Dean&Francis
ISSN 2959-6130

E-Commerce Customer Churn Prediction and Key Determinant Investigation Based on Machine Learning Algorithms

Mingyue Zheng

Department of Mathematics and Statistics, Shanxi University, Taiyuan, China
zhengmingyue@sxu.edu.cn

Abstract:

In response to the challenges of high dimensionality, class imbalance, and a large number of missing values in e-commerce customer data, this paper aims to construct a high-precision and interpretable customer churn prediction model to enhance the customer retention capabilities of e-commerce enterprises. This study is based on the XGBoost algorithm, using median imputation to handle missing values, applying oversampling techniques to alleviate class imbalance, and combining multi-dimensional feature selection to enhance model interpretability. By comparing the performance of three models - logistic regression, decision tree and XGBoost - it was found that XGBoost significantly outperformed the other models even without tuning. Further, after hyperparameter optimization, the model achieved an accuracy of 98.49% and an Area Under Curve (AUC) value of 0.99 on the test set, demonstrating excellent generalization ability. Feature importance analysis indicated that "customer tenure" and "whether to complain" were the core factors influencing churn. This study provides a robust and interpretable solution for e-commerce customer churn warning, with strong practical application value.

Keywords: Customer churn prediction; XGBoost; E-commerce; Class imbalance; Feature importance

1. Introduction

E-commerce refers to an electronic transaction model facilitated through the internet. As a modern business paradigm, it has reached a stage of relative maturity, with the growth of new customers gradually decelerating while customer acquisition costs and difficulties continue to escalate. Customer churn is often defined as the termination of service usage or commercial relationships by clients, has emerged as a critical factor impacting the sustainable development of enterprises. Consequently, proactively predicting churn risk and implementing targeted intervention measures hold significant importance for enhancing customer retention rates [1]. In recent years, the widespread application of big data and machine learning technologies, particularly advancements in ensemble learning and

deep learning methodologies, has substantially improved the accuracy and reliability of user behavior prediction [2]. However, e-commerce customer data typically exhibits characteristics such as high dimensionality, class imbalance, and numerous missing values, presenting considerable challenges in constructing robust and interpretable predictive models.

In previous studies on customer churn prediction, traditional machine learning models such as Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF) have been widely adopted [3]. More recently, XGBoost has gained attention due to its robustness to missing data and high predictive accuracy [4]. Previous studies on customer churn have predominantly focused on traditional sectors such as financial services and telecommunications. For instance, the Light Gradient Boosting Machine (LGBM) model has been employed to predict bank customer churn [5]. Additionally, a comprehensive analysis of the risk factors contributing to telecommunications customer churn has been conducted using a combination of random survival forests, Cox regression, and Kaplan-Meier survival curves [6]. Given the relatively early start of research in these fields, a relatively well-developed analytical framework and methodology system have been established. However, existing research still has obvious limitations in the e-commerce context: most methods do not systematically address the issues of class imbalance and missing values, resulting in insufficient generalization ability of the models [1]; feature selection usually relies on statistical tests or internal indicators of tree models, lacking multi - dimensional evaluation and business interpretability [7]; moreover, the overfitting risk of XGBoost in highly imbalanced churn prediction has not been fully explored. In response to these challenges, this study aims to develop a more robust framework by using mean imputation to handle missing values, introduces oversampling techniques to alleviate class bias, and enhances the interpretability and generalization performance of the model through multi- dimensional feature selection.

To tackle these issues, this study develops an e-commerce customer churn prediction model based on the XGBoost algorithm, systematically addressing issues of data imbalance and missing value processing. Through comparative analysis of the predictive performance among logistic regression, decision tree, and XGBoost models, the results demonstrate that XGBoost exhibits significant advantages, achieving a test accuracy of 98.49% after parameter optimization. Furthermore, feature importance analysis reveals that tenure and complain are the two most predictive core features. The study establishes a high-precision and interpretable customer churn early-warning framework, providing robust theoretical and practical support for customer retention strategies in the e-commerce sector, thereby effectively addressing the limitations of existing research in terms of practicality and stability.

2. Method

2.1 Dataset Preparation

The source of the dataset used in this study is from Kaggle [8]. The original data set consisted of 5,630 data and 20 features such as Tenure, PreferredLoginDevice, CityTier. The aim of this project is to develop a predictive model using Machine Learning (ML) algorithms for analysing customer attrition in the E-commerce industry.

The target variable in this dataset is customer churn, where a value of 1 indicates a churned customer and 0 indicates a non-churned customer. The features include 10 numerical variables and 6 categorical variables.

