Journal of
Theoretical and Applied Electronic Commerce Research

MDPI

Article

Feature Selection, Model Comparison, and Explainability for E-Commerce Customer Repurchase Prediction: A Robustness-Validated Empirical Study

Xiaoling You, Nipaporn Chutiman and Piyapatr Busababodhin *

Department of Mathematics, Faculty of Science, Mahasarakham University, Maha Sarakham 44150, Thailand; nipaporn.c@msu.ac.th (N.C.)
* Correspondence: piyapatr.b@msu.ac.th; Tel.: +66-92-542-6396

Abstract

Predicting customer repurchase behavior is critical in e-commerce analytics, yet the contributions of feature engineering and model selection remain insufficiently understood. This study investigates how feature engineering and algorithm choice influence repurchase prediction performance using a large-scale Tmall dataset (228,867 user–merchant records; 6.01% positive rate). Six feature-selection methods, including mutual information (MI), Light Gradient Boosting Machine (LightGBM) gain importance, stepwise logistic regression, least absolute shrinkage and selection operator (Lasso), ridge regression, and Elastic Net, were applied to 37 candidate features. All methods identified merchant historical repurchase rate as the strongest predictor. Seven feature schemes-six selected subsets plus the full set as benchmark-were evaluated via single-split testing and fifty-replication stability analysis. The 28-feature Elastic Net subset matched or exceeded the full set and ranked first in 15 of 50 replications. Seven classifiers were compared under class weighting, which outperformed unweighted training and the Synthetic Minority Oversampling Technique (SMOTE). LightGBM achieved the highest area under the receiver operating characteristic curve (AUC = 0.7222). SHapley Additive exPlanations (SHAP) analysis confirmed the dominant effect of merchant repurchase rate. The findings highlight that feature engineering, particularly merchant-level historical features, contributes more than algorithm selection, informing e-commerce practitioners, data scientists, and marketing researchers.

Keywords: customer repurchase prediction; feature optimization; explainable machine learning; feature selection; e-commerce analytics

Check for updates

Academic Editor: Hua Pang

Received: 26 August 2026
Revised: 20 September 2026
Accepted: 28 September 2026
Published: 3 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Repeat-purchase behavior has long been recognized as a critical determinant of customer loyalty, customer lifetime value, and firm profitability in marketing and electronic commerce research [1–3]. With the rapid development of online platforms, customer retention has become increasingly important because acquiring new customers is often more costly than maintaining existing relationships [4,5]. For e-commerce platforms and merchants, accurately identifying customers with a high likelihood of repurchase enables more efficient allocation of marketing resources, personalized retention strategies, and improved customer relationship management. However, the effectiveness of repurchase prediction depends not only on the selection of predictive algorithms but also on the quality of behavioral representations used to describe customers, products, brands, and merchants.

Repurchase behavior is inherently heterogeneous and may vary according to not only customer characteristics but also the types of products, brands, and merchants involved in previous transactions [6,7]. For example, frequently purchased products may exhibit different repurchase patterns from durable or occasional-use products, while brand loyalty and merchant attractiveness may further influence customers' repurchase decisions. Therefore, effective repurchase prediction requires comprehensive behavioral representations that capture multiple dimensions of customer–product–merchant interactions rather than relying solely on demographic characteristics. In this study, such heterogeneity is considered through extensive feature engineering. Specifically, the proposed feature set incorporates customer-level characteristics (e.g., age and gender), product-category characteristics (e.g., category diversity and category concentration), brand-related preferences (e.g., brand exploration and brand loyalty intensity), item-level behaviors (e.g., item exploration patterns), and merchant-level relationship characteristics (e.g., merchant interaction patterns and historical merchant repurchase rate). Although this study does not develop category-specific prediction models, it integrates product and merchant heterogeneity into a general repurchase prediction framework.

A substantial body of research has applied machine learning approaches to customer purchase and repurchase prediction in e-commerce environments [8]. Traditional statistical methods, such as logistic regression, have been widely used because of their interpretability and statistical advantages [9]. More recently, advanced machine learning methods, including ensemble learning and deep learning approaches, have demonstrated improved predictive capability by capturing nonlinear relationships and complex behavioral patterns [10–12]. Because customer repurchase decisions are influenced by the interaction of heterogeneous factors, including customer behavior, product preference, and merchant relationship characteristics, predictive models that can capture nonlinear interactions may provide advantages over traditional linear approaches. For example, deep forest architectures have been applied to consumer repurchase prediction [13,14], while recurrent neural network approaches, such as long short-term memory (LSTM) networks, have been used to model sequential purchasing behaviors [15]. However, existing studies have primarily focused on identifying high-performing algorithms based on predictive accuracy or area under the receiver operating characteristic curve (AUC), while the relative contribution of feature engineering and algorithm selection remains insufficiently explored.

Feature engineering and feature selection are fundamental components of predictive modeling, particularly in customer analytics where behavioral data are high-dimensional and heterogeneous [16]. Most existing studies rely on a single feature-selection approach or evaluate selected features using only one data split, which may lead to unstable conclusions caused by sampling variability [17]. Moreover, feature importance is often reported as a ranking without quantifying the marginal contribution of individual predictors or explaining their practical implications [18]. Explainable artificial intelligence (XAI) techniques, such as SHapley Additive exPlanations (SHAP), provide opportunities to interpret complex machine learning models and translate predictive outputs into actionable managerial insights [19,20]. However, an integrated framework combining systematic feature-selection comparison, robustness evaluation, model comparison, and explainable interpretation remains limited in customer repurchase prediction research. Therefore, this study addresses two research questions: (RQ1) which customer, product, brand, and merchant-related features provide the most reliable predictive information for customer repurchase behavior, and whether a reduced feature subset can achieve comparable predictive performance to the full feature set; and (RQ2) which machine learning classifier provides reliable predictive performance under imbalanced e-commerce data conditions, and how model predictions can be interpreted from a managerial perspective.

To address these gaps, this study develops an explainable machine learning framework for customer repurchase prediction using a large-scale Tmall e-commerce dataset, examining how feature engineering and model selection influence predictive performance. Specifically, this study makes four contributions.

First, addressing RQ1, rather than relying on a single criterion, this study systematically compares six independent feature-selection methods-mutual information (MI), Light Gradient Boosting Machine (LightGBM) gain importance, stepwise logistic regression, least absolute shrinkage and selection operator (Lasso), ridge regression, and Elastic Net regularization-applied to 37 candidate features, yielding cross-method convergent evidence for the identified predictors.

Second, also addressing RQ1, this study evaluates whether feature reduction can maintain or improve predictive performance relative to the full feature set. Seven feature configurations-six selected subsets and the original 37-feature set as a benchmark-are compared through single-split evaluation and a fifty-replication stability analysis, providing a more reliable assessment of feature-selection outcomes than a single train-test split allows.

Third, addressing RQ2, this study compares seven classifiers spanning linear, tree-based, ensemble, and neural network paradigms under imbalanced learning conditions, and applies SHAP-based interpretation to quantify feature contributions, moving beyond the static, aggregate rankings typical of prior studies.

Fourth, this study links machine learning findings with customer loyalty and customer lifetime-value perspectives by interpreting the dominant role of merchant-level historical features, thereby providing a theoretically informed explanation of predictive patterns and a direct empirical test of H1 rather than treating feature importance as purely statistical outcomes.

The empirical results show that merchant historical repurchase rate is consistently the most influential predictor across all six feature-selection methods, supporting H1 regarding the dominant role of merchant-level relational characteristics in explaining repurchase behavior. The 28-feature Elastic Net subset matches or exceeds the full feature set and demonstrates strong stability across repeated experiments. Among the classifiers, LightGBM achieves the highest predictive performance, closely followed by three other models within a narrow margin, while SHapley Additive exPlanations (SHAP) analysis reveals the dominant, near-linear contribution of merchant-level historical features. These findings indicate that feature engineering contributes more to predictive performance than algorithm selection, and that this merchant-driven pattern aligns with a relational view of customer loyalty, offering practical guidance for e-commerce managers, data scientists, and researchers.

The remainder of this paper is organized as follows. Section 2 reviews related literature and theoretical foundations. Section 3 describes the dataset, feature engineering procedures, machine learning methods, and evaluation framework. Section 4 presents the empirical results, including feature-selection comparisons, model evaluation, and interpretability analysis. Section 5 discusses theoretical contributions, managerial implications, and limitations. Section 6 concludes the study.

2. Literature Review

2.1. Theoretical Foundations: Customer Loyalty, Perceived Value, and Customer Lifetime Value

Repeat purchase behavior is not merely a data-mining outcome variable but a construct with deep roots in marketing and consumer behavior research. Customer loyalty theory distinguishes behavioral loyalty (i.e., repeated purchasing behavior) from attitudinal loyalty (i.e., psychological commitment toward a brand or merchant), suggesting that sustained repurchase reflects an enduring and favorable relationship between customers

and sellers rather than simple purchase repetition caused by convenience or inertia [21]. From this perspective, repurchase behavior is shaped not only by individual customer characteristics but also by relationship-related factors associated with merchants, implying that effective predictive models should incorporate both customer-level behavioral patterns and merchant-level relational characteristics.

A related theoretical foundation is customer perceived value theory, which defines perceived value as the consumer's overall evaluation of the benefits received relative to the costs incurred during an exchange relationship [22]. This perspective suggests that repurchase decisions are influenced by both transaction-specific attributes (e.g., product characteristics, price, and quality) and accumulated relational experiences (e.g., trust, service consistency, and satisfaction) [23]. In an e-commerce environment, such relational experiences are reflected not only in individual transactions but also in aggregated behavioral histories involving customers, products, brands, and merchants. Therefore, comprehensive behavioral representations that capture multiple dimensions of customer–merchant interactions may provide more informative signals for predicting future repurchase behavior.

Customer lifetime value (CLV) theory further explains the managerial importance of repeat purchase prediction by viewing repeated transactions as an important indicator of future customer profitability [24]. From this perspective, accurately identifying customers with a higher likelihood of repurchase enables firms to allocate retention resources more effectively and prioritize customer-merchant relationships with greater expected long-term value rather than applying uniform retention strategies [25]. Thus, repurchase prediction represents not only a technical modeling problem but also a strategic decision-support task for improving customer relationship management.

These theoretical perspectives collectively suggest that customer repurchase behavior may emerge from the interaction between individual behavioral tendencies and relationship-level characteristics. However, existing machine learning studies have rarely examined whether the predictive signals identified by data-driven models correspond to these established theoretical perspectives, particularly regarding the relative importance of customer-level behaviors versus merchant-level relational characteristics. This motivates further investigation into how behavioral features should be constructed and evaluated within an explainable repurchase prediction framework.

2.2. Repurchase Prediction in E-Commerce

Repeat-buyer prediction gained particular prominence following the release of the Tmall repeat-buyer dataset and the associated KDD Cup challenge [26], which established a widely used benchmark for evaluating repurchase prediction methods and remains the data source underlying the present study. Early research had already formulated repeat purchase prediction as a distinct analytical problem within e-commerce. Tian et al. [27] developed a practical predictive model for repeat purchasing behavior in consumer-to-consumer (C2C) e-commerce environments, demonstrating the feasibility of using behavioral data to support seller-side decision-making.

Building on this foundation, subsequent studies have applied various machine learning methods to customer repurchase prediction. Chou et al. [28] compared traditional probabilistic models with machine learning approaches for customer repurchase prediction and demonstrated that machine learning models can effectively exploit high-dimensional behavioral features to improve predictive performance and provide interpretable insights, while Yang et al. [29] proposed a hybrid random forest-Light Gradient Boosting Machine (RF-LightGBM) framework for predicting repurchase behavior in community e-commerce environments. Zhang et al. [30] further combined feature engineering with ensemble learn-

ing and reported that carefully constructed behavioral features contributed substantially to predictive performance.

Beyond e-commerce-specific applications, earlier studies in direct marketing and consumer analytics also demonstrated the ability of advanced modeling approaches to capture nonlinear patterns in repeat purchasing behavior. Baesens et al. [31] applied Bayesian neural network methods to model repeat-purchase incidence, providing early evidence that neural architectures could identify complex behavioral patterns beyond conventional parametric models. Meade and Islam [32] explored copula-based approaches for modeling repeat purchase behavior and demonstrated the potential of alternative statistical frameworks for capturing dependence structures in customer purchasing processes.

More recent studies have extended predictive modeling approaches to richer e-commerce behavioral data. Kuric et al. [33] examined the contribution of low-level interaction signals, such as clickstream information, to repeat purchase prediction, while Kuo and Cheng [34] incorporated repurchase considerations into content-based recommendation systems. More broadly, machine learning-based analysis of e-commerce consumer data and behavior has been discussed by Al Rahib et al. [35] and Li [36], reflecting the continued expansion of applied machine learning across e-commerce analytics tasks beyond repurchase prediction specifically.

