



Hybrid Learning for Automated E-commerce Churn Prediction with XGBoost and SHAP

Amir Acalapati Henry^{1*}, Abdul Hamid Arribathi², Henderi Henderi³

^{1,2,3}Faculty of Computer and Science, University of Raharja, Indonesia

E-Mail: ¹amir.acalapati@raharja.info, ²abdulhamid@raharja.info, ³henderi@raharja.info

Received May 23th 2026; Revised Jun 20th 2026; Accepted Jul 13th 2026; Available Online Jul 28th 2026

Corresponding Author: Amir Acalapati Henry

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Customer retention is a critical challenge in the e-commerce industry, yet many platforms often lack explicit labels to identify potential defectors (churn). This research proposes a hybrid unsupervised-supervised learning framework to perform automated churn classification using the high-dimensional Olist Brazilian E-commerce dataset. The first stage employs K-Means Clustering to objectively generate churn labels based on Recency, Frequency, and Monetary (RFM) features. The second stage applies an Extreme Gradient Boosting (XGBoost) model, augmented by the Synthetic Minority Over-sampling Technique (SMOTE), to address the inherent class imbalance typical of transactional data. Experimental results demonstrate that the proposed model achieved a robust accuracy of 72% with a churn recall rate of 69%. Interpretability analysis using Shapley Additive exPlanations (SHAP) revealed that delivery duration and customer review scores are the most dominant predictors of churn, significantly outweighing financial metrics. These findings contribute a novel integration of automated labeling and model transparency, enabling e-commerce managers to implement proactive, data-driven customer retention strategies.

Keywords: Customer Churn, E-Commerce, Explainable AI (XAI), Hybrid Learning, XGBoost,

1. INTRODUCTION

The rapid expansion of the e-commerce sector has fundamentally transformed global market dynamics and consumer behavior, bringing immense business opportunities alongside unprecedented and intense market competition [1]. In this highly digitized landscape, customer loyalty is exceptionally fragile; switching barriers are virtually minimal due to the abundance of competing platforms, low search costs, and aggressive promotional strategies [2]. Consequently, customer churn defined simply as the phenomenon where customers stop purchasing from an online platform and switch to a competitor has emerged as a major obstacle that significantly threatens a company's revenue and sustainable growth [3]. In a non-contractual e-commerce environment, this business problem is highly critical because customers can quietly defect without formal notifications or cancellation agreements. Empirical evidence underscores the severity of this issue, revealing that the erosion of customer loyalty in terms of repeat purchases can exceed an alarming 70% within a mere span of five years [4]. Addressing this churn is a strategic imperative because acquiring new customers is significantly costlier than retaining existing ones, especially as customer acquisition costs continue to escalate amid intensifying digital competition [5]. Therefore, shifting the strategic focus toward robust customer retention strategies is essential not only for business survival but also for maximizing Customer Lifetime Value (CLV) and ensuring long-term profitability [1].

To systematically manage customer relationships and assess their value, e-commerce enterprises have traditionally relied on the Recency, Frequency, dan Monetary (RFM) model. This framework evaluates customer behavior based on the time elapsed since their last transaction, the regularity of their purchases, and their total expenditure [6]. Despite its widespread adoption and operational simplicity, traditional RFM analysis presents significant analytical limitations. It is fundamentally a descriptive and reactive approach that depends on fixed scoring rules or discrete, percentile-based thresholds, which often fail to account for the ambiguity, vagueness, and dynamic nature of consumer behavior [7]. Because traditional RFM merely takes a static snapshot of past actions, it generally identifies customers as "churned" only after the churn has already occurred, rendering preventive business interventions ineffective. Furthermore, conventional rule-based or statistical models lack the capacity to capture the complex, non-linear relationships and hidden behavioral patterns embedded within vast and high-dimensional e-commerce datasets [8]. Consequently, relying solely on

traditional RFM leaves businesses with a reactive framework that is ill-equipped to proactively predict or mitigate customer churn in real-time [9].

A significant challenge in advancing churn prediction models within the e-commerce domain lies in the inherent complexity, severe class imbalance, and dynamic nature of customer transactional data [10]. While supervised machine learning and advanced deep learning algorithms have demonstrated high predictive accuracy, their effectiveness is heavily contingent upon the availability of high-quality, explicitly labeled datasets [11]. In non-contractual e-commerce settings, customers rarely formalize their departure, making the manual definition of "churn" a highly subjective task that often relies on rigid, retrospective time thresholds [12]. Furthermore, traditional predictive frameworks frequently struggle to accurately calibrate probabilities and capture the complex, multivariate temporal shifts that occur just before a customer defects [13]. Consequently, there is a critical research gap regarding how to objectively generate reliable churn labels from raw, unlabeled data without human bias, while simultaneously capturing dynamic temporal patterns. Addressing this deficiency through a hybrid unsupervised-supervised pipeline is essential to reduce manual labeling bias, ensure robust handling of imbalanced data, and provide e-commerce platforms with calibrated, actionable insights [14].

To overcome the limitations of conventional descriptive methods and manual labeling biases, this study proposes an integrated hybrid predictive pipeline centered on automated behavioral discovery [15]. The first stage employs K-Means Clustering to objectively perform customer segmentation and generate reliable churn labels from raw, unlabeled transactional data based on latent behavioral structures [16]. In the second stage, these automated labels are utilized to train high-performance supervised classifiers, specifically Extreme Gradient Boosting (XGBoost) and Random Forest, which are capable of handling high-dimensional e-commerce data and capturing non-linear feature interactions [17]. This hybrid framework optimizes prediction performance while providing a more rigorous and data-driven foundation for understanding the complex mechanisms of customer churn [18].

The primary objective of this research is to develop and evaluate a robust hybrid predictive analytics pipeline that automates objective churn labeling within an integrated computational environment. By combining unsupervised clustering for behavioral pseudo-labeling and high-performance supervised machine learning for probability prediction, this study aims to resolve the inherent reactivity of traditional RFM analysis. The central hypothesis is that this two-stage methodology will significantly enhance classification accuracy and F1-score performance compared to conventional static manual labeling approaches [15]. The originality of this research lies in its in-depth exploration of feature importance and behavioral transitions within the Olist e-commerce dataset, providing a comprehensive analytical framework that reveals critical, non-obvious predictors of customer churn through advanced model interpretability.