The data preprocessing procedure comprises five sequential steps. Initially, feature variables not involved in modeling—specifically customer ID—were eliminated. Subsequently, missing values were detected, identifying 1,856 instances with missing data, accounting for 32.97% of the total sample size. Given the proximity of means and medians across numerical variables and the requirement to preserve integer characteristics for certain features, this study employed median imputation to uniformly address missing values. The third step involved visualizing the distribution of customer churn. As illustrated in Fig. 1 and Fig. 2, the majority of customers retained their subscriptions, with average app usage durations ranging from 2 to 4 hours and order frequencies predominantly between 1 and 2 instances; conversely, customers who filed complaints or were identified as single exhibited notably higher churn rates. Fourth, due to the limited number of churned customers (948 instances, representing 16.8% of the total sample), a significant class imbalance was identified. To mitigate this, an oversampling technique was applied to augment the minority class, resulting in a balanced dataset with 4,682 instances each for both churned and non-churned categories. Finally, standardization was performed to transform all feature values into a distribution with a mean of 0 and standard deviation of 1, thereby enhancing model training efficacy.

ISSN 2959-6130

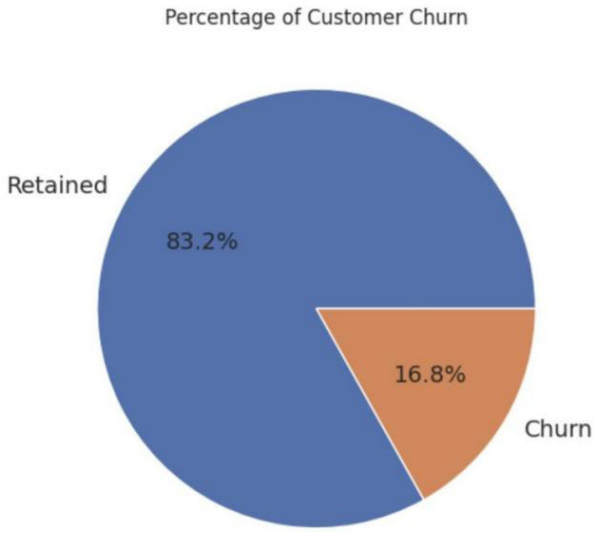

Fig. 1 Exited proportion (Photo/Picture credit : Original).