Despite these advances, existing studies have primarily focused on improving predictive performance through increasingly sophisticated algorithms. Comparatively less attention has been paid to systematically evaluating the relative contribution of feature engineering and algorithm selection, determining whether predictive improvements arise primarily from better behavioral representations or from more advanced classifiers.

2.3. Feature Engineering and Automated Feature Construction

Feature engineering plays a fundamental role in customer analytics because raw transactional records often contain limited direct information about customer intentions and relationship characteristics. Effective behavioral representation requires transforming fragmented interaction data into meaningful indicators that capture customer activity intensity, preference patterns, temporal engagement, and customer-merchant relationships.

Early research on relational data mining introduced aggregation-based approaches for transforming complex relational structures into flat feature representations suitable for conventional machine learning algorithms. Knobbe et al. [37] proposed propositionalisation techniques for extracting predictive attributes from relational databases, providing an important foundation for feature construction from behavioral records. Building on this idea, Kanter and Veeramachaneni [38] introduced Deep Feature Synthesis, an automated framework for generating candidate features from relational data, while Katz et al. [39] proposed ExploreKit to automate feature generation and selection processes.

Although automated feature construction approaches have attracted increasing attention, manually designed behavioral features remain widely used in applied e-commerce analytics because they allow researchers to incorporate domain knowledge into predictive representations. In particular, aggregated features such as activity frequency, behavioral diversity, preference concentration, temporal engagement, and merchant-level historical performance can capture latent aspects of customer relationships that may not be directly observable from individual transactions.

In the context of repurchase prediction, Zhang et al. [30] demonstrated that engineered behavioral features can substantially improve model performance. However, existing studies often rely on a single feature-selection strategy or report feature importance based on one predictive model, making it difficult to determine whether identified predictors are robust across different selection criteria. Therefore, systematic comparison of mul-

tiple feature-selection approaches and evaluation of feature stability remain important methodological challenges.

2.4. Class Imbalance and Cost-Sensitive Learning

Customer repurchase prediction is typically characterized by severe class imbalance because repeat purchasers usually represent only a small proportion of all customers. Such imbalance can bias predictive models toward the majority class and reduce their ability to identify potential repeat buyers, making appropriate imbalance-handling strategies essential.

Cost-sensitive learning provides one major approach for addressing this issue by assigning different misclassification costs to different classes. Elkan [40] established a theoretical foundation for cost-sensitive learning by demonstrating the relationship between cost weighting and classification decision boundaries, while Zadrozny et al. [41] further developed cost-proportionate weighting methods for practical applications. These approaches motivate the use of class-weighted learning strategies, which adjust model training objectives without modifying the original data distribution.

An alternative approach involves resampling methods that modify the training data distribution. The Synthetic Minority Oversampling Technique (SMOTE) [42] generates synthetic minority-class observations by interpolating between existing minority samples and their nearest neighbors, and has become one of the most widely applied oversampling techniques in imbalanced learning. He and Garcia [43] provided a comprehensive review of imbalanced learning methods and highlighted that resampling and reweighting approaches may produce different effects on decision boundaries and model generalization.

Although both strategies have been widely applied, their relative effectiveness can vary depending on dataset characteristics and model structures. Therefore, empirical comparison of imbalance-handling strategies remains necessary in specific predictive contexts such as e-commerce repurchase prediction.

2.5. Research Gaps, Research Questions, and Exploratory Hypothesis

Based on the above literature review, three major research gaps are identified.

First, although previous studies have successfully applied machine learning algorithms to repurchase prediction, most have emphasized improving predictive accuracy through algorithm development. The relative contribution of feature engineering and model selection has rarely been systematically examined within a unified framework. In particular, few studies have compared multiple independent feature-selection approaches to determine whether important predictors remain consistent across different methodological perspectives.

Second, existing feature-selection studies often evaluate selected features using a single train-test split, which may lead to unstable conclusions caused by sampling variability. Therefore, robustness evaluation through repeated experiments is necessary to determine whether selected feature subsets generalize beyond a specific data partition.

Third, while predictive models can identify important features, many studies report feature importance only as static rankings without explaining how these predictors relate to customer relationship theories or managerial decision-making. Explainable artificial intelligence (XAI) approaches, such as SHapley Additive exPlanations (SHAP), provide opportunities to bridge predictive modeling and practical interpretation.

To address these gaps, this study focuses on two research questions:

RQ1: Which customer-, product-, brand-, and merchant-related features provide the most reliable predictive information for customer repurchase behavior, and can a reduced feature subset achieve comparable or superior predictive performance compared with the complete feature set?

RQ2: Which machine learning classifier provides reliable predictive performance under imbalanced e-commerce data conditions, and how can model predictions be interpreted from a managerial perspective?

Drawing on customer loyalty theory and the relational perspective of customer retention discussed in Section 2.1, this study further proposes an exploratory hypothesis related to the first research question:

H1. *Merchant-level historical characteristics provide stronger predictive information for customer repurchase behavior than individual customer behavioral characteristics.*

This hypothesis does not imply a causal relationship; rather, it examines whether machine learning-derived predictive patterns are consistent with the theoretical perspective that customer retention is influenced not only by individual behavioral tendencies but also by relationship-level characteristics between customers and merchants.

No corresponding hypothesis is proposed for RQ2. Unlike the feature-importance question addressed by RQ1, which is directly informed by loyalty theory's prediction regarding the relative importance of merchant-level characteristics, the relative performance of competing machine learning classifiers on this task does not follow from an equally well-established theoretical prediction. The comparison across classifiers is therefore treated as an open empirical question rather than a theory-driven hypothesis, and RQ2 is addressed directly through the classifier evaluation and interpretation procedures described below.

The research questions and hypothesis are empirically examined through multi-method feature-selection comparison, repeated stability evaluation, classifier performance comparison under imbalanced learning conditions, and SHAP-based model interpretation analyses presented in Section 4.

3. Materials and Methods

3.1. Data Source, Preprocessing, and Feature Engineering

Tmall (Tmall.com) is a large business-to-consumer e-commerce platform operated by Alibaba Group, on which registered merchants sell branded products directly to consumers. The platform is one of the largest e-commerce marketplaces in China and hosts recurring, large-scale promotional campaigns that generate substantial transaction volume. The dataset used in this study is derived from Tmall user behavior logs collected during one such major promotional campaign period, comprising 228,867 user–merchant interaction records.

The dataset was selected because it represents a large-scale and publicly available benchmark for customer repurchase prediction, originally released through the Alibaba Cloud Tianchi platform in conjunction with the IJCAI-15 repeat-buyer prediction challenge. The campaign-period data provide a sufficiently large volume of customer-merchant interaction records, diverse behavioral signals, and a standardized evaluation setting, enabling reproducible model development and comparison with previous studies using the same benchmark dataset. Although the data were collected during a major promotional period, this setting is particularly valuable for studying repurchase prediction because promotional events generate intensive customer–merchant interactions and provide sufficient behavioral information for identifying patterns associated with subsequent purchasing decisions.

Each record aggregates a user's interactions with a given merchant, including click, add-to-cart, purchase, and favorite actions, together with the user's demographic attributes (age range and gender). The binary outcome variable indicates whether the user made a repeat purchase from the same merchant within a fixed observation window following an initial purchase.

The positive class (repeat purchase) accounts for 6.01% of all records, reflecting the severe class imbalance characteristic of repurchase prediction tasks. The dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve the positive class ratio across all three subsets.

Based on the 10 original behavioral and demographic features, an additional 25 statistical features, ratio features, and concentration features were engineered to capture the relative, rather than absolute, intensity and concentration of user behavior-for example, the proportion of a user's total interactions directed at their single most-visited merchant, and the ratio of distinct merchants interacted with relative to total interaction count. Two further features were constructed via target encoding: a merchant-level historical repurchase rate and an age-by-gender segment-level historical repurchase rate. Both target-encoded features were computed using five-fold out-of-fold encoding on the training set only, with validation and test set values derived by applying training-set statistics, thereby preventing label leakage. The resulting feature set comprises 37 variables in total. The complete processed dataset, comprising all 37 engineered features together with the corresponding train/validation/test partitions, is provided in the Supplementary Materials (Data S1). As shown in Table 1.

Table 1. Defines all 37 engineered features referenced in this study.

Feature Name	Definition
age_range	User's age bracket (demographic attribute).
gender	User's gender (demographic attribute).
user_cnt	Total number of historical actions between the user and the merchant.
user_cnt_0	Number of click actions between the user and the merchant.
user_cnt_1	Number of add-to-cart actions between the user and the merchant.
user_cnt_2	Number of favorite/collect actions between the user and the merchant.
user_cnt_3	Number of historical purchase actions between the user and the merchant.
action_type_1_cnt	Number of actions belonging to a specific behavioral stage, across the user's full history.
avg_actions_per_day	Average daily interaction frequency between the user and the merchant.
time_stamp_nunique	Number of distinct active days between the user and the merchant.
seller_nunique	Number of distinct merchants the user has interacted with (across full history).
cat_nunique	Number of distinct product categories the user has interacted with.
brand_nunique	Number of distinct brands the user has interacted with.
item_nunique	Number of distinct items the user has interacted with.
action_type_nunique	Number of distinct action types between the user and the merchant (max.4: click, cart, favorite, purchase).
seller_nunique_0	Number of distinct merchants interacted with, under a specific action-type condition.
seller_most_1_cnt	Number of actions directed at the user's single most-interacted merchant (full history).
cat_most_1_cnt	Number of actions directed at the user's single most-preferred category.
brand_most_1_cnt	Number of actions directed at the user's single most-preferred brand.
ratio_seller_most1	seller_most_1_cnt divided by the user's total historical actions.
ratio_cat_most1	cat_most_1_cnt divided by the user's total historical actions.
ratio_brand_most1	brand_most_1_cnt divided by the user's total historical actions.
ratio_action1	Share of a specific action type among the user–merchant pair's total actions.

Table 1. *Cont.*

Feature Name	Definition
seller_concentration	Degree of concentration of the user's actions across merchants (full history).
cat_concentration	Degree of concentration of the user's actions across categories.
brand_concentration	Degree of concentration of the user's actions across brands.
ratio_seller_nunique	seller_nunique divided by the user's total historical actions.
ratio_cat_nunique	cat_nunique divided by the user's total historical actions.
ratio_brand_nunique	brand_nunique divided by the user's total historical actions.
ratio_item_nunique	item_nunique divided by the user's total historical actions.
time_stamp_max	Timestamp of the most recent action between the user and the merchant.
time_stamp_min	Timestamp of the earliest action between the user and the merchant.
time_stamp_std	Standard deviation of action timestamps between the user and the merchant.
time_stamp_range	time_stamp_max minus time_stamp_min (duration of the user–merchant relationship).
activity_density	time_stamp_nunique divided by time_stamp_range (continuity vs.burstiness of engagement).
merchant_repurchase_rate	Merchant-level historical repurchase rate (out-of-fold target encoding).
age_gender_te	Historical repurchase rate of the user's age-by-gender segment (out-of-fold target encoding).

3.2. Feature Selection Criteria

Two independent feature-selection criteria were applied to the 37-feature set. Mutual information, a model-agnostic filter criterion, was computed between each feature and the binary outcome on the training set. Gradient boosting gain importance, an embedded criterion computed from a trained LightGBM classifier, was used to capture non-linear interactions that mutual information, as an approximately univariate measure, cannot detect. For both criteria, a top-k feature selection curve was constructed by incrementally adding features in ranked order and evaluating validation-set (for mutual information) or cross-validation (for gain importance) AUC, in order to identify whether a natural, economical cutoff point exists.

Beyond the two data-driven ranking criteria described above, four regularization-based methods were additionally applied to the 37-feature set to provide an independent, model-based line of validation: backward-elimination stepwise logistic regression, which iteratively removes the least significant predictor until all remaining predictors satisfy $p < 0.05$; L1-regularized (Lasso) logistic regression; L2-regularized (Ridge) logistic regression; and Elastic Net logistic regression, which combines L1 and L2 penalties via a mixing parameter α (following the parameterization of Zou and Hastie [44]). For the three penalized regression methods, the regularization strength (and, for Elastic Net, the value of α) was selected via validation-set AUC over a grid of candidate values. Because Ridge regression does not produce sparse coefficient estimates, a corresponding feature subset was instead obtained by ranking features by absolute coefficient magnitude and selecting the validation-optimal cutoff, analogous to the top-k procedure used for the mutual information and gain importance criteria.