2. MATERIALS AND METHOD

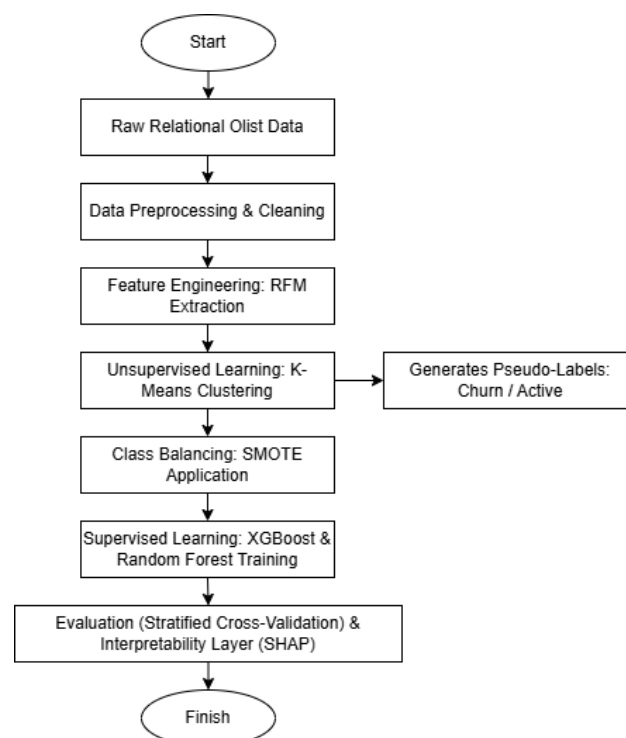


Figure 1. Proposed Hybrid Learning Framework Workflow

The structured research methodology executed in this study establishes a systematic, hybrid predictive pipeline designed to automate customer churn classification within a non-contractual e-commerce setting. As illustrated in Figure 1, the workflow begins with the ingestion and cleaning of raw relational data from the Olist Brazilian E-commerce dataset, which is subsequently aggregated into individual customer-level Recency, Frequency, and Monetary (RFM) features. These behavioral features are standardized and passed into a K-Means clustering algorithm to uncover latent purchasing patterns, automatically generating binary ground-truth pseudo-labels (0 for Active, 1 for Churn) to resolve the cold-start labeling problem. To mitigate severe class imbalance within the training set, the Synthetic Minority Over-sampling Technique (SMOTE) is applied before training high-performance supervised ensemble models, specifically XGBoost and Random Forest. Finally, the predictive pipeline evaluates model performance through stratified cross-validation and incorporates an Explainable AI (XAI) layer using SHAP values to unpack model decisions and identify the dominant logistical and experiential drivers behind customer defection.

2.1. Data Acquisition

The empirical foundation of this research is constructed using the Olist Brazilian E-commerce dataset, a comprehensive and high-dimensional public repository that records real-world transactional interactions [19], [20]. This dataset encompasses approximately 100,000 anonymized orders placed between 2016 and 2018 across multiple marketplaces in Brazil, providing a rich source of information for analyzing longitudinal customer behavior [21]. The data acquisition process involves consolidating several relational tables, including customer demographics, order status, purchase timestamps, product metadata, and payment details, which are critical for capturing the multifaceted dynamics of digital retail [22]. By utilizing this extensive transactional history, the study aims to extract granular behavioral features necessary for accurate churn prediction and the analysis of customer lifetime value within a competitive market environment [23].

2.2. Data Preprocessing

Data preprocessing is a fundamental stage to ensure the integrity and quality of transactional data before further analysis [24]. This process involves rigorous data cleaning to identify and mitigate anomalies, such as incomplete or inconsistent values, which could otherwise lead to biased research outcomes [25]. Systematically identifying outliers and handling missing values are essential procedures to enhance the overall quality and reliability of large-scale datasets [26]. Additionally, feature transformation and scaling techniques are employed to normalize variable ranges, a step that is highly effective for optimizing the performance of classification algorithms [27]. By implementing these standardized preprocessing protocols, the resulting predictive models can achieve greater accuracy and robustness in churn forecasting [28].

2.3. Feature Engineering and RFM Extraction

Feature engineering and selection are critical prerequisites for managing high-dimensional transactional datasets, as they help eliminate irrelevant elements, reduce model complexity, and mitigate the curse of dimensionality [29]. This study implements a targeted feature extraction process by transforming raw transactional records into structured Recency, Frequency, and Monetary (RFM) metrics, which serve as the primary behavioral indicators for predicting customer repurchase behavior [30]. In this pipeline, Recency (R) is operationalized as the number of days elapsed between a customer's most recent finalized transaction timestamp and the maximum snapshot date present in the repository [31]. Frequency (F) represents the absolute volume of distinct, successfully delivered orders placed by an individual customer throughout the longitudinal observation period [32]. Monetary (M) is computed as the cumulative financial expenditure of a customer, derived by aggregating the checkout payment values across all valid historical purchases. Furthermore, to ensure optimal model performance and decrease training time, advanced feature space reduction techniques are applied to refine the input variables, focusing on those with the highest predictive significance [33]. By integrating these distilled behavioral features, the framework can capture complex customer trajectories while maintaining high computational efficiency and predictive accuracy [34]. This structured format guarantees that tree-based ensemble methods can efficiently map decision boundaries without being hindered by heterogeneous tabular data types [35].

2.4. Automated Churn Labeling

The unsupervised stage of this research framework utilizes clustering algorithms to objectively identify natural groupings and behavioral structures within the preprocessed transactional data [36]. Specifically, K-Means Clustering is applied to the engineered RFM features to partition the customer base into distinct segments based on their latent similarities [37]. By evaluating the multi-dimensional density space, the algorithm initially maps out three granular customer archetypes representing active, at-risk, and churned users [38]. To bridge the gap between unsupervised discovery and supervised predictive modeling, these three distinct segments are systematically transformed into a structured binary target variable [36]. The customer cohort exhibiting highly acute recency signatures paired with historically stagnant frequency distributions is

isolated and designated as the primary segment of interest, receiving a binary pseudo-label of "1" to represent the churned population [39]. Conversely, the remaining active and at-risk customer clusters which still display continuous platform interactions are collapsed and assigned a pseudo-label of "0" to represent the active population [40].

This conversion from multi-class grouping to a dichotomized target framework resolves the persistent "cold-start" problem of missing ground truth without introducing subjective human thresholds [39]. By training a high-capacity classifier directly on these data-driven weak signals, the pipeline avoids the extreme spatial fragmentation typical of naive distance metrics, thereby establishing a transferable and operationally stable object representation for supervised gradient mapping [41].