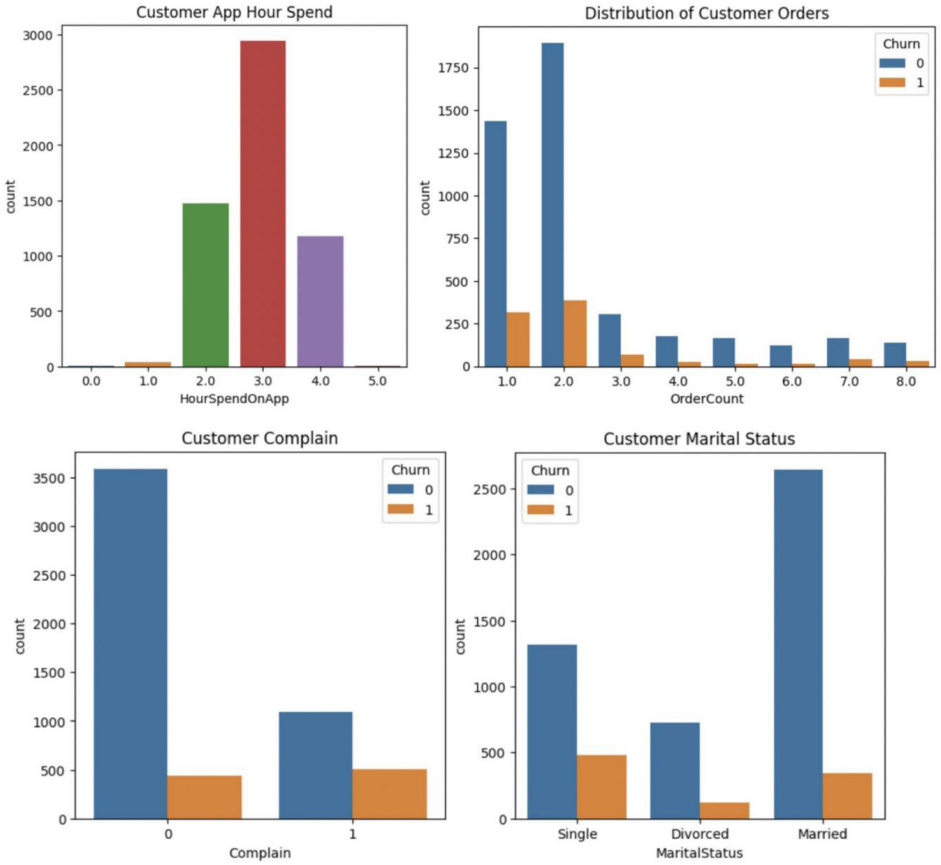

Fig. 2 Factors Influencing Customer Churn: Marital Status, Orders, Complaints and App Engagement (Photo/Picture credit : Original).

2.2 Machine Learning Models

This study used logistic regression, decision tree and XGBoost models implemented with sklearn to predict e-commerce customer churn and identify important features. Model evaluation was based on training accuracy, test accuracy, the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC) score. More details are as followed:

2.2.1 Logistic Regression

Logistic regression is a classical linear model designed for binary classification tasks. It employs a specialized function known as the sigmoid function to transform the linear combination of input features into a probability value between 0 and 1, thereby predicting the likelihood of a particular event occurring [9].

The operational pipeline of logistic regression commences with the linear combination of features and model coefficients:

$$Z = \beta_0 + \beta_1 X_1 + ... + \beta_n X_{n.} \quad (1)$$

Subsequently, this linear combination z is processed through the sigmoid function, yielding a probability value p(x) bounded between 0 and 1.

$$P = \frac{1}{1 + e^{-Z}} . \quad (2)$$

This probability quantifies the likelihood of a sample belonging to the positive class. To convert this continuous probability into a discrete class label (0 or 1), a threshold is typically established. Samples with predicted probabilities equal to or exceeding this threshold are classified as positive (1), while others are assigned to the negative class (0). The model parameters (β) are generally estimated by maximizing the likelihood function, often implemented through optimization algorithms such as gradient descent.

The foremost advantages of this model include its simplicity, computational efficiency, and the high interpretability of its coefficients. However, its primary limitation lies in its inherent inability to autonomously capture complex nonlinear relationships.

2.2.2 Decision Tree

Decision trees simulate the decision-making process through a series of sequential "if-then-else" rules, ultimately forming a tree-structured predictive model. Starting from the root node, the algorithm recursively partitions the data based on feature values, aiming to maximize the homogeneity of samples within each resulting subset [10].

At each node, the algorithm evaluates all features and their potential split points according to predefined criteria, selecting the feature and threshold that best separate the data into subsets with maximal class purity. This process iterates recursively on newly generated child nodes until predetermined stopping conditions are met. Eventually, each leaf node is assigned a class label.

The advantages of decision trees include intuitive interpretability, minimal data preprocessing requirements, and the ability to handle mixed data types. However, individual trees are highly prone to overfitting, which is why they are often employed as base models in ensemble learning methods.

2.2.3 XGBoost

XGBoost is an advanced gradient boosting framework that builds a powerful ensemble model by sequentially constructing multiple weak decision trees. Each new tree is dedicated to correcting the prediction errors left by the previous tree and controls model complexity by adding regularization terms, thereby achieving outstanding predictive performance [1].

The XGBoost algorithm starts with a simple initial model for prediction. Then, new trees are iteratively added. In each round of iteration, the construction goal of the new tree is no longer to fit the original labels, but to fit the residuals between the current model's predicted values and the true values. Additionally, XGBoost introduces regularization terms and second-order Taylor expansions in the objective function to more accurately approximate the loss function.

XGBoost minimizes the loss function through step-by-step optimization and model regularization, while striving to enhance computational efficiency and accuracy. The built-in parallel processing, missing value handling mechanism, and support for various custom loss functions in XGBoost make it a fast, accurate, and flexible powerful tool.

3. Results and Discussion

3.1 Machine Learning Performance Comparison

The performance evaluation results of the three machine learning models indicate that the XGBoost model demonstrates the most superior performance. As illustrated in Table 1 and Fig. 3, this model achieved a test accuracy of 97.87% and an AUC score of 0.96. Across all evaluation metrics, XGBoost significantly outperformed both the logistic regression and decision tree models.

