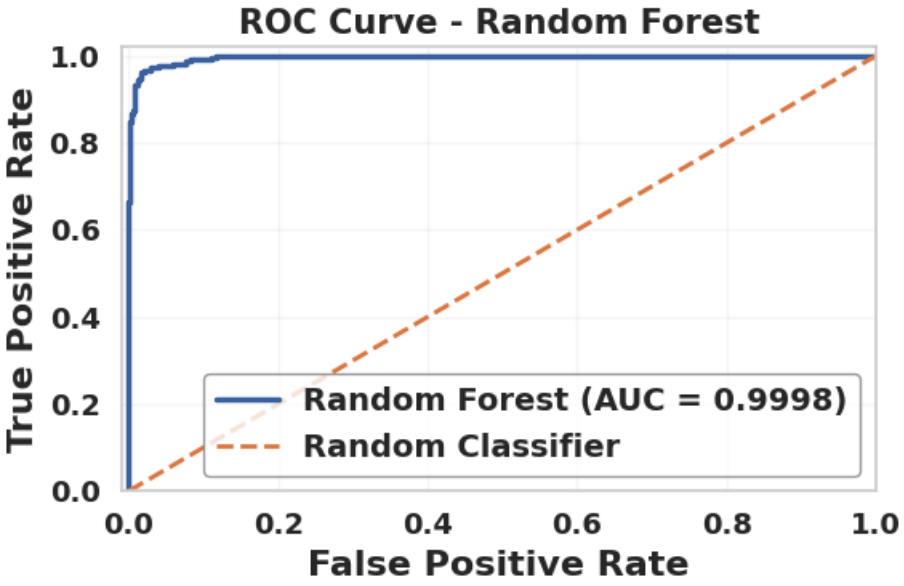
Classification results of the proposed model for Customer Churn Prediction

Metric	Testing	Training
Accuracy	99.95	100
Precision	99.93	100
Recall	98.98	100
F1-score	98.96	100



Confusion Matrix for the proposed RF Model

Fig. 6 evaluates the Random Forest churn prediction model. The model correctly classifies 22,569 churn cases (True Positives) and 5,198 non-churn cases (True Negatives). It records only 65 False Positives and 35 False Negatives, indicating very few classification errors. Overall, the results suggest that the model can accurately distinguish between churn and non-churn customers.



4.1 ROC Curve for Proposed RF Model

Fig. 7 illustrates the classification performance of the proposed RF model in distinguishing customer classes. The curve remains close to the upper-left corner, indicating a high true-positive rate with a low false-positive rate. The model achieves an AUC of 0.9998, demonstrating excellent classification performance. Overall, the ROC curve confirms that the RF model effectively differentiates between the target customer classes.

4.2 Comparative Analysis

Table III presents a comparison of the suggested Random Forest model with various ML models in order to evaluate its efficacy. KNN, XGBoost, and Logistic Regression achieved varying performance across their respective datasets, while K-Means was evaluated as an unsupervised segmentation method. The proposed RF achieved 99.95% accuracy, showing strong performance on the proposed dataset and preprocessing pipeline. The compared studies use different data-cleaning, feature-engineering, transformation, and scaling procedures, which can affect model performance. Therefore, the reported results provide a general cross-study comparison rather than a direct comparison under identical experimental conditions.

4.3 Comparison of Different ML Models for Customer Churn Prediction

Model	Accuracy	Precision	Recall	F1-score
KNN[16]	67	74	67	70
XGBoost[17]	81	79	83	81
LR [18]	89.11	71.47	48.55	57.70
K-Means [19]	-	85	80	82
Proposed RF	99.95	99.93	98.98	98.96

This research is significant because it provides a data-driven approach to understanding customer churn and behavioral patterns in e-commerce. By identifying customers at risk of leaving, the study can support timely retention efforts and more personalized customer engagement. The findings also provide insights into purchasing behavior, customer preferences, and delivery-related experiences that influence customer relationships. The strong predictive performance demonstrates the potential of ML to transform transactional data into actionable business insights, supporting customer segmentation, targeted marketing, improved retention, and more informed business decision-making.

5. Conclusion and Future Research

Customer segmentation is critical to corporate success because it provides detailed insights into customer behaviours and preferences, allowing for more effective and personalised marketing campaigns. However, traditional segmentation methods often fall short of accurately analyzing large and complex datasets. This research paper presents an ML-based framework for intelligent customer segmentation and personalized marketing. The proposed RF model achieved an accuracy of 99.95%. Its reported performance was higher than the values reported for several models in related studies, although direct comparison is limited by differences in datasets and experimental settings. These findings show that the proposed technique is effective in identifying churn and non-churn customers and can provide data-driven insights for personalized marketing and customer retention strategies.

5.1 Limitations and Future Work

The study is limited by the use of a single Brazilian e-commerce dataset, which may not fully represent customer behavior across different markets, platforms, and demographic groups. The analysis is also based on historical transactional data and therefore

may not capture changes in customer preferences over time. Future work can validate the framework on multiple e-commerce datasets and incorporate behavioral and temporal features. Advanced ensemble methods, deep learning, explainable AI, and real-time churn prediction can also be explored to improve customer retention and personalized marketing. Future research will include a controlled comparison using the same Olist dataset, train-test split, and preprocessing pipeline to provide a fair and consistent evaluation of model performance.

Funding: This research received no external funding.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers.

References

[1] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models," Algorithms, vol. 17, no. 6, 2024, https://doi.org/10.3390/a17060231.