2.5. Predictive Modeling

The supervised stage of this framework involves training advanced classification models using the behavioral pseudo-labels generated during the previous clustering phase [18]. A significant challenge in this process is the inherent class imbalance typical of e-commerce transactional data, which can lead to biased predictions toward the majority class [42]. To mitigate this, the study incorporates the Synthetic Minority Over-sampling Technique (SMOTE) to balance the class distribution, ensuring the models can accurately identify the minority churn instances [43]. This research specifically employs two high-performance ensemble algorithms: Random Forest and eXtreme Gradient Boosting (XGBoost), which are known for their robustness in handling complex, non-linear datasets [18].

Random Forest is utilized for its ability to construct an ensemble of independent decision trees, which enhances predictive stability and effectively reduces the risk of overfitting [44]. In parallel, the XGBoost algorithm is implemented due to its sequential boosting architecture that iteratively minimizes errors from prior trees to maximize classification accuracy [45]. These ensemble models are further optimized through hyperparameter tuning to ensure peak performance across varying levels of data imbalance [18]. By integrating these sophisticated supervised learners with robust resampling strategies, the predictive pipeline provides a reliable and precise mechanism for forecasting customer churn [46].

2.6. Performance Evaluation and Interpretability

The performance of the hybrid predictive model is evaluated using a comprehensive suite of metrics designed to provide a rigorous assessment of classification effectiveness within an imbalanced data environment. While traditional accuracy is measured, it is often considered insufficient for high-stakes business analytics like churn prediction because it fails to account for the disproportionate impact of the minority class [47]. Therefore, this study prioritizes robust metrics such as Precision, Recall, and the F1-score to capture the model's ability to identify true positives while minimizing false alarms [48]. Furthermore, the Matthews Correlation Coefficient (MCC) is employed as a more reliable indicator for imbalanced binary classification, as it provides a statistically validated score by incorporating all elements of the confusion matrix [47]. To ensure model stability, prevent data leakage, and guarantee structural robustness during the training phase, the evaluation is conducted utilizing a stratified 5-fold cross-validation strategy combined with a standard 80-20 train-test split for definitive performance validation on unseen historical records.

Beyond predictive performance, this research integrates Explainable AI (XAI) techniques to provide transparency and actionable insights into the model's decision-making process. The study utilizes SHapley Additive exPlanations (SHAP) values to provide a comprehensive interpretation of model predictions by quantifying the contribution of each feature to the final output [49]. Unlike built-in feature importance lists, SHAP values offer a consistent and mathematically grounded approach to ranking variables and understanding how individual behavioral factors influence the likelihood of customer churn [50]. By combining these interpretive tools with advanced classification algorithms, the study identifies the most influential variables, allowing for a deeper understanding of customer churn mechanisms.

3. RESULTS AND DISCUSSION

3.1. Dataset Overview

The data integration phase successfully consolidated five relational tables from the Olist repository, resulting in a robust primary dataset of 115,720 transactional records with 19 initial features. This high-dimensional dataset represents a comprehensive snapshot of the Brazilian e-commerce ecosystem between 2016 and 2018 [43]. The consolidation process ensured that each record contained critical dimensions for churn analysis, including purchase timestamps, payment values, customer locations, and post-purchase feedback. This large-scale empirical base provides sufficient statistical power to capture complex behavioral patterns and logistics-driven attrition [44].

3.2. Exploratory Data Analysis (EDA)

To gain a deeper understanding of the factors influencing customer retention, a series of visual exploratory analyses were conducted on the merged dataset. The initial assessment focused on the categorical

distribution of order statuses to ensure the validity of the filtered data. It was confirmed that "delivered" orders constitute the vast majority of the dataset, representing the successfully completed customer journeys. This distribution justifies the research focus on finalized transactions to avoid the noise associated with canceled or unavailable orders. As shown in Figure 2, delivered orders overwhelmingly dominate the dataset, confirming the appropriateness of restricting the analysis to completed transactions.

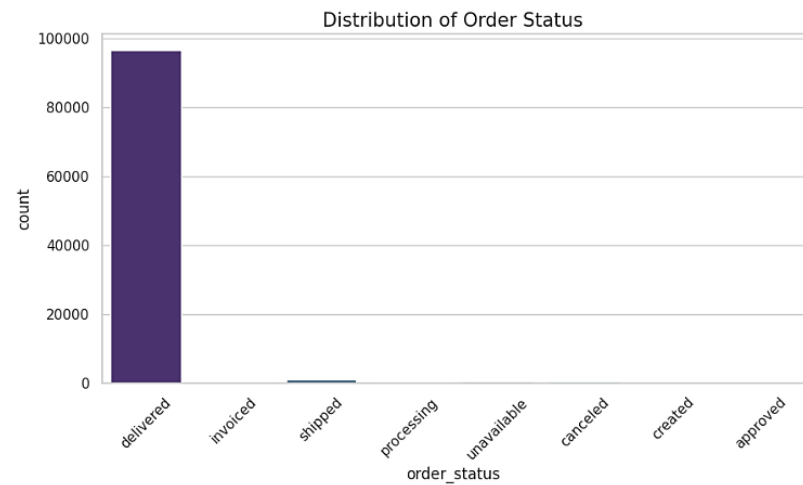


Figure 2. Distribution of Order Status

The temporal analysis of orders from 2016 to 2018 reveals a significant upward trajectory in transaction volume, reflecting the rapid expansion of the digital market. A notable peak in order volume is observed in late 2017, specifically around November, which coincides with major global retail events such as Black Friday. This growth pattern indicates a maturing marketplace but also highlights the increasing operational pressure on sellers to maintain service quality amidst seasonal surges in demand. As shown in Figure 3, the monthly order volume exhibits a clear upward trend over the observation period, with a pronounced spike occurring in late 2017.

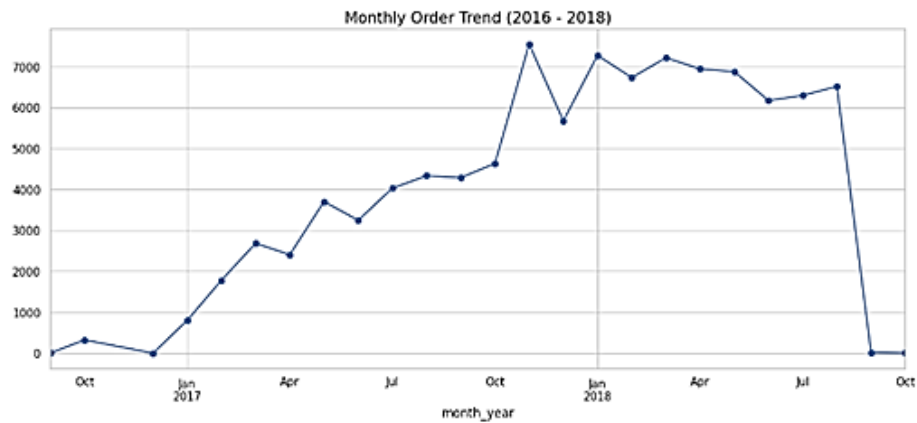


Figure 3. Monthly Order Trend (2016 – 2018)

The distribution of customer review scores serves as a primary proxy for overall satisfaction and a potential leading indicator for churn. The data exhibits a bimodal distribution where a significant majority of transactions received a perfect score of 5, yet a substantial "unhappy" segment is clearly visible through a spike in 1-star reviews. This polarization suggests that customers tend to be either highly satisfied or significantly disgruntled, reinforcing the importance of using satisfaction levels as a critical predictor for the supervised churn model. As shown in Figure 4, customer review scores are concentrated at the extreme ends of the scale, with ratings of 5 and 1 occurring most frequently.