3.3. Class Imbalance Handling

Three strategies for handling class imbalance were compared using an identical LightGBM configuration (100 estimators, default tree depth, applied to the stratified 70% training and 15% validation split, each preserving the 6.01% positive-class ratio) to isolate the effect

of the imbalance-handling strategy itself from other sources of performance variation: no adjustment, oversampling via the Synthetic Minority Oversampling Technique (SMOTE) [42] applied to the training set only, and class weighting (implemented as scale_pos_weight for gradient boosting models and class_weight = "balanced" for scikit-learn estimators). SMOTE was configured with its default number of nearest neighbors (k = 5) and applied to achieve a balanced training distribution (a 1:1 ratio between positive and negative classes), consistent with its standard implementation in the imbalanced-learning literature. The class-weighting strategy follows the cost-sensitive learning framework established by Elkan [40] and extended by Zadrozny et al. [41], under which reweighting the training loss according to class frequency is treated as the theoretically grounded counterpart to resampling the training distribution; this theoretical distinction between resampling-based and reweighting-based approaches, and the conditions under which each is preferable, is discussed further in Section 2.4. All three strategies were evaluated on the same held-out validation set using identical performance metrics, and all subsequent modeling in this study adopts whichever strategy proves superior in this comparison.

3.4. Model Development and Evaluation Protocol

Seven classifiers were developed and compared: logistic regression [45] and a decision tree [46–48] as single-model baselines; random forest [49], XGBoost [50], LightGBM [51], and CatBoost [52] as tree-based ensemble methods; and a multilayer perceptron (MLP) neural network with a single hidden layer, included to introduce a decision-boundary paradigm structurally distinct from both linear and tree-based models. Hyperparameters for each model were tuned via grid or randomized search with three-fold cross-validation on the training set, with final model selection confirmed on the validation set; the test set was used exactly once, after all tuning decisions were finalized, to avoid optimistic bias from repeated test-set evaluation.

3.5. Mathematical Formulation of the Compared Models

All data processing, feature engineering, model development, and statistical analyses in this study were conducted in Python 3.13, using scikit-learn (logistic regression, decision tree, random forest, and the Lasso/Ridge/Elastic Net regularization models), XGBoost, LightGBM, and CatBoost (their respective official Python packages) for the gradient-boosting classifiers, and the SHAP package for the model interpretability analysis reported in Section 4.4.

This section formalizes the classifiers and evaluation criteria introduced in Sections 3.4 and 3.5. The equations are presented only for the key methodological components directly related to this study, including class weighting, representative model formulations, and evaluation metrics.

Class weighting. The Tmall repurchase prediction dataset is characterized by substantial class imbalance, where repeat buyers represent only a small proportion of all customer–merchant interactions. To mitigate the bias toward the majority class, this study adopts class-weighted learning strategies for models that support cost-sensitive optimization. The positive-class weight is calculated according to the ratio between negative and positive training samples:

$$w_{\text{pos}} = \frac{N_{\text{neg}}}{N_{\text{pos}}} = \frac{215{,}111}{13{,}756} \approx 15.64 \tag{1}$$

where N_{neg} and N_{pos} denote the numbers of negative and positive samples in the training set, respectively, and w_{pos} represents the weight assigned to positive samples. In this study, the calculated weight is approximately 15.64, indicating that misclassification of positive

repurchase cases receives a substantially higher penalty during model training. This weighting strategy is implemented through the scale_pos_weight parameter in gradient boosting models and the class_weight = "balanced" option in scikit-learn estimators.

Logistic regression. As a linear baseline because of its interpretability and widespread use in binary classification problems. The model estimates the probability of customer repurchase through a sigmoid transformation of a linear combination of input features:

$$p_i = \sigma\left(w^\top x_i + b\right) = \frac{1}{1 + e^{-(w^\top x_i + b)}} \quad (2)$$

where p_i represents the predicted probability of repurchase for customer i, x_i denotes the feature vector, w represents the vector of model coefficients, b is the intercept term, and $\sigma(\cdot)$ denotes the sigmoid activation function. Because the dataset is imbalanced, logistic regression is trained with class-weighted optimization to increase sensitivity toward minority-class repurchase cases.

Tree-based ensemble models. This study evaluates several tree-based models, including decision tree, random forest, XGBoost, LightGBM, and CatBoost. These models are capable of capturing nonlinear relationships and complex interactions among customer, product, brand, and merchant-related features.

Among these models, gradient boosting algorithms share a common additive learning framework, where each newly added weak learner improves the prediction of the existing ensemble:

$$F_m(x) = F_{m-1}(x) + \eta f_m(x) \quad (3)$$

where $F_m(x)$ denotes the ensemble prediction after the m-th boosting iteration, $F_{m-1}(x)$ represents the prediction from the previous iteration, $f_m(x)$ is the newly added decision tree, and η is the learning rate controlling the contribution of each tree.

Based on this general boosting framework, XGBoost, LightGBM, and CatBoost introduce different improvements in optimization and tree construction. XGBoost incorporates regularization to control model complexity, LightGBM employs histogram-based and leaf-wise tree growth for computational efficiency, and CatBoost applies ordered boosting to reduce potential target leakage when handling categorical and target-encoded variables. These characteristics make them suitable candidates for modeling complex repurchase patterns.

Multilayer perceptron. The neural network is included as a neural network baseline capable of learning nonlinear feature transformations. The hidden layer representation is calculated as:

$$h = \phi(W_1 x + b_1) \quad (4)$$

where h represents the hidden-layer output, W_1 denotes the weight matrix connecting input features to the hidden layer, b_1 represents the bias vector, and $\phi(\cdot)$ is the rectified linear unit (ReLU) activation function. The resulting representation is then passed to an output layer to estimate repurchase probability. Similar to other models, MLP training incorporates class weighting to address the imbalance between repeat buyers and non-repeat buyers.

Evaluation metrics. Model performance is evaluated on the independent test set using threshold-dependent classification metrics, including accuracy, precision, recall, and F1-score, together with the threshold-independent area under the receiver operating characteristic curve (AUC). Because the dataset is highly imbalanced, accuracy alone may provide misleading conclusions; therefore, multiple complementary evaluation metrics are considered.

The classification metrics are calculated based on four fundamental outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP and TN

represent correctly classified positive and negative samples, whereas FP and FN represent incorrectly classified negative and positive samples, respectively.

Accuracy measures the proportion of correctly classified observations among all evaluated samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

Precision evaluates the proportion of correctly identified repeat purchasers among all predicted positive cases:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

Recall measures the proportion of actual repeat purchasers correctly identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

F1-score provides a harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

4. Results

4.1. Class Imbalance Handling Comparison

The comparative performance of the three imbalance-handling strategies on the test set is summarized in Table 2. The no-adjustment baseline achieved a nominally high AUC (0.7213) but a recall of exactly zero, indicating that under the default 0.5 classification threshold, the model failed to identify a single true repeat-purchase customer among 2064 positive test cases-illustrating that AUC alone can be a misleading indicator of practical model utility on severely imbalanced tasks. SMOTE oversampling improved recall (0.3929) but reduced AUC substantially (0.6401), consistent with the interpretation that interpolated synthetic minority samples can distort the learned decision boundary away from the true data distribution. Class weighting achieved an AUC (0.7197) comparable to the no-adjustment baseline while substantially improving recall (0.6289) and achieving the highest F1 score (0.1987) among the three strategies. Class weighting was therefore adopted for all subsequent modeling. To formally assess whether the observed AUC differences reflect genuine performance differences rather than sampling variability, pairwise DeLong tests for correlated ROC curves were conducted. The AUC difference between the no-adjustment baseline and class weighting was not statistically significant ($p = 0.193$), confirming that reweighting the loss function preserves the baseline's discriminative ability, whereas SMOTE's AUC reduction relative to class weighting was highly significant ($p < 0.001$). Figure 1 plots the corresponding ROC curves: the SMOTE curve is visibly and uniformly below the other two, while class weighting tracks the no-adjustment curve almost exactly, confirming that reweighting the loss recovers baseline ranking ability while the recall gain in Table 1 comes from a shifted operating threshold rather than a degraded ROC curve.

Table 2. Comparison of class imbalance handling strategies (test set).

Method	Accuracy	Precision	Recall	F1	AUC	*p*-Value (DeLong vs. Class Weighting)
No adjustment	0.9399	0.0000	0.0000	0.0000	0.7213	0.193
SMOTE oversampling	0.7512	0.1001	0.3929	0.1596	0.6401	<0.001
Class weighting	0.6951	0.1180	0.6289	0.1987	0.7197	—

Note: All threshold-dependent metrics were calculated using a fixed cutoff of 0.5.

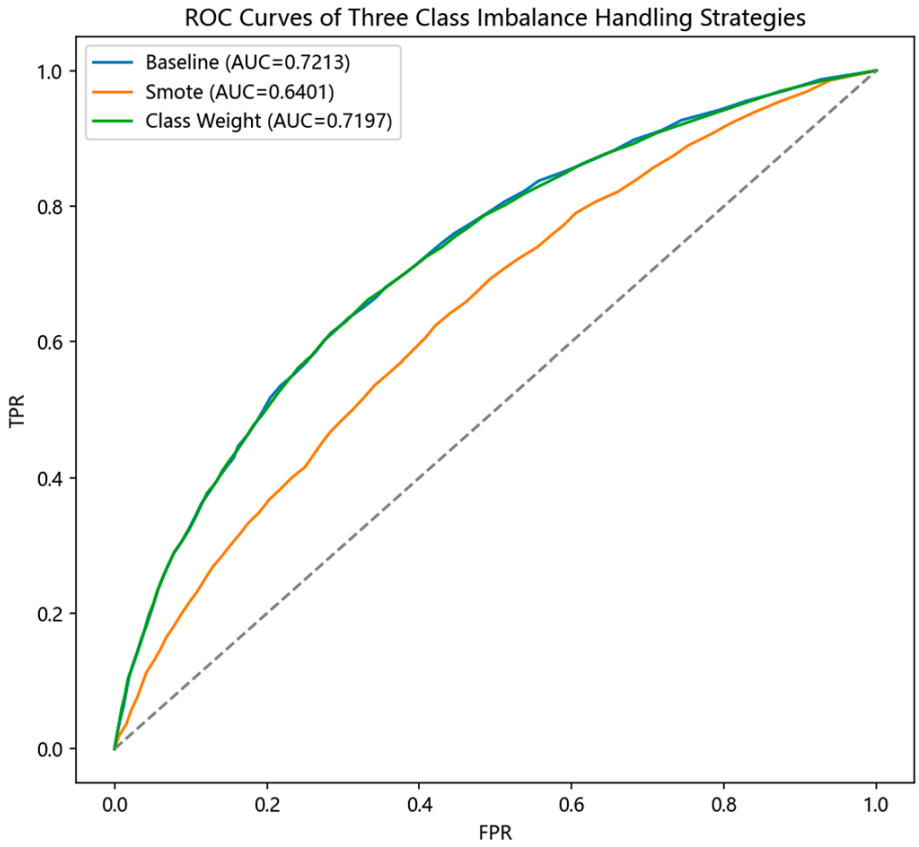


Figure 1. ROC curves comparing three class-imbalance handling strategies on the test set. The dashed diagonal line represents the ROC curve of a random classifier.

4.2. Feature Selection Results

This section presents the empirical findings addressing RQ1: which features carry the greatest predictive weight, and whether a reduced feature subset can match the performance of the full feature set.

The purpose of feature selection is to examine whether this feature system contains informational redundancy, and thereby to determine whether reducing the feature set prior to modeling is warranted. This section applies six feature-selection methods in turn-mutual information, gradient-boosting gain importance, stepwise logistic regression, Lasso regularization, Ridge regularization, and Elastic Net regularization-to assess the importance of the 37 candidate features. The results are shown below.

4.2.1. Feature Selection Based on Mutual Information

Mutual information scores were computed for all 37 features on the training set. Figure 2 illustrates the ranking of all 37 features according to their mutual information scores calculated from the training set. Table 3 presents the top-15 features with the highest mutual information values. The top-ranked features were time_stamp_max (the time of a user's most recent action), time_stamp_range (the temporal span of a user's actions), action_type_nunique (the number of distinct action types a user engaged in), and merchant_repurchase_rate (the merchant's historical repurchase rate). This ranking structure indicates that the temporal distribution of user behavior and the merchant's own repurchase performance are the two categories of information most strongly associated with the repurchase label.