[2] A. Manzoor, M. Atif Qureshi, E. Kidney, and L. Longo, "A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners," IEEE Access, vol. 12, no. May, pp. 70434–70463, 2024, https://doi.org/10.1109/ACCESS.2024.3402092.

[3] X. Shi, Y. Zhang, M. Yu, and L. Zhang, "Revolutionizing market surveillance: customer relationship management with machine learning," PeerJ Comput. Sci., vol. 10, p. e2583, Dec. 2024, https://doi.org/10.7717/peerj-cs.2583.

[4] G. Dash, C. Sharma, and S. Sharma, "Sustainable Marketing and the Role of Social Media: An Experimental Study Using Natural Language Processing (NLP)," Sustainability, vol. 15, no. 6, p. 5443, Mar. 2023, https://doi.org/10.3390/su15065443.

[5] M. Barari, S. Quach, and P. Thaichon, "New developments in artificial intelligence (AI)-powered products in marketing," in Artificial Intelligence for Marketing Management, London: Routledge, 2022, pp. 55–75. https://doi.org/10.4324/9781003280392-7.

[6] N. S. Jayawardena, A. Behl, P. Thaichon, and S. Quach, "Artificial intelligence (AI)-based market intelligence and customer insights," in Artificial Intelligence for Marketing Management, London: Routledge, 2022, pp. 120–141. https://doi.org/10.4324/9781003280392-10.

[7] T. Davenport, A. Guha, D. Grewal, and T. Bressgott, "How artificial intelligence will change the future of marketing," J. Acad. Mark. Sci., vol. 48, no. 1, pp. 24–42, Jan. 2020, https://doi.org/10.1007/s11747-019-00696-0.

[8] N. Gautam and N. Kumar, "Customer segmentation using k-means clustering for developing sustainable marketing strategies," Bus. Informatics, vol. 16, no. 1, pp. 72–82, Mar. 2022, https://doi.org/10.17323/2587-814X.2022.1.72.82.

[9] A. Hemalatha, AI-Driven Marketing: Leveraging Artificial Intelligence for Enhanced Customer Engagement. Jupiter Publications Consortium, 2023. https://doi.org/10.47715/JPC.B.978-93-91303-61-7.

[10] Y. Nigam, N. Mishra, and Y. P. S. Maravi, "*AI-Driven Dynamic Customer Behavioral Intelligence Framework Using Deep Learning*," in 2026 International Conference on Emerging Trends in Information, Communication & Systems (ICETICS), IEEE, Jul. 2026, pp. 1–6. https://doi.org/10.1109/ICETICS69505.2026.11651121.

[11] P. K. Khaing, T. Kaewlarksee, P. Choeiban, N. Kurubanjerdjit, and S. Nupap, "Integrating Machine Learning into Data-Driven Marketing: A Case Study of Beverage Brand Consumer Segmentation," in 2026 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering (ECTI DAMT & NCON), IEEE, Feb. 2026, pp. 66–71. https://doi.org/10.1109/ECTIDAMTNCON67592.2026.11460056.

[12] T. C. Garidzira, C. Matobobo, and W. T. Vambe, "Customer Segmentation in E-Commerce Through Data-Driven Customer Behavior Analysis Employing Unsupervised Machine Learning Models," in 2025 5th International Multidisciplinary Information Technology and Engineering Conference (IMITEC), IEEE, Nov. 2025, pp. 1–7. https://doi.org/10.1109/IMITEC67386.2025.11410443.

[13] S. Tiwari, G. Vijh, N. Sharma, S. Vijh, S. Kumar, and K. Rastogi, "Financial Institution Customer Segmentation Using Unsupervised Learning – A Data-Driven Approach to Customer Profiling," in 2025 2nd Global AI Summit - International Conference on Artificial Intelligence and Emerging Technology (AI Summit), IEEE, Nov. 2025, pp. 90–95. https://doi.org/10.1109/AISummit66170.2025.11411318.

[14] M. Xu, "Data-Driven Customer Segmentation in E-Commerce using K-Means Clustering for Personalized Marketing Strategies," in 2025 3rd International Conference on Data Science and Information System (ICDSIS), 2025, pp. 1–8. https://doi.org/10.1109/ICDSIS65355.2025.11071130.

[15] A. Chakraborty, K. Dey, S. Ghosh, R. Chakraborty, I. Mitra, and P. Nandy, "A Novel Approach for Customer Segmentation and Product Recommendation to Boost Sales using Machine Learning," in 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), 2023, pp. 97–103.

[16] M. Ehsan, "Predictive Modeling for Customer Segmentation in E-Commerce," no. July, 2024.

[17] M. Zeinali, L. Ramezani Asli, and M. A. Khalili, "Integrating Business Intelligence and CRM Systems With a Machine Learning Approach for Predictive Customer Retention in E-Commerce," Sci. World J., vol. 2026, no. 1, 2026, https://doi.org/10.1155/tswj/1946904.

[18] M. Chaudhary, M. A. Alam, and S. Zafar, "Leveraging Machine Learning for Customer Segmentation and Personalized Client Targeting in E-commerce Systems," Recent Adv. Electr. Electron. Eng. (Formerly Recent Patents Electr. Electron. Eng., 2025, https://doi.org/10.2174/0123520965380645250505070035.

[19] S. Liu and S. Yang, "Machine Learning-Based Market Segmentation and Consumer Behavior Prediction Models," in Proceedings of the International Conference on Decision Science & Management, New York, NY, USA: ACM, Apr. 2024, pp. 122–126. https://doi.org/10.1145/3686081.3686100.