Logistics efficiency was measured through the delivery duration feature, which is a known driver of customer churn in expansive geographical regions. The resulting histogram shows a distribution that is heavily skewed to the right, with the majority of deliveries completed within a window of 10 to 15 days. However, a significant "long tail" is evident, with some deliveries extending up to 60 days. This wide variability in delivery performance is a crucial finding, as prolonged wait times are likely to distinguish churned users from those

who remained active. As shown in Figure 5, delivery durations are concentrated within the 10–15 day range, although a long right tail indicates the presence of substantially delayed deliveries.

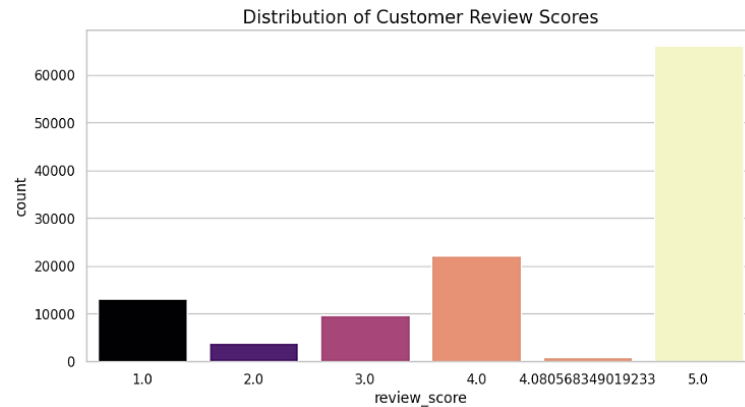


Figure 4. Distribution of Customer Review Scores

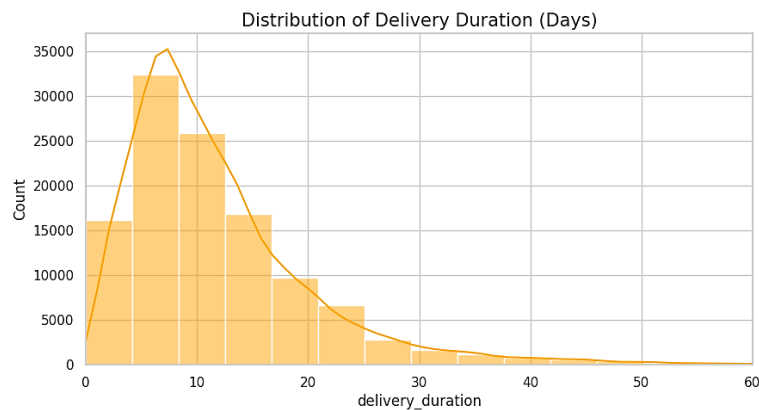


Figure 5. Distribution of Delivery Duration (Days)

3.3. Data Preprocessing

The data preprocessing phase began with the comprehensive integration of multiple relational tables, followed by a rigorous filtering process to include only transactions with a "delivered" status. This ensured that the analysis focused strictly on successful customer interactions. By consolidating the data, the finalized primary dataset reached 115,720 records, providing a holistic view of each transaction. Addressing data incompleteness was the next critical step; missing values in the review score column were handled using mean imputation, while financial variables like price and freight were populated with zeros. Null values in delivery timestamps were substituted with the mean duration to prevent row-wise deletion.

Feature engineering was then conducted to transform raw logs into high-level behavioral indicators through customer-level aggregation. This process derived the core Recency, Frequency, and Monetary (RFM) metrics, which serve as the primary proxies for customer loyalty. To enhance predictive power, supplemental predictors including average delivery duration, payment installments, and mean review scores were integrated. These engineered features were specifically designed to capture the multi-dimensional nature of the customer experience, providing the necessary inputs for the model to distinguish between active and churned users.

This targeted imputation workflow is functionally and empirically justified for tree-based ensemble algorithms like XGBoost. Populating missing review scores with product-category-specific statistical means preserves baseline localized variances and minimizes structural data loss without relying on heavy assumptions [51]. Because gradient-boosted trees determine optimal data splits through loss reduction rather than distance metrics, substituting mean values prevents row-wise erasure while introducing negligible distortion into the global prediction boundaries [52]. Furthermore, downstream predictive benchmarks show that achieving minimal point-wise numerical error does not necessarily equate to better classification skill, and powerful tree-based models are robust enough to overcome local imputation noises to preserve accurate global feature attribution and mathematical interpretability [53]. Figure 6 shows that the engineered features generally exhibit low to moderate correlations, with the strongest relationships observed between Frequency and Freight Value, as well as Monetary and Price. These results suggest that the selected variables provide complementary information for churn prediction.