Table 1. Model Performance

	Test accuracy	AUC Score
Logistic Regression	89.17%	0.87
Decision Tree	96.98%	0.95
XGBoost	97.87%	0.96

ISSN 2959-6130

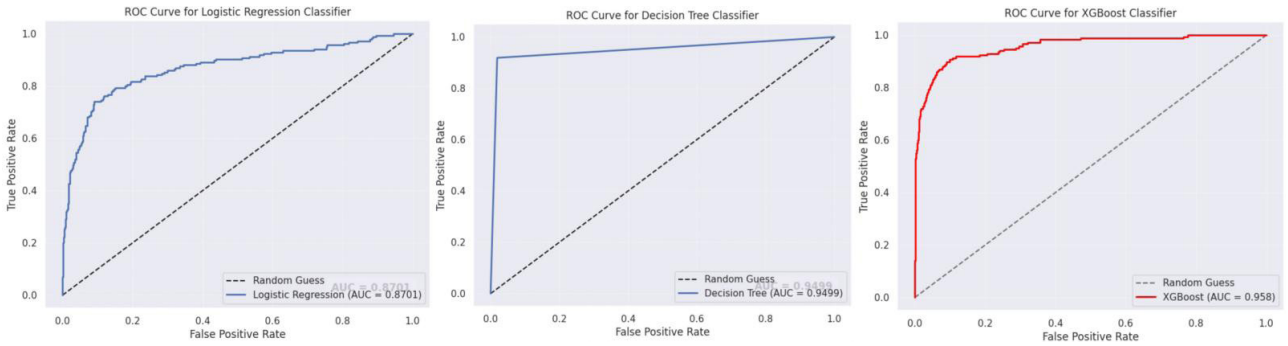

Fig. 3 AUC Score (Photo/Picture credit : Original)

As shown in Table 1 and Fig. 3, the XGBoost model demonstrates significantly superior predictive performance compared to both logistic regression and single decision tree models. The following analysis elucidates the underlying reasons for these performance differences:

Logistic regression, as a generalized linear model, operates under the core assumption of a linear relationship between features and the target variable. However, e-commerce customer churn data often involve complex nonlinear feature interactions, which logistic regression struggles to capture effectively. This limitation is reflected in its lowest test accuracy (89.17%) and AUC value (0.87) among the evaluated models.

Although decision tree models can capture nonlinear relationships through recursive partitioning and achieved relatively high test accuracy (96.98%) and AUC (0.95), their performance remains slightly inferior to that of XGBoost.

XGBoost successfully overcomes the linear constraints of logistic regression and the overfitting issues inherent in decision trees through its sophisticated ensemble learning framework, regularization techniques, and optimization algorithms. Consequently, it exhibits optimal performance in the customer churn prediction task examined in this study.

3.2 The Influence of Hyperparameters on Model Performance

This study employed a systematic parameter scanning approach to conduct optimization experiments on three core parameters: maximum tree depth, learning rate, and the number of weak learners.

As illustrated in Fig. 4, experimental results regarding maximum tree depth (max_depth) indicate that the model achieved optimal test accuracy when max_depth was set to 10. At shallower depths (<8), the model exhibited underfitting and failed to adequately capture complex features within the data; conversely, excessive depth (>15) led to pronounced overfitting. The optimization process for the learning rate (eta) revealed that a setting of 0.4 yielded the best model performance. It was observed that higher learning rates (>0.6) resulted in unstable training processes and premature convergence, whereas lower rates (<0.2) necessitated a substantial increase in the number of trees to achieve comparable performance, thereby significantly raising computational costs. Experiments on the number of weak learners (n_estimators) demonstrated that model performance reached a stable plateau when n_estimators was set to 400.

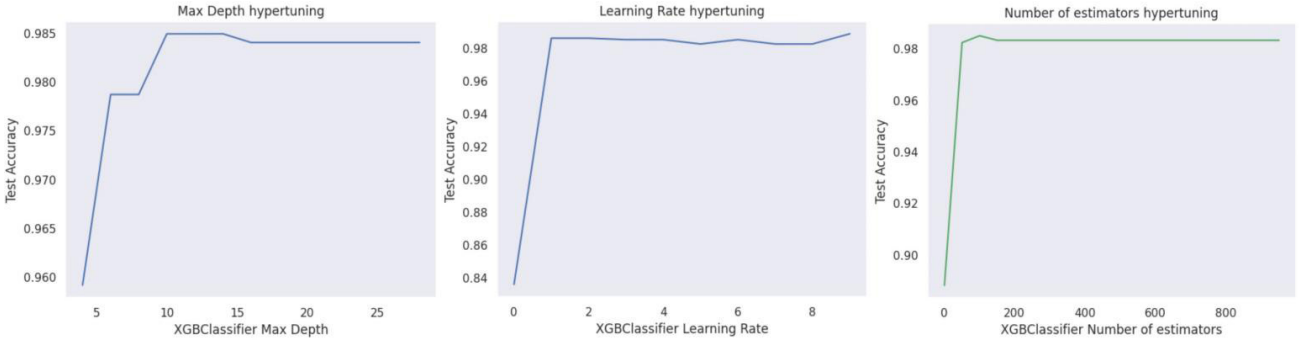

Fig. 4 Performance Evaluation of XGBoost Classifier through Hyperparameter Optimization (Picture credit: Original)