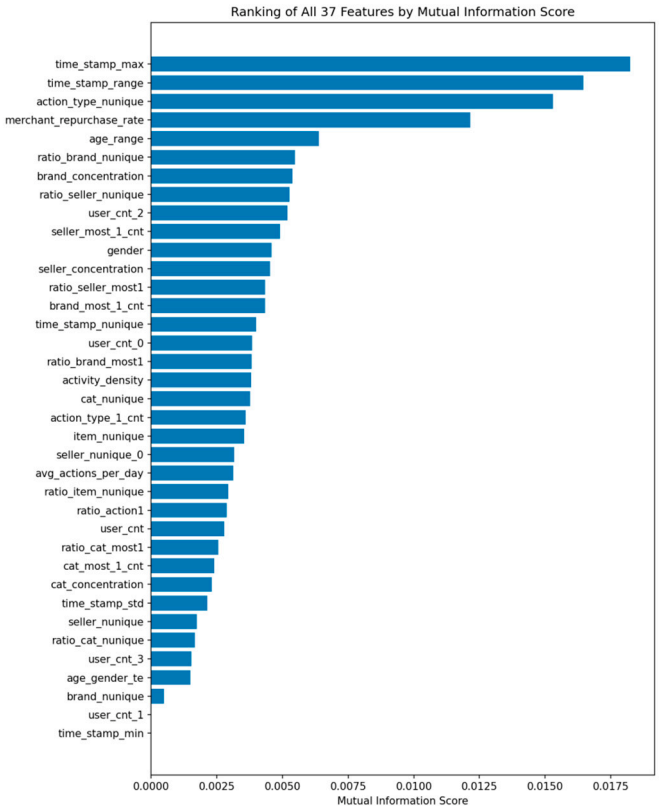


Figure 2. Ranking of all 37 features by mutual information score.

Table 3. Top-15 features ranked by mutual information score.

Rank	Feature	Mutual Information Score
1	time_stamp_max	0.0182
2	time_stamp_range	0.0165
3	action_type_nunique	0.0153
4	merchant_repurchase_rate	0.0121
5	age_range	0.0064
6	ratio_brand_nunique	0.0055
7	brand_concentration	0.0054
8	ratio_seller_nunique	0.0053
9	user_cnt_2	0.0052
10	seller_most_1_cnt	0.0049
11	gender	0.0046
12	seller_concentration	0.0045
13	ratio_seller_most1	0.0043
14	brand_most_1_cnt	0.0043
15	time_stamp_nunique	0.0040

Note: Full definitions of all feature names listed above are provided in Table 1.

To determine a reasonable feature-selection cutoff, features were incrementally added in mutual-information rank order to train a logistic regression model, and AUC was evaluated on an independent validation set at each step, tracing the relationship between the number of features and model discriminative power. As shown in Figure 3, validation AUC rose approximately monotonically with the number of features, reaching a maximum of 0.7143 at k = 25, after which it plateaued without a clearly identifiable inflection point. This pattern indicates that mutual information alone does not yield a sufficiently compelling, economical cutoff point, and that the curve's shape itself suggests that adding more features tends to be beneficial overall.

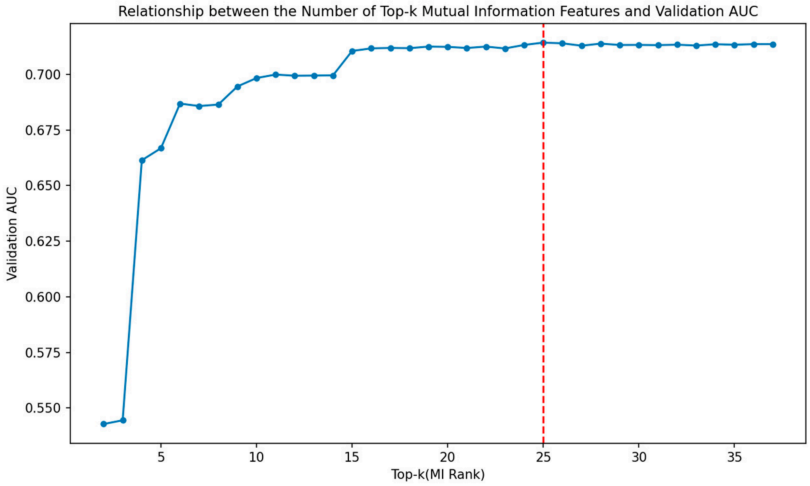


Figure 3. Validation-set AUC as a function of the number of top-k mutual-information-ranked features.

4.2.2. Feature Selection Based on LightGBM Gain Importance

A LightGBM classifier (with class weighting to address imbalance) was trained on the training set, and gain importance was extracted for each feature. Figure 4 illustrates the relative importance ranking of all 37 features based on LightGBM gain importance, while Table 4 presents the top-15 features with the highest gain importance scores. It can be seen that merchant historical repurchase rate accounted for 41.741% of total gain, far exceeding the second-ranked feature, seller_most_1_cnt (9.072%), indicating that a merchant's own historical repurchase level is the single largest source of predictive information for repurchase behavior in this feature system.

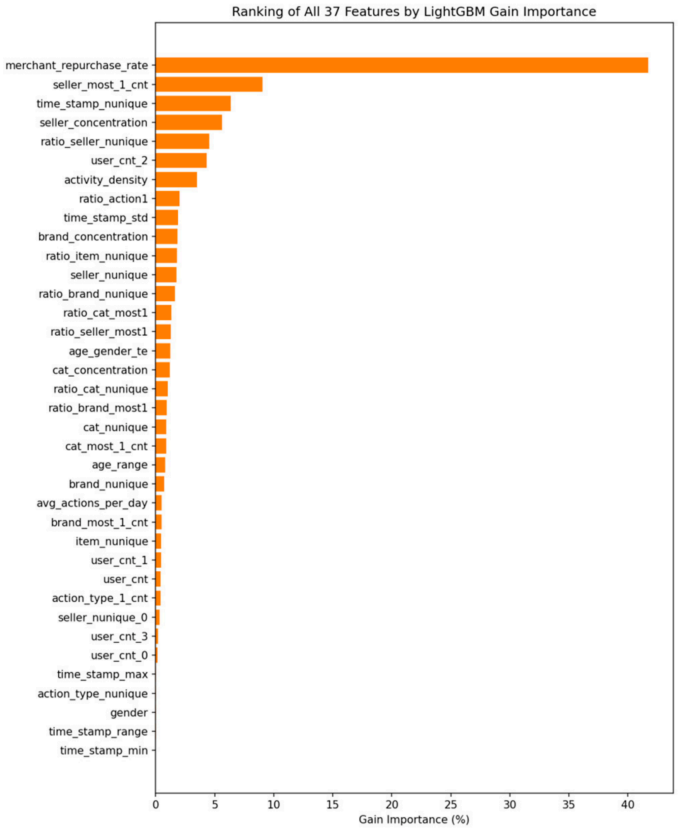


Figure 4. Ranking of all 37 features by LightGBM gain importance.

Table 4. Top-15 features ranked by LightGBM gain importance.

Rank	Feature	Gain Share (%)
1	merchant_repurchase_rate	41.741
2	seller_most_1_cnt	9.072
3	time_stamp_nunique	6.358
4	seller_concentration	5.626
5	ratio_seller_nunique	4.537
6	user_cnt_2	4.302
7	activity_density	3.513
8	ratio_action1	2.037
9	time_stamp_std	1.906
10	brand_concentration	1.859
11	ratio_item_nunique	1.804
12	seller_nunique	1.743
13	ratio_brand_nunique	1.638
14	ratio_cat_most1	1.331
15	ratio_seller_most1	1.297

Note: Full definitions of all feature names listed above are provided in Table 1.

Following the same logic as in Section 4.2.1, features were incrementally added in gain-importance rank order, and the resulting mean AUC was evaluated via 5-fold cross-validation within the training set. Figure 5 illustrates the relationship between cross-validation AUC and the number of top-k features selected according to LightGBM gain importance ranking. Cross-validation AUC rose rapidly to approximately 0.700 by around 8 features, reaching a maximum of 0.7142 at k = 30.

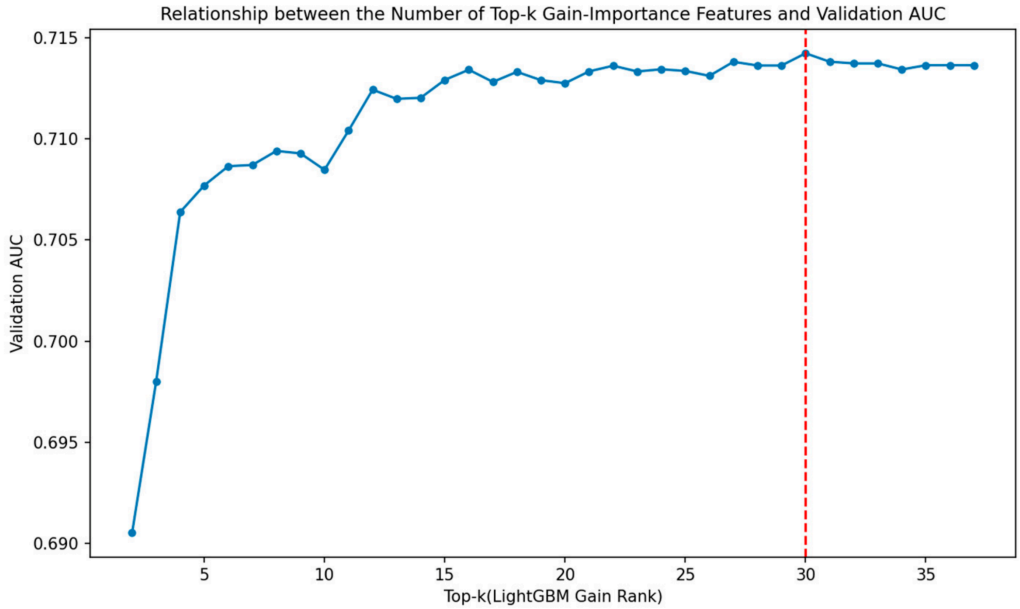


Figure 5. Cross-validation AUC as a function of the number of top-k gain-ranked features.

4.2.3. Feature Selection Based on Stepwise Logistic Regression

Backward-elimination stepwise logistic regression was applied to the standardized 37-feature set, retaining only predictors significant at $p < 0.05$. As shown in Table 5 presents the top-10 statistically significant variables retained by stepwise logistic regression, ranked according to the absolute magnitude of their z-values ($|z|$). The z-value for merchant_repurchase_rate reached 49.77, making it the most strongly significant variable among all 15 retained predictors—consistent with the conclusions established indepen-

dently by the mutual information analysis and the LightGBM gain-importance analysis, providing a third, independent statistical method of cross-validation for this variable's dominant role. It should be noted that several retained variables in Table 5 show odds ratios relatively close to 1 while remaining statistically significant ($p < 0.0001$). This pattern is mainly attributable to the large sample size used in the stepwise selection procedure (160,206 training observations), which enables small coefficients to be estimated with high precision. In logistic regression, statistical significance reflects whether the estimated coefficient differs significantly from zero, whereas the odds ratio represents the magnitude and direction of the association. Therefore, an odds ratio close to 1 does not indicate that a variable lacks predictive information; rather, it suggests that the marginal effect of that variable is relatively limited. This 15-variable subset is included, together with the other selection schemes, in the LightGBM-based performance comparison reported in Section 4.2.7.

Table 5. Top-10 significant variables from stepwise logistic regression (ranked by |z|).

Rank	Variable	Coefficient	z-Value	*p*-Value	Odds Ratio
1	merchant_repurchase_rate	0.4072	49.77	<0.0001	1.503
2	ratio_cat_most1	−0.2236	−12.21	<0.0001	0.800
3	ratio_seller_most1	0.2123	12.04	<0.0001	1.237
4	user_cnt_2	0.1208	11.22	<0.0001	1.128
5	ratio_seller_nunique	−0.2410	−9.80	<0.0001	0.786
6	ratio_item_nunique	0.1078	7.66	<0.0001	1.114
7	age_gender_te	0.0831	7.56	<0.0001	1.087
8	ratio_action1	−0.0996	−7.39	<0.0001	0.905
9	ratio_cat_nunique	−0.1149	−4.88	<0.0001	0.891
10	user_cnt_1	−0.0485	−4.02	0.0001	0.953

Note: Full definitions of all feature names listed above are provided in Table 1.

4.2.4. Feature Selection Based on Lasso Regularization

L1-regularized (Lasso) logistic regression was further applied as an embedded linear feature-selection method. And after calibrating the parameter scale and searching over $C \in \{0.02, 0.01, 0.005, 0.002, 0.001\}$, the validation-optimal configuration was $C = 0.005$ (corresponding to regularization strength $\lambda \approx 0.0012$), achieving a validation AUC of 0.7048 and retaining 25 non-zero coefficients, with 12 variables compressed to zero. Table 6 presents the top-10 features identified by Lasso regression, ranked according to the absolute magnitude of their estimated coefficients. Merchant_repurchase_rate remained the largest-magnitude non-zero coefficient (0.4358), consistent with the conclusions of the three preceding methods. Figure 6 illustrates the Lasso regularization path and the corresponding validation AUC performance under different numbers of non-zero features. The validation AUC reached its maximum when 25 features were retained, indicating that the Lasso-based feature selection procedure identified a 25-feature subset as the optimal balance between model complexity and predictive performance. Therefore, the 25 non-zero coefficient features were selected as the Lasso-based feature subset for subsequent analysis.

Table 6. Top-10 features by Lasso (ranked by |coefficient|).