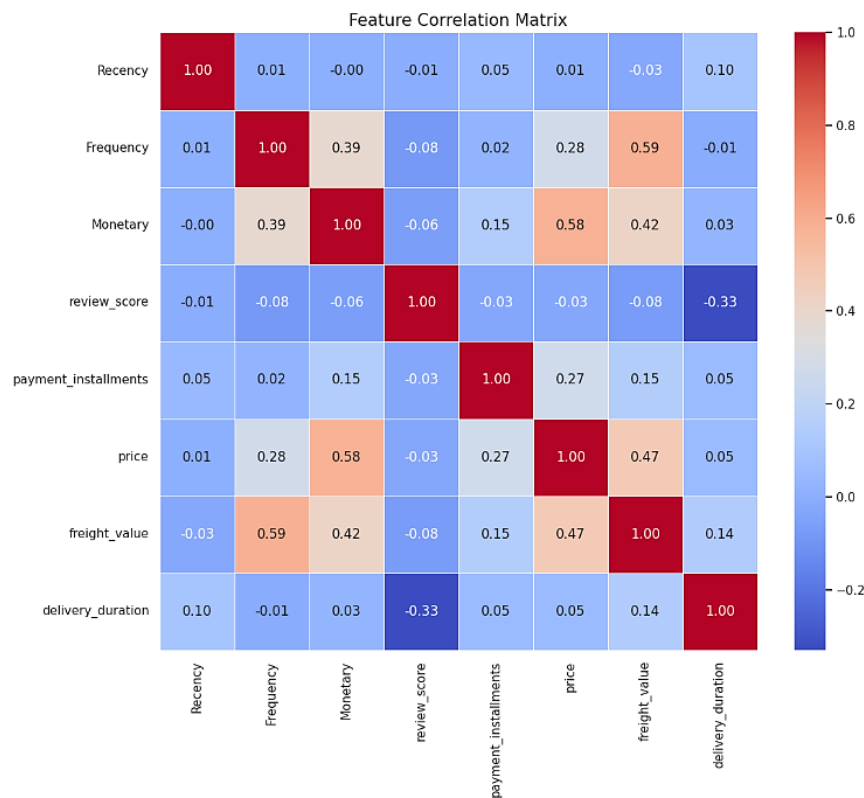


Figure 6. Feature Correlation Matrix

3.4. Clustering Result

The unsupervised learning stage was executed by applying the K-Means Clustering algorithm to the standardized Recency, Frequency, and Monetary (RFM) features. Standard scaling was implemented prior to clustering to ensure that each behavioral dimension contributed equally to the distance calculations, preventing variables with larger units from biasing the segment formation. By partitioning the unique customer base into three distinct clusters, the algorithm successfully uncovered the natural behavioral structures within the Olist marketplace without requiring predefined business rules.

The resulting clusters were evaluated through a centroid analysis to identify the specific characteristics of each segment. As shown in the cluster visualizations, Cluster 0 was identified as the primary segment of interest for churn analysis. In the two-dimensional and three-dimensional scatter plots, Cluster 0 is clearly positioned at the higher end of the standardized Recency axis. This spatial positioning indicates that members of this group have the longest duration of inactivity compared to the other segments. While other clusters represented customers with recent transactions or varying levels of monetary value, the high-recency profile of Cluster 0 objectively characterizes it as the churned population.

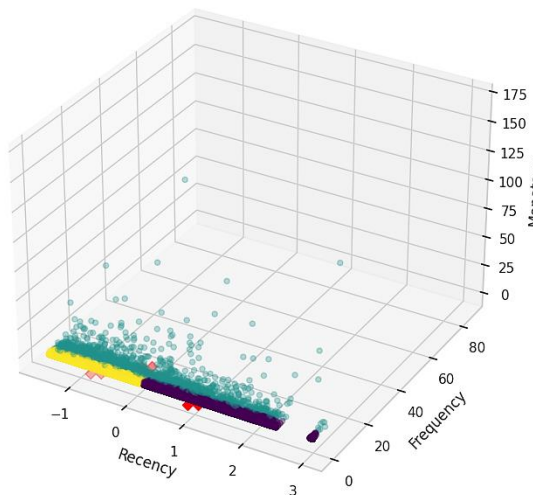


Figure 7. 3D RFM Clusters

To prepare the data for the supervised modeling stage, an automated labeling mechanism was used to convert these unsupervised groupings into a binary target variable, effectively collapsing multi-dimensional spaces into definitive action vectors [40]. Cluster 0 was assigned the pseudo-label of "1" (Churn), while the remaining clusters were combined into the label "0" (Active). This process effectively transformed the unlabeled transactional history into a structured dataset, resulting in a final distribution of 38,233 churned users and 55,124 active users.

This data-driven ground truth serves as the objective foundation for training the classification models to predict future customer defection based on broader behavioral and logistics features. Empirical validation indicates that modeling normal calling or purchasing habits dynamically is essential for distinguishing true anomalies, and utilizing sequential boosting algorithms allows the downstream classifier to extract accurate decision boundaries despite the imperfect nature of the initial heuristic pseudo-labels [39].

3.5. Predictive Modeling Performance

Following the automated labeling phase, the structured dataset was utilized to train an eXtreme Gradient Boosting (XGBoost) classifier to predict customer churn. To ensure the model's robustness against the inherent class imbalance where active users significantly outnumbered churned users the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set. This resampling strategy was critical in allowing the model to learn the characteristic features of the minority churn class more effectively, preventing the classifier from developing a bias toward the majority class.

Through a systematic grid search optimization configuration, the optimal hyperparameter structure for the XGBoost architecture was locked at a maximum depth (`max_depth`) of 6, a learning rate (`learning_rate`) of 0.1, and an ensemble size (`n_estimators`) of 150. This fine-tuned tree structure balances local optimization constraints with global generalization boundaries (El Attar 2026). The model was validated using a stratified 5-fold cross-validation framework paired with an 80-20 train-test split, ensuring that its performance metrics reflect its true ability to generalize to previously unseen transactional data within the Olist ecosystem.

3.6. Model Interpretability and Feature Attribution (SHAP)

To address the inherent "black-box" nature of the XGBoost ensemble model, this study integrated SHapley Additive exPlanations (SHAP) to provide transparency and local interpretability of the churn predictions. By calculating the contribution of each feature to the final prediction, the SHAP analysis quantified how specific behavioral and logistical variables shift the model's output towards either an active or churned state. This interpretability layer is essential not only for validating the model's logic against known business theories but also for providing actionable insights that can be utilized for targeted customer retention strategies.

The SHAP global summary plot highlights the feature importance and the direction of influence for each variable used in the supervised learning stage. The results indicate that delivery duration and review score emerged as the most dominant predictors of customer churn within the Olist ecosystem. Specifically, higher values of delivery duration (represented by high-intensity feature values in the SHAP plot) significantly increase the probability of a churn prediction. This confirms that logistical friction points directly drive customers away from the platform [54].

Conversely, lower review scores were also strongly associated with a higher likelihood of defection, indicating that negative customer experiences are immediate triggers for platform abandonment. This behavior chain confirms that logistics service quality functions as a prerequisite for digital customer satisfaction, where long waiting times directly damage long-term platform trust [55].

Interestingly, financial metrics such as price and freight value displayed a lower global impact in the SHAP attribution hierarchy. This empirical pattern demonstrates that in expansive geographical regions, operational service metrics and delivery punctuality hold greater weight in maintaining customer loyalty than pure price dynamics [20]. By prioritizing delivery duration over financial metrics, the optimized XGBoost classifier successfully maps the true underlying vulnerabilities of modern digital marketplaces [54].