The optimal hyperparameter combination derived from these experimental results (max_depth=10, eta=0.4, n_estimators=400) substantially enhanced model performance. As shown in Table 2 and Fig. 5, the optimized XGBoost model achieved a training accuracy of 100%, a test accuracy of 98.49%, and an AUC score of 0.988. These results robustly validate the critical role of hyperparameter optimization in improving both the generalization capability

and predictive accuracy of the model.

Table 2. XGBoost Model Performance

	Test accuracy	AUC Score
XGBoost	98.49%	0.99

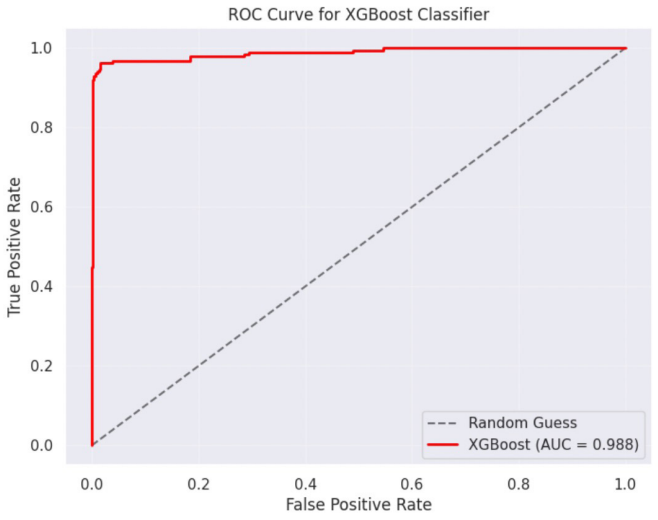

Fig. 5 AUC Score (Picture credit: Original)

3.3 Feature Importance Analysis

Based on the optimized XGBoost model, this study conducted an importance analysis of key predictive features for e-commerce customer churn. As illustrated in Fig. 6, the feature importance ranking reveals that "Tenure" and "Complain" constitute two core determinants of churn, demonstrating significantly higher importance weights compared to other variables.

The predominance of "Tenure" as the most critical predictive feature indicates that the duration of customer-platform relationship establishment plays a pivotal role in churn behavior. Customers with shorter tenure exhibit elevated churn risks, potentially due to unestablished usage patterns or platform loyalty, whereas long-term customers demonstrate higher retention rates and satisfaction levels with comparatively lower churn probability. "Complain" as the second most influential feature underscores the critical role of service quality in maintaining customer relationships. Customers who have filed complaints show substantially increased churn propensity, highlighting the essential nature of timely complaint resolution and effective remediation of negative experiences for customer retention.