Rank	Variable	Coefficient
1	merchant_repurchase_rate	0.4358
2	activity_density	0.3598
3	ratio_seller_nunique	−0.2396
4	ratio_seller_most1	0.1849

Table 6. *Cont.*

Rank	Variable	Coefficient
5	ratio_cat_most1	−0.1361
6	ratio_item_nunique	0.1097
7	user_cnt_2	0.1046
8	seller_nunique	−0.0896
9	cat_nunique	−0.0868
10	time_stamp_range	0.0848

Note: Full definitions of all feature names listed above are provided in Table 1.

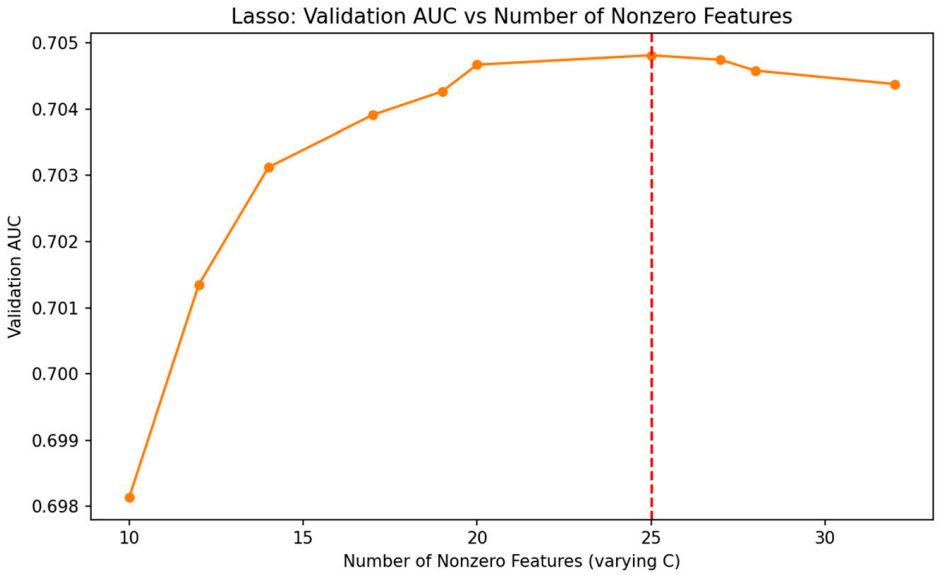


Figure 6. Lasso regularization path: validation AUC as a function of the number of non-zero features.

4.2.5. Feature Selection Based on Ridge Regularization

Unlike Lasso and Elastic Net, Ridge regularization does not shrink coefficients exactly to zero and therefore cannot directly produce a sparse feature subset. Accordingly, Ridge logistic regression was used as a coefficient-based feature ranking method rather than a strict selection approach. The regularization parameter was optimized using the same procedure as other regularized logistic regression models, and features were ranked by the absolute magnitude of their coefficients. The top-10 ranked features are presented in Table 7.

Table 7. Top-10 features by Ridge (ranked by |coefficient|).

Rank	Variable	Coefficient
1	merchant_repurchase_rate	0.4196
2	activity_density	0.2092
3	ratio_seller_nunique	−0.1513
4	ratio_cat_most1	−0.1466
5	ratio_seller_most1	0.1435
6	time_stamp_nunique	0.1416
7	ratio_item_nunique	0.1066
8	seller_nunique	−0.1043
9	user_cnt_2	0.1025
10	brand_nunique	−0.0951

Note: Full definitions of all feature names listed above are provided in Table 1.

To determine the optimal number of retained features, the top-k features from the Ridge ranking were incrementally evaluated using LightGBM on the validation set. As shown in Figure 7, the highest validation AUC (0.7144) was achieved with the top 30 features, indicating that this configuration provided the most effective balance between feature retention and predictive performance.

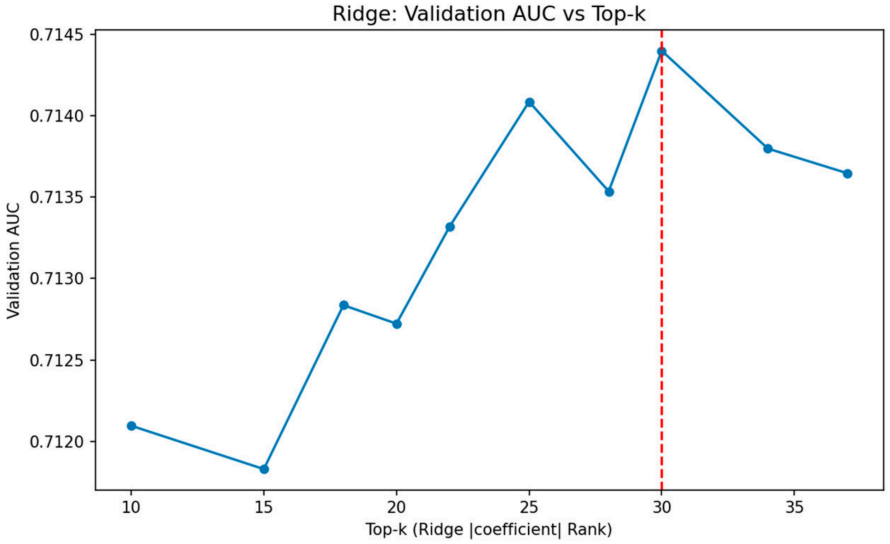


Figure 7. Ridge: validation AUC as a function of Top-k (ranked by |coefficient|).

4.2.6. Feature Selection Based on Elastic Net Regularization

Given the selection instability of Lasso under highly collinear variables, Elastic Net, which combines L1 and L2 penalties, was further applied. A grid search over $\alpha \in \{0.05, 0.15, 0.30, 0.50, 0.70, 0.85\}$ and the calibrated regularization-strength range identified the optimal configuration as $\alpha = 0.15$ and $\lambda \approx 0.0062$ ($C = 0.001$), achieving a validation AUC of 0.7050, the highest among Stepwise, Lasso, and Ridge. This configuration retained 28 non-zero coefficients and eliminated 9 variables. The top-ranked features are presented in Table 8 based on absolute coefficient magnitude. Merchant_repurchase_rate remained the most influential feature (coefficient = 0.4281), confirming its dominance across all six feature-selection methods: mutual information, LightGBM gain importance, stepwise logistic regression, Lasso, Ridge, and Elastic Net.

Table 8. Top-10 features by Elastic Net (ranked by |coefficient|).

Rank	Variable	Coefficient
1	merchant_repurchase_rate	0.4281
2	activity_density	0.2272
3	ratio_seller_nunique	−0.1788
4	ratio_seller_most1	0.1565
5	ratio_cat_most1	−0.1371
6	time_stamp_nunique	0.1250
7	ratio_item_nunique	0.1072
8	user_cnt_2	0.1019
9	seller_nunique	−0.0941
10	ratio_brand_nunique	−0.0838

Note: Full definitions of all feature names listed above are provided in Table 1.

The relationship between the number of retained non-zero features and predictive performance was further examined through validation-set AUC evaluation. As shown in Figure 8, the validation AUC generally increased as more informative features were retained

and reached its maximum when 28 non-zero features were included. Beyond this point, additional features provided limited predictive improvement, suggesting diminishing returns from retaining more variables. Therefore, the 28-feature configuration was selected as the final Elastic Net feature subset.

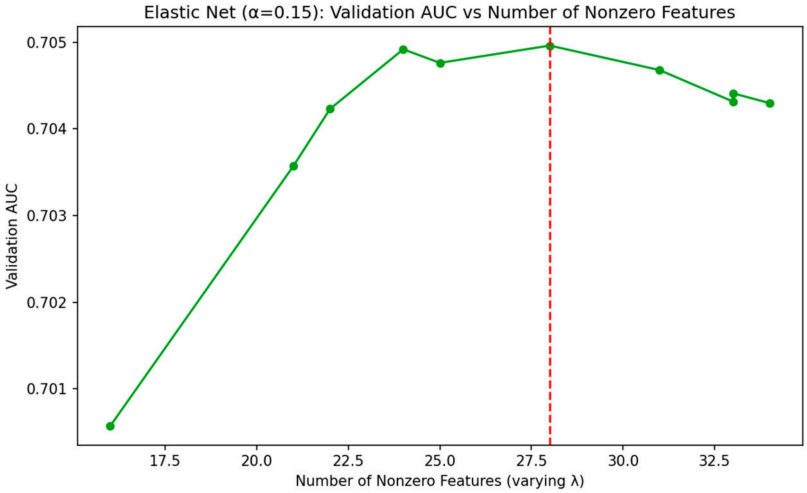


Figure 8. Elastic Net ($\alpha = 0.15$) regularization path: validation AUC as a function of the number of non-zero features.

4.2.7. Comprehensive Comparison of Six Feature-Selection Schemes

To further examine how different feature-selection methods affect predictive performance, the feature subsets obtained from mutual information (MI), LightGBM gain importance, stepwise logistic regression, Lasso, Ridge, and Elastic Net were each compared against the full 37-feature scheme. To ensure a fair comparison across selection strategies, all feature subsets were fed into the same LightGBM configuration used in Section 3.3, with validation-set and test-set AUC computed for each; results are reported in Table 9.

Table 9. Comparison of six feature-selection schemes against the full feature set.

Scheme	Number of Features	Validation AUC	Test AUC
Full feature set	37	0.7136	0.7220
Mutual information (top-25)	25	0.7143	0.7195
LightGBM importance (top-30)	30	0.7142	0.7221
Stepwise logistic regression (15)	15	0.7028	0.7137
Lasso (25)	25	0.7140	0.7203
Elastic Net (28)	28	0.7138	0.7222
Ridge (30)	30	0.7144	0.7218

On the validation set, the full 37-feature scheme achieved an AUC of 0.7136, while all feature-selection schemes maintained comparable performance. Ridge achieved the highest validation AUC (0.7144), followed by mutual information (0.7143) and LightGBM importance (0.7142), with the overall AUC variation limited to approximately 0.002, indicating that feature reduction did not result in meaningful predictive loss.

On the test set, Elastic Net achieved the highest AUC (0.7222), marginally exceeding LightGBM importance (0.7221) and the full-feature benchmark (0.7220). Other selection methods also remained close to the benchmark, with stepwise logistic regression achieving 0.7137 despite retaining only 15 features. These results suggest that the original feature set contains some redundancy and that carefully selected subsets can preserve predictive information while reducing feature complexity.

Among the six methods, Elastic Net, Ridge, and LightGBM importance achieved relatively strong performance, suggesting that both regularization-based methods and tree-based importance criteria can effectively identify informative predictors. Considering predictive performance, feature reduction, and model complexity, the Elastic Net-selected subset was further evaluated through repeated simulation experiments to examine its stability under data perturbations.

4.2.8. Robustness Check: Fifty-Replication Stability Analysis and Final Feature Set Determination

To further evaluate the stability and reliability of different feature-selection methods under data perturbations, a 50-replication simulation analysis was conducted. In each trial, the dataset was randomly re-partitioned, and an identically configured LightGBM model was trained using the feature subset selected by each method. The mean AUC, AUC standard deviation, and the number of times each scheme achieved the highest AUC were recorded to jointly assess predictive performance and stability. The results are summarized in Table 10.

Table 10. Mean AUC, standard deviation, and win count across fifty simulation replications for six feature-selection schemes and the full feature set.

Strategy	Mean AUC	Standard Deviation of AUC	Number of Wins (of 50 Runs)
Elastic Net	0.6543	0.0122	15
All 37 Features	0.6538	0.0133	7
Stepwise Logistic Regression	0.6533	0.0118	4
Ridge	0.6532	0.0131	8
Lasso	0.6530	0.0121	7
LightGBM Importance	0.6528	0.0130	3
Mutual Information	0.6487	0.0156	6

Unlike the fixed-test-set evaluation in Table 9, the simulation analysis estimates expected performance under repeated sample perturbations. Therefore, the slightly lower mean AUC observed across simulations is a normal consequence of evaluating performance across varying data splits rather than evidence of reduced predictive ability, and instead reflects the robustness of each feature-selection strategy.

As shown in Table 10, all feature-selection schemes achieved comparable predictive performance across the 50 simulation trials. Elastic Net obtained the highest mean AUC (0.6543), slightly exceeding the full 37-feature scheme (0.6538), while the other methods showed similarly close performance. These results indicate that the selected subsets retained the essential predictive information, and feature reduction did not materially impair model performance.

The AUC standard deviations across methods ranged from 0.0118 to 0.0156, demonstrating stable performance under different data partitions. Stepwise logistic regression showed the lowest variability (SD = 0.0118), whereas Elastic Net achieved a favorable balance between predictive performance and stability (SD = 0.0122).

In terms of trial-level performance, Elastic Net achieved the highest AUC in 15 out of 50 trials, followed by the full 37-feature scheme and Lasso (both 7 trials). This result suggests that Elastic Net provides a more stable feature subset by considering both individual feature contributions and inter-variable correlations through combined L1 and L2 regularization.