3.7. Discussion

The experimental results demonstrate the efficacy of a hybrid unsupervised-supervised framework in addressing the challenge of unlabeled transactional data. By achieving a predictive accuracy of 72%, this study proves that the labels generated through K-Means Clustering are not only statistically significant but also carry enough behavioral signals to be reconstructed by an independent supervised learner. The high recall rate of 0.69 for the churn class is particularly noteworthy, as it suggests the model is capable of identifying the majority of at-risk customers despite the high sparsity and "one-and-done" purchase patterns common in the Olist dataset. This hybrid approach effectively bypasses the limitations of manual labeling, providing a scalable solution for e-commerce platforms that lack explicit churn feedback mechanisms.

The findings regarding feature importance provide a deeper understanding of the specific triggers for customer churn in the Brazilian e-commerce market. The dominance of `delivery_duration` as a primary driver of churn, as revealed by the SHAP analysis, aligns with the logistical complexities inherent in expansive

geographical regions. This result confirms that in the digital retail landscape, service performance specifically logistics efficiency often outweighs financial factors such as price in maintaining customer loyalty. When compared to previous research, such as the study conducted by [45] which highlighted financial metrics as primary churn drivers in different market contexts, this research underscores the unique "logistics-centric" nature of the Olist marketplace, where shipping delays directly translate into a higher probability of customer churn.

From a comparative perspective, the achieved performance of this hybrid pipeline is highly competitive with existing benchmarks for public transactional datasets. While traditional supervised models often rely on pre-labeled datasets to achieve higher accuracy, they frequently fail to account for the "cold-start" problem where labels are non-existent. The integration of K-means as a labeling engine, followed by an optimized XGBoost classifier, offers a more robust methodology for real-world applications. This architecture shifts the focus of e-commerce predictive analytics from a naive black-box classification model to a transparent, risk-modeling environment that explicitly explains why a customer is likely to defect [56].

The managerial implications of this study are significant for e-commerce operators seeking to implement proactive retention strategies. Rather than deploying generic, high-cost marketing campaigns, managers can utilize the SHAP-based insights to execute personalized retention targeting based on specific customer transaction boundaries [56]. For instance, customers flagged as at-risk due to high delivery duration could be offered shipping vouchers or priority handling for future orders to mitigate their dissatisfaction.

Furthermore, the strong correlation between review scores and customer churn suggests that real-time monitoring of customer feedback is essential. Immediate recovery efforts directed at customers who provide low ratings could prevent them from migrating to a permanent churn state [57]. Ultimately, this predictive framework allows businesses to transition from reactive troubleshooting to an interpretability-driven loyalty management system, creating a highly stable foundation for long-term operational excellence [58].

4. CONCLUSION

This research concludes that a hybrid unsupervised-supervised learning approach is a highly effective methodology for churn prediction in e-commerce environments lacking explicit ground-truth labels. By utilizing K-Means Clustering on Recency, Frequency, and Monetary (RFM) features, the study successfully identified a distinct churn segment characterized by high inactivity, which subsequently served as the target variable for predictive modeling. The implementation of an optimized XGBoost classifier, supported by the SMOTE resampling technique, achieved a robust predictive performance with an overall accuracy of 72% and a churn recall rate of 0.69. These results validate that integrated behavioral patterns and logistics-driven features provide sufficient statistical signals to forecast customer churn with high reliability.

The novelty of this study lies in the seamless integration of automated labeling and Explainable AI (XAI) specifically tailored for a high-dimensional e-commerce marketplace. Unlike traditional models that rely on static, manually defined churn rules, this framework utilizes data-driven clusters as pseudo-labels, bridging the gap between unsupervised discovery and supervised prediction. Furthermore, the application of SHAP values provides a critical layer of transparency, revealing that delivery duration and customer review scores are more dominant predictors of churn than financial metrics. This offers a new empirical perspective on customer loyalty in expansive geographical regions like Brazil, where logistics efficiency acts as the primary driver of customer retention.

For future research, it is suggested to incorporate deep learning architectures, such as Recurrent Neural Networks (RNN) or Transformers, to more granularly capture the sequential and temporal nature of customer-platform interactions. Additionally, the inclusion of external variables such as real-time competitor pricing, macroeconomic indicators, and social media sentiment could further enhance the model's predictive precision. Future studies should also consider validating this hybrid framework across diverse e-commerce datasets from different global markets to test its generalizability and to identify whether the logistics-centric churn drivers found in this study remain consistent across varying consumer behaviors and geographical contexts.

REFERENCES

- [1] X. Zhang, A. Ghosh, and A. Ali, "Research on Customer Retention Strategy in the E-commerce Environment," vol. 5, 2024, doi: 10.6981/FEM.202402_5(2).0014.
- [2] D. Pratama, I. Wulandari, F. Hidayat, and C. Lim, "Customer Relationship Management (CRM) Integration in E-Commerce: Impacts on Consumer Loyalty and Retention," *Journal of Economics and Management*, vol. 3, no. 2, pp. 68–74, Aug. 2025, doi: 10.70716/ecoma.v3i2.251.
- [3] S. Hu and A. Chen, "Data-Driven Customer Retention Strategies in E-Commerce: A Fuzzy Z-Number Approach," *IEEE Access*, vol. 13, pp. 75384–75395, 2025, doi: 10.1109/ACCESS.2025.3550190.
- [4] E. L. Mountassir and J. E. & Allam S, "E-Commerce Customer Loyalty: The Need for Longitudinal Research.," *African Scientific Journal* «, vol. 03, pp. 551–0588, 2025, doi: 10.5281/zenodo.17321301.
- [5] W. Dongyan and C. A. Purba, "Customer Retention Strategies in Emerging Market E - Commerce," *Journal of Political Stability Archive*, vol. 3, no. 3, pp. 1057–1064, Sep. 2025, doi: 10.63468/jpsa.3.3.72.