ISSN 2959-6130

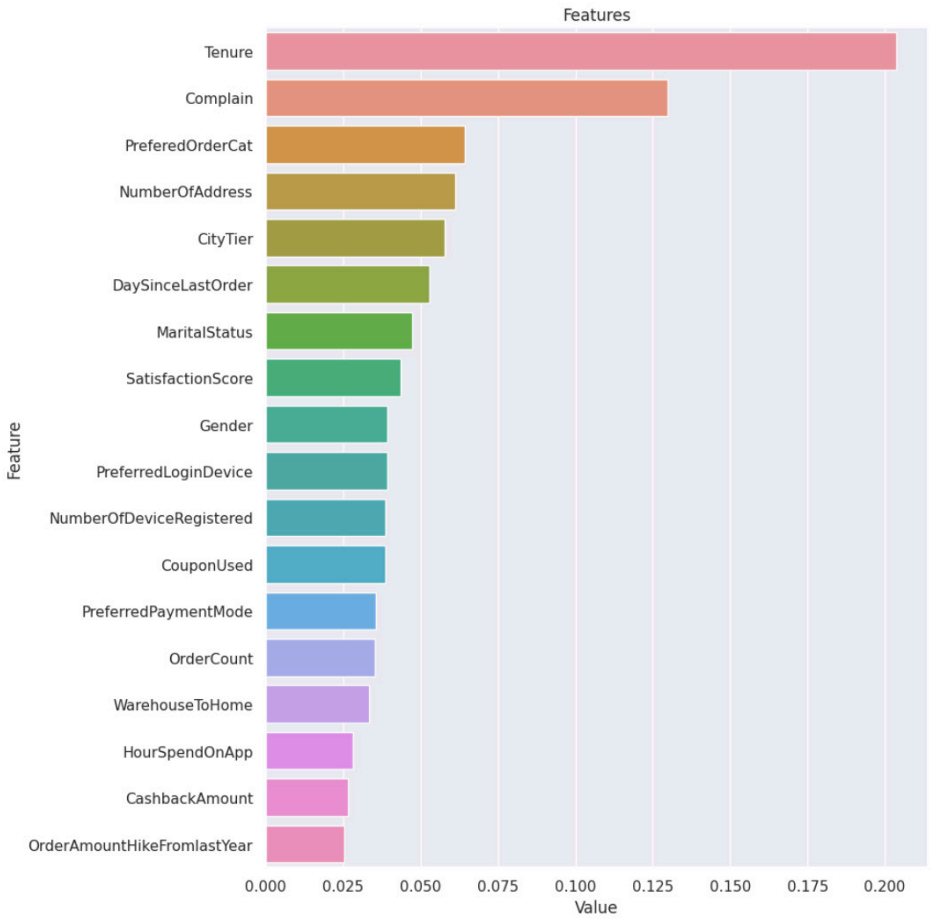

Fig. 6 Feature Importance (Picture credit: Original)

The feature importance analysis not only enhances model interpretability but also provides strategic guidance for e-commerce enterprises to optimize customer retention initiatives. Organizations should prioritize improving onboarding processes for new users, enhancing early-stage service experiences to foster loyalty, while simultaneously establishing robust complaint response and remediation mechanisms to effectively mitigate customer churn.

4. Conclusion

This work constructed an e-commerce customer churn prediction model based on the XGBoost algorithm, systematically addressing challenges such as data imbalance, missing value handling, and model interpretability. By employing median imputation, oversampling techniques, and multi-dimensional feature selection, this paper established a robust and interpretable prediction framework. Experimental results demonstrated that the XGBoost model significantly outperformed logistic regression and decision tree models in terms of prediction performance. After systematic hyperparameter optimization, the model achieved an accuracy of 98.49% and an AUC value of 0.99 on the test set, showcasing excellent generalization ability. Feature importance analysis further revealed that "Tenure" and "Complain" were the core factors influencing customer churn. The limitations of this study lie in the data being sourced from a single platform and the limited sample size. Future work will consider incorporating multi-source data, exploring advanced methods such as deep learning, and extending the model to real-time churn warning scenarios to further enhance its applicability and stability.

References

[1] Liu YT. Customer churn prediction in e-commerce based on machine learning (in Chinese). Southwest University; 2023.

[2] Zheng K, Liang Z. The impact of CNN MHAM-enhanced WRF and BPNN models for user behavior prediction. Sci Rep. 2025;15(1):29999-29999.

[3] Phumchusri N, Amornvetchayakul P. Machine learning models for predicting customer churn: a case study in a software-as-a-service inventory management company. Int J Bus

Intell Data Min. 2024;24(1):74-106.
[4] Boozary P, et al. Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction. Int J Inf Manag Data Insights. 2025;5(1):100331-100331.
[5] Shaik S, et al. Experimental data analysis to recognize and visualise the factors contributing to bank customer churn prediction using ensemble learning models. Asian J Res Comput Sci. 2025;18(5):480-495.
[6] Pflughoeft KA, Butz NT, Corbley A. Customer churn prediction for fixed wireless access: The case of a regional internet service provider. Telecommun Policy. 2025;49(4):102929-102929.
[7] Xiao Q. Research on customer churn prediction based on machine learning (in Chinese). Chongqing University of Technology; 2022.
[8] Kaggle. Ecommerce customer churn analysis and prediction dataset [Internet]. Available from: https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/data
[9] Ashbaugh L, Zhang Y. A comparative study of sentiment analysis on customer reviews using machine learning and deep learning. Computers. 2024;13(12):340-340.
[10] Liu C, et al. Do individual differences in self-regulated learning predict digital health intervention engagement? A decision tree analysis approach. J Technol Behav Sci. 2025;Epub ahead of print:1-10.