Taken together, the fixed-test-set and simulation results demonstrate that Elastic Net achieved the strongest overall performance, with the highest test AUC (0.7222), highest mean simulation AUC (0.6543), and most frequent wins across repeated trials. Therefore,

the 28-feature Elastic Net subset was selected as the final feature set for subsequent model construction and interpretation.

4.3. Model Comparison Results

This section presents the empirical findings addressing RQ2: which classifier achieves the most reliable predictive performance under class imbalance, and how its predictions can be interpreted from a managerial perspective.

This section evaluates and compares seven classification models using the 28 features selected by Elastic Net in Section 4.2: two single models (logistic regression and decision tree), four tree-based ensemble models (random forest, XGBoost, LightGBM, and CatBoost), and one multilayer perceptron (MLP) neural network. All models were trained under the same class-imbalance correction framework, using either class weighting or an equivalent sample-weighting strategy. Hyperparameters were optimized through 3-fold cross-validation within the training set, and the final model performance was evaluated on the independent test set, which was used only once after all tuning procedures were completed to avoid test-set overfitting.

Different models were tuned according to their algorithmic characteristics. For logistic regression, the regularization strength C was searched over {0.01, 0.1, 1, 10}, with C = 0.01 selected as the optimal value. The decision tree model was optimized over tree depth and minimum leaf size, resulting in max_depth = 5 and min_samples_leaf = 20. The MLP neural network used a single hidden layer, with the number of neurons and L2 regularization coefficient optimized, resulting in 64 hidden neurons and $\alpha = 1 \times 10^{-5}$. For ensemble models, tree complexity, learning rate, and ensemble size were tuned according to model-specific parameter spaces. The final configurations and test-set performance of all seven classifiers are summarized in Table 11.

Table 11. Hyperparameter configurations and test-set performance of seven classifiers.

Model	Optimal Hyperparameters	Imbalance Handling	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	C = 0.01	class_weight = balanced	0.6829	0.1139	0.6303	0.1929	0.7141
Decision Tree	max_depth = 5; min_samples_leaf = 20	class_weight = balanced	0.6489	0.1038	0.6337	0.1783	0.6908
Random Forest	n_estimators = 150; max_depth = 10; min_samples_leaf = 30	class_weight = balanced	0.7536	0.1278	0.5320	0.2061	0.7160
XGBoost	n_estimators = 200; max_depth = 4; learning_rate = 0.02; subsample = 1.0; colsample_bytree = 0.9	scale_pos_weight = 15.6379	0.6753	0.1139	0.6492	0.1938	0.7197
LightGBM	n_estimators = 200; num_leaves = 15; learning_rate = 0.03; min_child_samples = 50	scale_pos_weight = 15.6379	0.6804	0.1160	0.6516	0.1969	0.7222
CatBoost	iterations = 250; depth = 6; learning_rate = 0.05	auto_class_weights = Balanced	0.6813	0.1150	0.6420	0.1950	0.7204
MLP	hidden_layer_sizes = (64,); alpha = 1×10^{-5}	sample_weight equivalent to scale_pos_weight = 15.6379	0.6736	0.1135	0.6507	0.1933	0.7175

The comparative results in Table 11 indicate that model performance varied across different evaluation metrics. LightGBM achieved the highest test-set AUC (0.7222), followed closely by CatBoost (0.7204) and XGBoost (0.7197), demonstrating the advantage of gradient-boosting approaches in capturing nonlinear relationships among heterogeneous behavioral features. Among all models, LightGBM also achieved the highest recall (0.6516), indicating its stronger capability to identify potential repeat purchasers under severe class imbalance. In the context of e-commerce repurchase prediction, recall is particularly important because identifying a larger proportion of actual repeat purchasers

can help merchants improve customer retention activities and reduce the risk of missing valuable returning customers.

Although Random Forest achieved the highest accuracy (0.7536), precision (0.1278), and F1 score (0.2061), its AUC (0.7160) was lower than those of the gradient-boosting models. This difference reflects the influence of class imbalance on threshold-dependent metrics: a model may achieve relatively favorable classification accuracy while providing weaker ranking ability for distinguishing repeat purchasers from non-repeat purchasers. Therefore, accuracy, precision, recall, and F1-score were jointly considered to provide complementary perspectives on practical classification performance. Among these metrics, F1-score is particularly informative in imbalanced settings because it balances precision and recall, reflecting the trade-off between identifying potential repeat purchasers and avoiding excessive marketing efforts toward customers with a lower likelihood of repurchase. However, the absolute F1 values remain relatively low in this study because repeat purchasers represent only a small minority of the customer population (6.01%), making the identification of positive cases inherently challenging. Therefore, F1-score is interpreted primarily as a measure of the balance between customer identification and false-positive control rather than as an absolute indicator of model quality. In contrast, AUC is used as the primary evaluation metric because it evaluates the model's ranking ability across different classification thresholds and is less affected by the specific decision threshold selected under imbalanced conditions.

The logistic regression model achieved an AUC of 0.7141, providing a strong linear baseline. However, its lower performance compared with tree-based ensemble models suggests that the relationship between engineered behavioral features and repurchase outcomes contains nonlinear patterns that cannot be fully captured through additive linear assumptions. Similarly, the decision tree model produced the lowest AUC (0.6908), indicating that a single tree structure was insufficient to represent the complex feature interactions contained in the dataset.

The MLP neural network achieved an AUC of 0.7175, comparable to several ensemble models, suggesting that nonlinear representation learning can capture meaningful behavioral patterns even with a relatively simple architecture. However, its performance remained slightly below gradient-boosting approaches, possibly because tree-based models are particularly effective for structured tabular data containing heterogeneous behavioral, categorical, and aggregated features.

The ROC curves of all seven classifiers are presented in Figure 9. Consistent with the AUC comparison, LightGBM, CatBoost, and XGBoost showed the strongest overall discrimination, whereas the decision tree exhibited comparatively weaker ranking capability.

Overall, the seven models rank as LightGBM > CatBoost > XGBoost > MLP neural network > random forest > logistic regression > decision tree, with the top four models (LightGBM, CatBoost, XGBoost, and the MLP neural network) all within an AUC range of 0.005-indicating comparable discriminative power-while the decision tree trails the remaining models by a substantially larger margin, making it the weakest performer among the seven. This result indicates that, on the feature system constructed in this study, gradient-boosting ensemble models perform best overall, while a neural network capable of non-linear transformation demonstrates comparable competitiveness.

From an e-commerce operations perspective, the narrow AUC margin separating LightGBM, CatBoost, XGBoost, and the MLP carries a direct practical implication for algorithm selection: because no single classifier offers a decisive accuracy advantage over its closest competitors, platforms need not treat model choice as the primary lever for improving repurchase prediction performance. Instead, the choice among these closely matched models can reasonably be guided by secondary, deployment-relevant criteria-

training, and inference speed and compatibility with interpretability tools such as SHAP rather than by marginal gains in predictive accuracy alone.

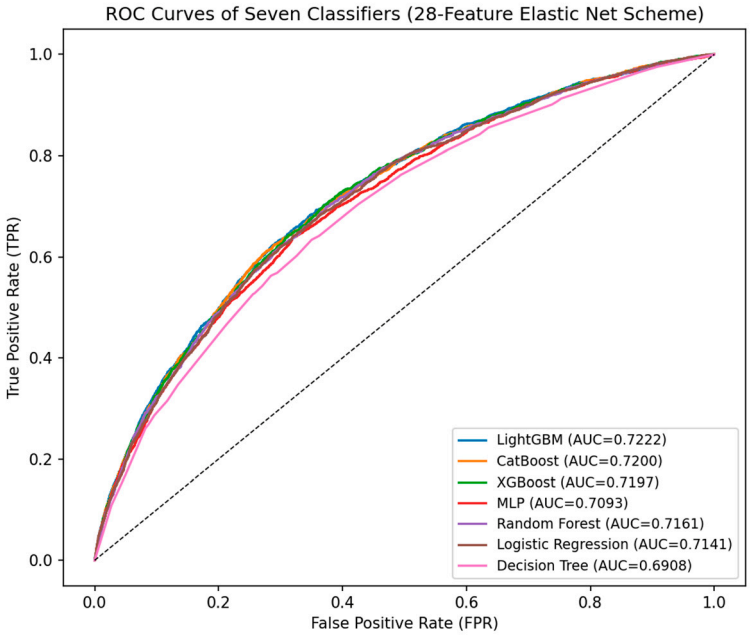


Figure 9. ROC curves of all seven classifiers on the test set (28-feature Elastic Net scheme). The diagonal dashed line denotes the ROC curve for random guessing as a baseline reference.

4.4. Model Interpretability Analysis

To further explain the mechanism underlying model predictions, and to translate feature importance into quantifiable, business-interpretable marginal effects, a SHAP (SHapley Additive exPlanations) analysis was conducted on the best-performing model (LightGBM under the 28-feature Elastic Net scheme). SHAP values were computed using TreeExplainer on a random sample of 5000 test-set observations.

Figure 10 shows that the SHAP value range for merchant_repurchase_rate (−1.5 to +1.6) far exceeds that of any other feature, with a clear separation between high values (red, driving predictions upward) and low values (blue, driving predictions downward). This corroborates, at the level of the prediction mechanism rather than aggregate ranking alone, the dominant role of this feature already established independently by all six feature-selection criteria in Section 4.2. The average absolute SHAP values in Figure 10 further quantify this: merchant_repurchase_rate (0.394) exceeds the second-ranked feature, time_stamp_nunique (0.175), by more than a factor of two, and exceeds the fifteenth-ranked feature, ratio_brand_most1 (0.022), by a factor of roughly 18-indicating that its contribution to model predictions is not merely the largest among the 15 features shown, but disproportionately so. time_stamp_nunique, seller_concentration, and activity_density form a second tier of contribution (mean |SHAP| of 0.175, 0.141, and 0.085, respectively), consistent with their rankings under gain importance.

To quantify the marginal effect of the leading predictors, SHAP dependence plots were constructed for merchant_repurchase_rate, activity_density, and time_stamp_nunique (Figure 11). merchant_repurchase_rate exhibits an almost monotonic, near-linear positive relationship with its SHAP contribution: as the merchant's historical repurchase rate rises from near 0 to approximately 0.25, its contribution to the predicted log-odds rises correspondingly from about −1.5 to +1.6, indicating that this single feature can, on its own, shift the predicted outcome across nearly the model's entire dynamic range. activity_density and time_stamp_nunique instead display a pattern of rapid initial increase followed by a

plateau, indicating diminishing marginal returns once a customer's activity level exceeds a moderate threshold.

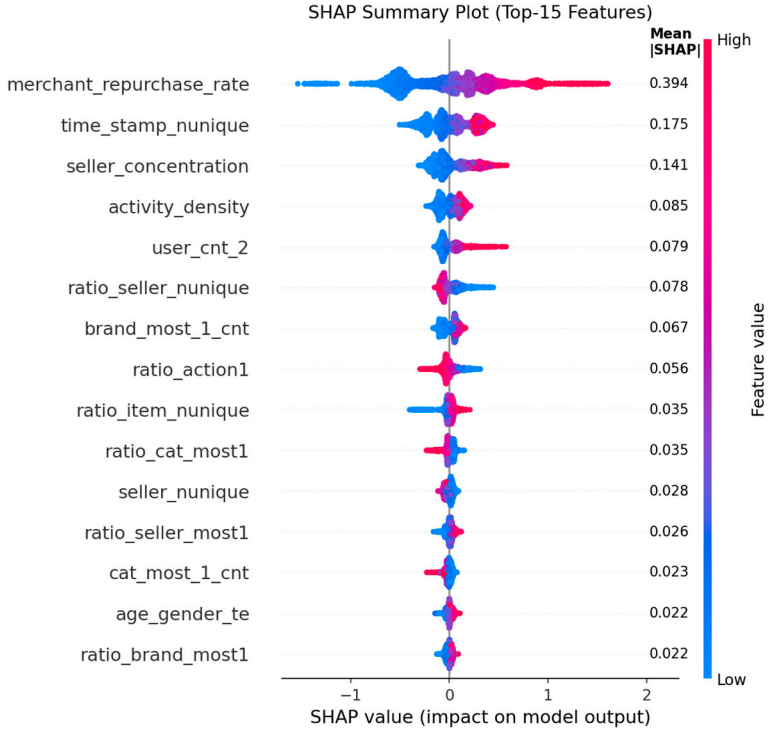


Figure 10. SHAP summary plot showing the distribution and direction of each feature's contribution to predicted repurchase probability (top-15 features).