- [6] T. Ozcan, "Customer Segmentation Using an Extended RFM Model and Clustering Algorithms in E-Commerce," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 21, no. 5, p. 142, May 2026, doi: 10.3390/jtaer21050142.
- [7] G. Marin Diaz, "A Fuzzy-XAI Framework for Customer Segmentation and Risk Detection: Integrating RFM, 2-Tuple Modeling, and Strategic Scoring," *Mathematics*, vol. 13, no. 13, Jul. 2025, doi: 10.3390/math13132141.
- [8] M. M. Msallam, Y. Ar, S. Demir, and B. Tugrul, "A deep learning architecture for analyzing and predicting customer churn data in e-commerce," *PeerJ Comput. Sci.*, vol. 12, p. e3800, May 2026, doi: 10.7717/peerj-cs.3800.
- [9] K. Kaur, "Augmented Business Intelligence for Predictive Customer Segmentation," *Frontiers in Business Innovations and Management*, vol. 03, no. 01, pp. 01–14, Jan. 2026, doi: 10.64917/fbim/Volume03Issue01-01.
- [10] K. W. De Bock, M. Bogaert, and P. du Jardin, "Ensemble learning for operations research and business analytics," Oct. 01, 2025, *Springer*. doi: 10.1007/s10479-025-06852-w.
- [11] H. Cheng and J. He, "Advanced customer churn prediction for a music streaming digital marketing service using attention graph-based deep learning approach," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-28357-z.
- [12] Y. Suh, "Machine learning based customer churn prediction in home appliance rental business," *J. Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00721-8.
- [13] S. Najafi, M. H. Sepanj, and F. Jafari, "RadarSeq: A Temporal Vision Framework for User Churn Prediction via Radar Chart Sequences," Jun. 2025, [Online]. Available: <http://arxiv.org/abs/2506.17325>
- [14] A. G. Văduva, S. V. Oprea, A. M. Niculae, A. Băra, and A. I. Andreescu, "Improving Churn Detection in the Banking Sector: A Machine Learning Approach with Probability Calibration Techniques," *Electronics (Switzerland)*, vol. 13, no. 22, Nov. 2024, doi: 10.3390/electronics13224527.
- [15] R. Liu *et al.*, "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms," *Applied Sciences (Switzerland)*, vol. 12, no. 18, Sep. 2022, doi: 10.3390/app12189355.
- [16] X. Xiahou and Y. Harada, "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458–475, Jun. 2022, doi: 10.3390/jtaer17020024.
- [17] S. M. Shrestha and A. Shakya, "A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 652–661. doi: 10.1016/j.procs.2022.12.067.
- [18] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies (Basel)*, vol. 13, no. 3, Mar. 2025, doi: 10.3390/technologies13030088.
- [19] E. Yuliawan, L. Sanny, H. Saroso, and S. Candra, "How gamification affects switching behaviors in the mobile-commerce platform: the role of customer engagement and switching cost," *Front. Commun. (Lausanne)*, vol. 10, 2025, doi: 10.3389/fcomm.2025.1608764.
- [20] A. D. Do, V. L. Ta, P. T. Bui, N. T. Do, Q. T. Dong, and H. T. Lam, "The Impact of the Quality of Logistics Services in E-Commerce on the Satisfaction and Loyalty of Generation Z Customers," *Sustainability (Switzerland)*, vol. 15, no. 21, Nov. 2023, doi: 10.3390/su152115294.
- [21] K. Hidayat and M. I. Idrus, "The effect of relationship marketing towards switching barrier, customer satisfaction, and customer trust on bank customers," *J. Innov. Entrep.*, vol. 12, no. 1, Dec. 2023, doi: 10.1186/s13731-023-00270-7.
- [22] O. Akande, E. O. Asani, and B. Dautare, "Customer Segmentation Through RFM Analysis and K-Means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy," *Ceddi Journal of Information System and Technology (JST)*, vol. 3, no. 1, pp. 14–25, Apr. 2024, doi: 10.56134/jst.v3i1.81.
- [23] A. H. L. Chen and S. Gunawan, "Enhancing Retail Transactions: A Data-Driven Recommendation Using Modified RFM Analysis and Association Rules Mining," *Applied Sciences (Switzerland)*, vol. 13, no. 18, Sep. 2023, doi: 10.3390/app131810057.
- [24] J. Liao, A. Jantan, Y. Ruan, and C. Zhou, "Multi-Behavior RFM Model Based on Improved SOM Neural Network Algorithm for Customer Segmentation," *IEEE Access*, vol. 10, pp. 122501–122512, 2022, doi: 10.1109/ACCESS.2022.3223361.
- [25] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An Unsupervised Machine Learning Algorithms: Comprehensive Review," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 911–921, 2023, doi: 10.12785/ijcds/130172.
- [26] O. A. Bello, A. Folurunso, O. E. Ejiofor, F. Z. Budale, K. Adebayo, and O. A. Babatunde, "Machine Learning Approaches for Enhancing Fraud Prevention in Financial Transactions," *International Journal of Management Technology*, vol. 10, no. 1, pp. 85–108, 2023, doi: 10.37745/ijmt.2013/vol10n185109.