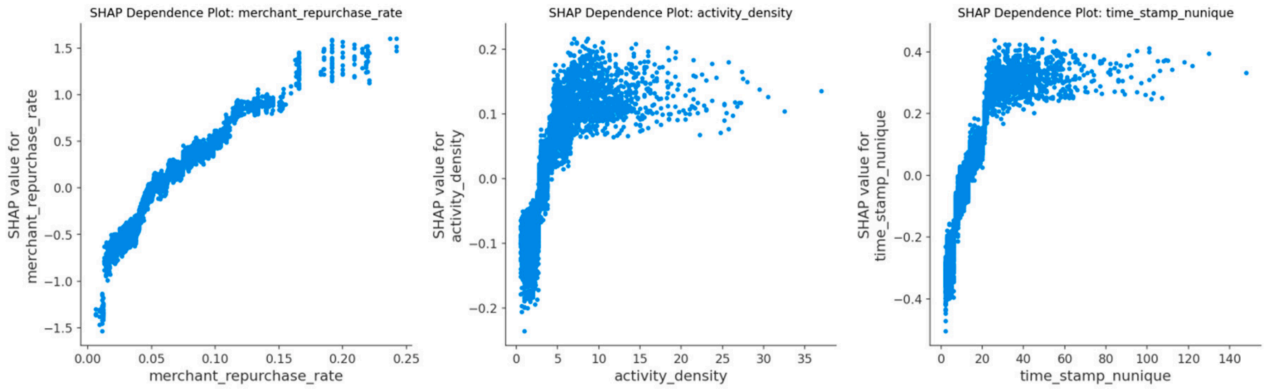


Figure 11. SHAP dependence plots for merchant_repurchase_rate, activity_density, and time_stamp_nunique.

To illustrate these mechanisms concretely, two representative test-set observations were selected: Case A, with the highest predicted repurchase probability in the test set (p = 0.872), and Case B, with the lowest (p = 0.060). Figure 12 decomposes each prediction into individual feature contributions via SHAP waterfall plots. In Case A, merchant_repurchase_rate alone contributes +1.11 to the predicted log-odds, exceeding the combined contribution of the next five features; in Case B, the same feature contributes −1.60, again the single largest contribution among all features considered. The two customers differ across numerous behavioral attributes, yet the dominant driver of the difference in their predicted outcomes is, in both cases, the historical repurchase rate of the merchant with whom they interacted-providing direct, case-level evidence that repurchase

prediction in this setting is substantially merchant-driven rather than purely a function of individual customer behavior.

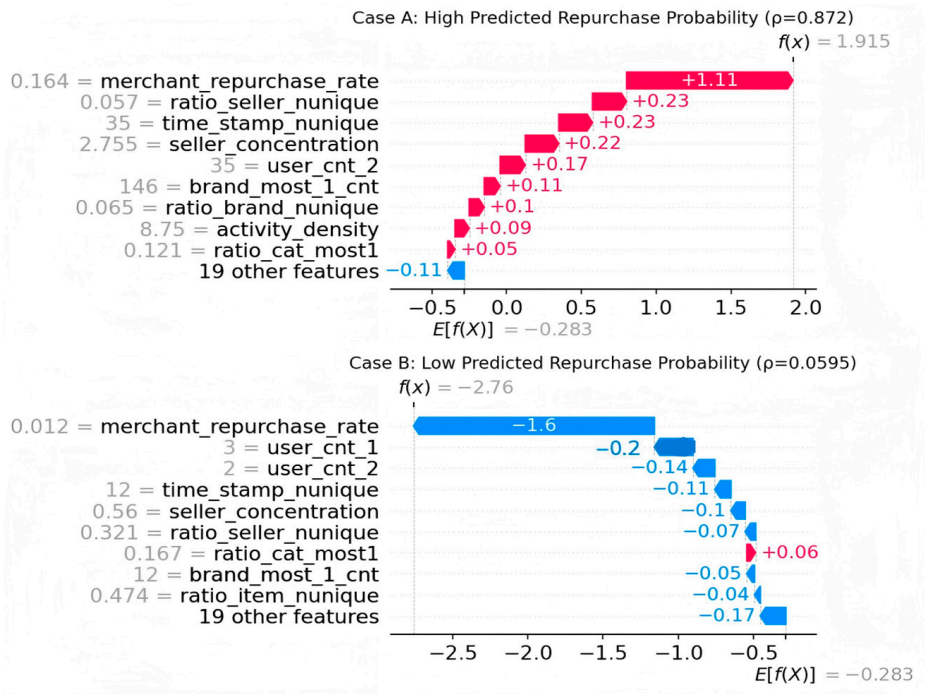


Figure 12. SHAP waterfall decomposition for a high-probability case (**Case A**) and a low-probability case (**Case B**). Red bars represent positive feature contributions that increase the predicted repurchase probability, and blue bars denote negative contributions that decrease the predicted probability.

Taken together, the SHAP analysis provides mechanism-level, rather than purely rank-based, evidence for the central empirical finding of this study: that merchant-level historical repurchase performance is the dominant driver of predicted repurchase behavior, with a quantifiable, near-monotonic marginal effect that can be communicated to non-technical stakeholders in concrete terms.

5. Discussion

This study set out to examine how feature representation and model selection jointly shape customer repurchase prediction in e-commerce, rather than treating predictive accuracy improvements as a matter of algorithm sophistication alone. By evaluating behavioral feature construction, feature-selection robustness, and classifier choice within a single framework, the findings speak to both the theoretical determinants of repurchase behavior and the practical design of data-driven retention strategies.

5.1. Theoretical Implications: The Role of Merchant-Level Characteristics in Repurchase Behavior

A central finding of this study is that merchant-level historical characteristics carry substantial predictive weight for customer repurchase behavior—evidence that repeat purchasing is not determined solely by individual customer traits, but also reflects the quality of the customer–merchant relationship built up through prior interactions. This directly confirms H1: the empirical dominance of merchant-level characteristics over individual customer-level behavioral variables is consistent with the theoretical expectation derived from customer loyalty theory.

Traditional loyalty theory places trust, perceived value, service consistency, and accumulated exchange experience at the center of long-term customer relationships [53–55]. These relational factors are difficult to measure directly in e-commerce, where customer

experience is entangled with products, merchants, and platform dynamics. This study offers a potential means of addressing this challenge: merchant-level historical repurchase indicators can serve as a quantitative proxy for accumulated relationship quality and a merchant's retention capability.

This complements prior e-commerce prediction work, which has focused mainly on customer-level signals—purchase frequency, browsing activity, transaction history [56,57]. Such signals remain important, but the present results show that repurchase decisions are also shaped by the broader merchant environment in which transactions occur. Merchant-level characteristics likely capture latent aspects of experience—product reliability, service consistency, fulfillment performance—that customer-level variables alone do not represent.

The same logic extends to customer lifetime value (CLV). Existing CLV frameworks center on customer profitability, purchase frequency, and individual-level relationship development [24]; the present results suggest that future customer value also depends on the retention capability of the merchants with whom a customer interacts. Retention analytics may therefore benefit from combining customer-side behavioral information with merchant-side relational characteristics, rather than treating each customer as an independent decision-making unit.

5.2. Implications for Feature Representation and Machine Learning-Based E-Commerce Analytics

A second finding is that feature representation matters more than algorithm selection for repurchase-prediction performance. Gradient-boosting models were competitive, but once informative behavioral representations had been constructed, the performance gap among the strongest classifiers narrowed considerably—consistent with the broader machine learning observation that predictive performance depends as much on representation quality as on model complexity [13].

E-commerce activity records are generated from heterogeneous interactions among customers, products, brands, and merchants, and raw transaction logs rarely capture the underlying behavioral mechanisms behind future repurchase on their own. Transforming these records into meaningful relational features—engagement intensity, behavioral diversity, temporal activity patterns, merchant-level historical performance—is what makes large-scale behavioral data predictive in the first place.

Well-constructed relational features can deliver greater predictive value than adopting a more sophisticated algorithm [58]. Once a sufficiently informative representation is in place, further gains from algorithmic refinement alone are modest. Future e-commerce analytics research may therefore gain more by investing in feature construction and behavioral interpretation than by pursuing increasingly complex model architectures.

Predictive performance in this setting should also not be judged by AUC alone. In a severely imbalanced task like repurchase prediction, threshold-dependent metrics reveal deployment challenges that AUC does not. LightGBM achieved the highest AUC in this study, yet its precision at the default threshold remained low (0.1160)—a direct consequence of picking out a small minority of repeat purchasers from a much larger pool of non-repeat purchasers. The practical role of these models is therefore best understood as ranking and prioritization, with the decision threshold set according to business objectives and the relative cost of false positives versus false negatives [59].

5.3. Managerial Implications for E-Commerce Customer Retention

These findings carry several practical implications for platforms, merchants, and customer relationship management teams. Repurchase prediction need not be treated purely as a classification problem: it can also inform broader retention management by flagging both high-potential customers and the operational factors behind repeat purchasing.

The strong predictive contribution of merchant-level historical performance points to a gap in current retention practice, which tends to focus on identifying individual high-probability customers and targeting them with promotions. Improving retention, the results suggest, requires attention not only to who the customer is, but to the merchant and service environment from which they purchase. Platform operators could operationalize this by setting a merchant-level repurchase-rate threshold—for example, flagging any merchant whose repurchase rate falls more than one standard deviation below the category-specific average for two consecutive monthly reporting periods—to automatically trigger an account-management review rather than leaving underperformance to be noticed informally. Once flagged, a merchant could be required to submit a corrective action plan against specific, trackable metrics such as return rate, average delivery lead time, and customer-service response time, with continued eligibility for promotional co-funding or preferential placement made conditional on measurable improvement in these metrics over a defined review window (e.g., 60–90 days), rather than relying solely on customer-facing discounts, which address the symptom of low repurchase rather than its operational cause.

The economic cost of false positives also deserves attention. Given the low precision observed at the default threshold, a meaningful share of customers flagged as likely repeat purchasers will not, in fact, repurchase; treating these predictions as automatic targeting rules risks wasted promotional spend and customer fatigue from unwanted contact. The more defensible use of the model is customer prioritization rather than binary classification: customers can be ranked by predicted repurchase probability and grouped into deciles, with the top decile allocated the highest-cost interventions (e.g., personalized product recommendations, early access to new inventory, or loyalty-tier upgrades), the middle deciles allocated lower-cost, semi-automated engagement (e.g., templated re-engagement emails or app notifications timed to the customer's typical purchase cycle), and the bottom deciles limited retention investment, since the marginal cost of contacting them is unlikely to be recovered given their low conversion likelihood.

Retention strategy should also be adapted to purchasing context. Because the data analyzed in this study were collected during a major promotional period, observed repurchase behavior likely reflects both underlying merchant–customer relationships and campaign-driven purchasing that would not otherwise have occurred. During promotions, the model's customer-ranking output could be used to restrict discount eligibility to customers in the middle deciles of predicted repurchase probability—those plausibly responsive to a price incentive but not likely to repurchase regardless—while withholding discounts from the top decile, where repurchase is already highly likely without a price incentive, and from the bottom deciles, where a discount is unlikely to convert a one-off purchase into a repeat one. Outside promotional periods, when campaign-driven noise is absent, merchant-level repurchase indicators offer a cleaner signal of relationship quality, and the same account-management review process described above (triggered by sustained merchant-level underperformance rather than by a single low-conversion campaign) is likely to be a more productive use of retention resources than price-based incentives.

Finally, platform operators and individual merchants face different levers, and the specific processes above can be assigned accordingly. Platforms are positioned to operate the merchant-level flagging and review mechanism at the ecosystem scale, since it depends on cross-merchant benchmarking (i.e., comparing a given merchant's repurchase rate against its category average) that individual merchants cannot compute on their own. Individual merchants, by contrast, are better positioned to act on the resulting corrective action plan directly—for example, auditing specific return-rate drivers by product line, renegotiating delivery-carrier service-level agreements where lead time is the flagged issue, or adding staffing to customer-service channels where response time is the flagged

issue—and to apply the same decile-based customer ranking internally to prioritize their own limited relationship-management resources (e.g., personal outreach from a dedicated account manager) toward their top-decile customers rather than distributing outreach effort uniformly across their full customer base.

Taken together, effective e-commerce retention management is best understood not as customer-only targeting, but as an integrated approach spanning customer behavior, merchant capability, and purchasing context, implemented through a concrete decile-based prioritization rule for customers and a threshold-triggered review process for merchants. Machine learning models therefore function not merely as predictive tools, but as decision support for more targeted and resource-efficient retention strategies.

5.4. Limitations and Future Research

These findings should be interpreted in light of two limitations.

First, the dataset comes from a single platform (Tmall) and a specific definition of the repeat-purchase observation window; although the data are publicly available via Alibaba Cloud's Tianchi platform and widely used as a benchmark in this literature, generalizing to other platforms, product categories, or window definitions will require further validation.