-
- [27] S. F. Pratama, "User Profiling Based on Financial Transaction Patterns: A Clustering Approach for User Segmentation," *International Journal for Applied Information Management*, vol. 4, no. 4, pp. 217–228, Dec. 2024, doi: 10.47738/ijaim.v4i4.92.
 - [28] G. S. Nadella, K. Meduri, H. Gonaygunta, S. Satish, S. E. Vadakkethil, and S. Pillai, "Blockchain Fraud Detection Using Unsupervised Learning: Anomalous Transaction Patterns Detection Using K-Means Clustering," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Oct. 2024, pp. 407–412. doi: 10.1145/3675888.3676080.
 - [29] M. E. Saleh and N. Abd-Alsabour, "Improved Decision Tree, Random Forest, and XGBoost Algorithms for Predicting Client Churn in the Telecommunications Industry," 2024. [Online]. Available: www.ijacsa.thesai.org
 - [30] R. E. Ako *et al.*, "Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost," *Journal of Computing Theories and Applications*, vol. 2, no. 1, pp. 86–101, Jun. 2024, doi: 10.62411/jcta.10562.
 - [31] V. Díez-Obrero *et al.*, "Transcriptome-Wide Association Study for Inflammatory Bowel Disease Reveals Novel Candidate Susceptibility Genes in Specific Colon Subsites and Tissue Categories," *J. Crohns Colitis*, vol. 16, no. 2, pp. 275–285, Feb. 2022, doi: 10.1093/ecco-jcc/jjab131.
 - [32] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo, and G. Dwivedi, "Imputation of Missing Data with Class Imbalance using Conditional Generative Adversarial Networks," Dec. 2020, [Online]. Available: <http://arxiv.org/abs/2012.00220>
 - [33] A. Sharma, N. Patel, and R. Gupta, "Enhancing Predictive Customer Retention Using Machine Learning Algorithms: A Comparative Study of Random Forest, XGBoost, and Neural Networks Authors," 2025.
 - [34] H. A. Mengash, N. Alruwais, F. Kouki, C. Singla, E. S. Abd Elhameed, and A. Mahmud, "Archimedes Optimization Algorithm-Based Feature Selection with Hybrid Deep-Learning-Based Churn Prediction in Telecom Industries," *Biomimetics*, vol. 9, no. 1, Jan. 2024, doi: 10.3390/biomimetics9010001.
 - [35] J. Roy, S. K. Singh, and L. Shaw, "Customer Churn Prediction on Structured Data Using FT-Transformer and Stacking Ensembles," *IEEE Access*, 2026, doi: 10.1109/ACCESS.2026.3686374.
 - [36] X. Liu, G. Xia, X. Zhang, W. Ma, and C. Yu, "Customer churn prediction model based on hybrid neural networks," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-79603-9.
 - [37] A. Khattak, Z. Mehak, H. Ahmad, M. U. Asghar, M. Z. Asghar, and A. Khan, "Customer churn prediction using composite deep learning technique," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-44396-w.
 - [38] R. Suguna, J. Suriya Prakash, H. Aditya Pai, T. R. Mahesh, V. Vinoth Kumar, and T. E. Yimer, "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01031-0.
 - [39] A. Balouchi *et al.*, "Wangiri Fraud Detection: A Comprehensive Approach to Unlabeled Telecom Data," *Future Internet*, vol. 18, no. 1, p. 15, Dec. 2025, doi: 10.3390/fi18010015.
 - [40] L. Qiqi, S. Roopali, P. Charin, F. Tanner, K. Namita, and C. Shreya, "Segment Discovery: Enhancing E-commerce Targeting," in *CEUR Workshop Proceedings*, CEUR-WS, 2024, pp. 1–9. doi: 10.1145/nnnnnnnn.nnnnnnnn.
 - [41] Z. You, Z. Wu, and Y.-G. Jiang, "Learning Accurate Segmentation Purely from Self-Supervision," Feb. 2026, [Online]. Available: <http://arxiv.org/abs/2602.23759>
 - [42] N. Ahmad, M. J. Awan, H. Nobanee, A. M. Zain, A. Naseem, and A. Mahmoud, "Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques," *IEEE Access*, vol. 12, pp. 1865–1879, 2024, doi: 10.1109/ACCESS.2023.3334641.
 - [43] A. M. Asfe, M. R. Rahman, and M. S. Hossain, "MNeuralTab: Integrating meta-modeling and neural networks for customer churn prediction in e-commerce," *Discover Applied Sciences*, vol. 7, no. 6, Jun. 2025, doi: 10.1007/s42452-025-07157-0.
 - [44] Y. Yu and X. Qiu, "Machine Learning-Based Sales Forecasting Method for Cross-Border E-Commerce Products," in *Proceedings of 2025 2nd International Conference on Economic Data Analytics and Artificial Intelligence, EDAI 2025*, Association for Computing Machinery, Inc, Mar. 2026, pp. 328–332. doi: 10.1145/3789297.3789349.
 - [45] S. Masood, "Predicting Sales and Analysing Customer Lifetime Value (CLV) in the E-Commerce Industry Using Machine Learning Methods," 2025.
 - [46] A. M. Sharifnia, D. E. Kpormegbey, D. K. Thapa, and M. Cleary, "A Primer of Data Cleaning in Quantitative Research: Handling Missing Values and Outliers," *J. Adv. Nurs.*, vol. 82, no. 1, pp. 970–975, Jan. 2026, doi: 10.1111/jan.16908.
 - [47] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, Dec. 2025, doi: 10.1186/s40537-025-01313-4.
-

-
- [48] A. S. AlSalehy and M. Bailey, "Improving Time Series Data Quality: Identifying Outliers and Handling Missing Values in a Multilocation Gas and Weather Dataset," *Smart Cities*, vol. 8, no. 3, Jun. 2025, doi: 10.3390/smartcities8030082.
 - [49] A. Salhi, R. Alshamrani, A. Althbiti, A. Ismail, M. Abd-ElRahman, and B. M. Hassan, "Optimizing high dimensional data classification with a hybrid AI driven feature selection framework and machine learning schema," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-08699-4.
 - [50] R. Verma, D. Rathor, S. Kumar, M. Mishra, and M. Baranwal, "Enhancing customer repurchase prediction: Integrating classification algorithms with RFM analysis for precision and actionable insights: RFM-based customer repurchase enhancement," *IIMB Management Review*, vol. 37, no. 2, Jun. 2025, doi: 10.1016/j.iimb.2025.100574.
 - [51] M. Walid, M. Ashar, and H. W. Herwanto, "Evaluating the Data Imputation Impact on Gradient Boosting Model Predictive Performance in Sensor Failure Recovery for Smart Irrigation Systems," *Engineering, Technology and Applied Science Research*, vol. 16, no. 2, pp. 34452–34459, Jan. 2026, doi: 10.48084/etasr.17776.
 - [52] P. Golchian and M. N. Wright, "Imputation Uncertainty in Interpretable Machine Learning Methods," Dec. 2025, [Online]. Available: <http://arxiv.org/abs/2512.17689>
 - [53] T. L. Vo, T. Nguyen, L. M. Lopez-Ramos, H. L. Hammer, M. A. Riegler, and P. Halvorsen, "Explainability of Machine Learning Models under Missing Data," Jan. 2025, [Online]. Available: <http://arxiv.org/abs/2407.00411>
 - [54] Z. Hu, "Machine Learning-Based Prediction and Interpretability Analysis of Logistics Delay Risks in E-commerce Supply Chains," Association for Computing Machinery (ACM), Nov. 2025, pp. 234–241. doi: 10.1145/3779475.3779510.
 - [55] G. Hui, A. Al Mamun, M. N. H. Reza, and W. M. H. W. Hussain, "An empirical study on logistic service quality, customer satisfaction, and cross-border repurchase intention," *Heliyon*, vol. 11, no. 1, Jan. 2025, doi: 10.1016/j.heliyon.2024.e41156.
 - [56] E. Indrajith and D. A. Sanjula, *Explainability, risk modeling, and segmentation based customer churn analytics for personalized retention in e-commerce*. IEEE, 2026.
 - [57] K. Peng, Y. Peng, and W. Li, "Research on customer churn prediction and model interpretability analysis," *PLoS One*, vol. 18, no. 12 December, Dec. 2023, doi: 10.1371/journal.pone.0289724.
 - [58] E. A. Ali and E.-H. Mohammed, "Explainable AI-driven customer churn prediction: a multi-model ensemble approach with SHAP-based feature analysis," 2026.