Second, the fifty-replication stability analysis tests sensitivity to random data partitioning, but not to temporal distribution shift. Consumer preferences, purchasing patterns, competitive dynamics, and platform strategy can all evolve over time in ways that affect predictive performance and feature importance alike. Temporally held-out validation or longitudinal data would be a natural next step for assessing how stable these repurchase-prediction results are under changing market conditions.

6. Conclusions

This study conducted a systematic empirical investigation of customer repurchase prediction in e-commerce by examining the relative contributions of feature representation, feature selection, and machine learning model choice within an integrated predictive framework. Using a large-scale Tmall behavioral dataset, the study addressed two research questions: RQ1, concerning the most informative predictors of repeat purchase behavior and the viability of a reduced feature subset; and RQ2, concerning the predictive capability of different machine learning classifiers under severe class imbalance and the managerial interpretability of their predictions.

The empirical results demonstrate that merchant-level historical characteristics provide substantial predictive information for customer repurchase behavior. Across six independent feature-selection approaches, merchant historical repurchase rate consistently emerged as one of the most influential predictors, and its importance was further supported by model interpretation analysis. This finding extends existing repurchase prediction research, which has predominantly emphasized customer-level behavioral signals, by demonstrating that repeat purchasing is not solely determined by individual customer activity but is also shaped by the retention capability and accumulated relationship quality of the merchant. Therefore, this study contributes a relational perspective to e-commerce customer retention analytics by incorporating merchant-level characteristics into predictive modeling frameworks. This finding provides empirical support for H1, derived from customer loyalty theory, that merchant-level historical characteristics carry stronger predictive information than individual customer behavioral characteristics. To the authors' knowledge, this study represents one of the few attempts in the repurchase-prediction literature to integrate theoretically motivated relational constructs with machine-learning-based feature analysis, rather than treating predictive modeling and marketing theory as separate analytical perspectives. By linking feature importance patterns with customer–merchant relationship theory, this study

provides a theoretically informed interpretation of machine learning results and advances understanding of the drivers underlying e-commerce repurchase behavior.

From a methodological perspective, this study contributes a robustness-oriented feature-selection and model evaluation framework for repurchase prediction. Unlike approaches that compare algorithms based on a single data partition, this study evaluates feature-selection stability across repeated simulations and examines whether predictive gains remain consistent across alternative feature configurations. The results show that the Elastic Net-selected 28-feature subset achieved predictive performance comparable to the full 37-feature representation, while LightGBM achieved the strongest single-model performance (AUC = 0.7222) among the evaluated classifiers. These findings suggest that once informative behavioral representations have been constructed, improvements from additional algorithmic complexity become relatively limited, highlighting the importance of feature engineering in e-commerce predictive analytics.

The practical implications of this study are twofold. First, e-commerce platforms and merchants should prioritize the construction and monitoring of merchant-level retention indicators rather than focusing exclusively on customer-level targeting. Second, predictive models should be viewed primarily as customer prioritization tools rather than deterministic classification systems, allowing firms to adjust decision thresholds according to marketing objectives and operational costs. In particular, retention strategies may benefit from distinguishing promotional and routine purchasing contexts, thereby enabling more efficient allocation of customer relationship management resources.

Despite these contributions, further research is needed to examine the generalizability of the proposed framework across different platforms, product categories, and temporal contexts. Future studies may incorporate temporally held-out validation, cross-platform datasets, and more advanced ensemble strategies to further investigate the stability of merchant-level predictive signals and their role in long-term customer retention management.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jtaer21100347/s1, Data S1: The processed dataset comprising the 37 engineered features and the binary repurchase-outcome label used throughout this study (train, validation, and test partitions, consistent with the 70%/15%/15% split described in Section 3.1).

Author Contributions: X.Y.: conceptualization, methodology, formal analysis, data curation, writing—original draft. P.B.: conceptualization, supervision, validation, writing—review and editing. N.C.: Supervision, validation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by Mahasarakham University. No external grant number is available for this internal university research fund.

Institutional Review Board Statement: Not applicable. This study did not involve human participants or animals; it relies exclusively on secondary, de-identified transactional log data from a publicly released e-commerce benchmark dataset.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset analyzed in this study is the publicly available Tmall repeat-buyer prediction dataset, released in conjunction with the IJCAI-15/KDD Cup repeat-buyer prediction challenge. Restrictions may apply to the availability of the processed/engineered feature files, which were derived by the authors and can be made available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank Nader Tajvidi and Yingying Song for their valuable guidance and suggestions during the course of this study.

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

Abbreviations

The following abbreviations are used in this manuscript:

MI	Mutual Information
LightGBM	Light Gradient Boosting Machine
Lasso	Least Absolute Shrinkage and Selection Operator
SMOTE	Synthetic Minority Oversampling Technique
AUC	Area Under the Receiver Operating Characteristic Curve
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence
CLV	Customer Lifetime Value
LSTM	Long Short-Term Memory
VIF	Variance Inflation Factor
MLP	Multilayer Perceptron
TP/TN/FP/FN	True/False Positive/Negative
C2C	Consumer-to-Consumer

References

1. Liu, M.; Jia, W.; Yan, W. Factors Influencing Consumers' Repurchase Behavior on Fresh Food e-Commerce Platforms: An Empirical Study. *Adv. Eng. Inform.* **2023**, *56*, 101936. [CrossRef]
2. Ge, R.; de Haan, E.; Verhoef, P.C. Location Still Matters: How Geographic Proximity between Customers Shapes Repeat Purchase Behavior in Online Food Delivery Services. *J. Retail.* **2025**, *102*, 315–332. [CrossRef]
3. Xie, K.; Kwok, L.; Wu, J. Are Consumers Loyal to Home-Sharing Services? *Int. J. Contemp. Hosp. Manag.* **2019**, *31*, 1066–1085. [CrossRef]
4. Thangeda, R.; Kumar, N.; Majhi, R. A Neural Network-Based Predictive Decision Model for Customer Retention in the Telecommunication Sector. *Technol. Forecast. Soc. Change* **2024**, *202*, 123250. [CrossRef]
5. Lamrhari, S.; El Ghazi, H.; Oubrich, M. A Social CRM Analytic Framework for Improving Customer Retention, Acquisition, and Conversion. *Technol. Forecast. Soc. Change* **2021**, *174*, 121275. [CrossRef]
6. Owusu, P.K.; Li, Z.; Mensah, I.A. Consumer Response to E-Commerce Service Failure: Leveraging Repurchase Intentions through Strategic Recovery Policies. *J. Retail. Consum. Serv.* **2024**, *82*, 104137. [CrossRef]
7. Kumar, V.; Ogunmola, G.A. E-Retail Factors for Customer Activation and Retention: An Empirical Study from Indian e-Commerce Customers. *J. Retail. Consum. Serv.* **2020**, *59*, 102399. [CrossRef]
8. Chaudhuri, N.; Gupta, G.; Vamsi, V. On the Platform but Will They Buy? Predicting Customers' Purchase Behavior Using Deep Learning. *Decis. Support Syst.* **2021**, *149*, 113622. [CrossRef]
9. Kirk, A. Binary Logistic Regression. In *Political Analysis: A Guide to Data and Statistics*; SAGE Publications: London, UK, 2023.
10. Yang, Y.; Lv, H.; Chen, N. A Survey on Ensemble Learning under the Era of Deep Learning. *Artif. Intell. Rev.* **2022**, *56*, 5545–5589. [CrossRef]
11. Ganaie, M.A.; Hu, M.; Malik, A.K. Ensemble Deep Learning: A Review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
12. Dong, X.; Yu, Z.; Cao, W. A Survey on Ensemble Learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
13. Zhang, W.; Wang, M. An Improved Deep Forest Model for Prediction of E-Commerce Consumers' Repurchase Behavior. *PLoS ONE* **2021**, *16*, e0255906. [CrossRef] [PubMed]
14. Zhang, J.; Song, H.; Zhou, B. SAR Target Classification Based on Deep Forest Model. *Remote Sens.* **2020**, *12*, 128. [CrossRef]
15. Zhu, C.; Wang, X. Prediction of Consumer Repurchase Behavior Based on LSTM Neural Network Model. *Int. J. Syst. Assur. Eng. Manag* **2022**, *13*, 1042–1053. [CrossRef]
16. Lee, G.K.K.; Kasim, H.; Sirigina, R.P. Smart Robust Feature Selection (SoFt) for Imbalanced and Heterogeneous Data. *Knowl.-Based Syst.* **2021**, *236*, 107197. [CrossRef]
17. Chen, J.; Yang, C.; Zhu, H. A Novel Variable Selection Method Based on Stability and Variable Permutation for Multivariate Calibration. *Chemom. Intell. Lab. Syst.* **2018**, *182*, 188–201. [CrossRef]
18. Bertolini, R.; Finch, S.J.; Nehm, R.H. Enhancing Data Pipelines for Forecasting Student Performance: Integrating Feature Selection with Cross-Validation. *Int. J. Educ. Technol. High. Educ.* **2021**, *18*, 44. [CrossRef] [PubMed]

19. Alaoui, E.A.A.; Sahlaoui, H.; Sallah, A. DE-SHAP: A Meta-Optimization Framework Leveraging Differential Evolution for Efficient and Scalable Kernel SHAP Explanations. *Inf. Sci.* **2026**, *739*, 123128. [CrossRef]
20. Cheng, X.; Wang, T.; Zhu, D. Uncertainty Explanation of Artificial Intelligence Models by SHAP. *Knowl.-Based Syst.* **2026**, *337*, 115437. [CrossRef]
21. Cossío-Silva, F.J.; Revilla-Camacho, M.Á.; Vega-Vázquez, M. Value Co-Creation and Customer Loyalty. *J. Bus. Res.* **2015**, *69*, 1621–1625. [CrossRef]
22. Blut, M.; Chaney, D.; Lunardo, R. Customer Perceived Value: A Comprehensive Meta-Analysis. *J. Serv. Res.* **2023**, *27*, 501–524. [CrossRef]
23. Sullivan, Y.; Kim, D.J. Assessing the Effects of Consumers' Product Evaluations and Trust on Repurchase Intention in e-Commerce Environments. *Int. J. Inf. Manag.* **2017**, *39*, 199–219. [CrossRef]
24. Valentini, T.; Roederer, C.; Castéran, H. From Redesign to Revenue: Measuring the Effects of Servicescape Remodeling on Customer Lifetime Value. *J. Retail. Consum. Serv.* **2023**, *77*, 103681. [CrossRef]
25. Artelt, A.; Gregoriades, A. Supporting Organizational Decisions on How to Improve Customer Repurchase Using Multi-Instance Counterfactual Explanations. *Decis. Support Syst.* **2024**, *182*, 114249. [CrossRef]
26. Liu, G.; Nguyen, T.T.; Zhao, G.; Zha, W.; Yang, J.; Cao, J.; Wu, M.; Zhao, P.; Chen, W. Repeat Buyer Prediction for E-Commerce. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Volume KDD'16.
27. Tian, Y.; Ye, Z.; Yan, Y.; Sun, M. A Practical Model to Predict the Repeat Purchasing Pattern of Consumers in the C2C E-Commerce. *Electron. Commer. Res.* **2015**, *15*, 571–583. [CrossRef]
28. Chou, P.; Chuang, H.H.C.; Chou, Y.C.; Lin, T.Y. Predictive Analytics for Customer Repurchase: Interdisciplinary Integration of Buy till You Die Modeling and Machine Learning. *Eur. J. Oper. Res.* **2021**, *296*, 635–651. [CrossRef]
29. Yang, L.; Niu, X.; Wu, J. RF-LightGBM: A probabilistic ensemble way to predict customer repurchase behaviour in community e-commerce. *arXiv* **2021**, arXiv:2109.00724.
30. Zhang, M.; Lu, J.; Ma, N.; Cheng, T.E.; Hua, G. A Feature Engineering and Ensemble Learning Based Approach for Repeated Buyers Prediction. *Comput. Commun. Control* **2022**, *17*, 4988. [CrossRef]
31. Baesens, B.; Viaene, S.; Poel, D.; Vanthienen, J.; Dedene, G. Bayesian Neural Network Learning for Repeat Purchase Modelling in Direct Marketing. *Eur. J. Oper. Res.* **2002**, *138*, 191–211. [CrossRef]
32. Meade, N.; Islam, T. Using Copulas to Model Repeat Purchase Behaviour—An Exploratory Analysis via a Case Study. *Eur. J. Oper. Res.* **2010**, *200*, 908–917. [CrossRef]
33. Kuric, E.; Pilarikova, A.; Demcak, P.; Mensatorisova, D. Effect of Low-Level Interaction Data in Repeat Purchase Prediction Task. *Int. J. Hum.-Comput. Interact.* **2024**, *40*, 2515–2533. [CrossRef]
34. Kuo, R.J.; Cheng, H.R. A Content-Based Recommender System with Consideration of Repeat Purchase Behavior. *Appl. Soft Comput* **2022**, *127*, 109361. [CrossRef]
35. Al Rahib, M.A.S.N.; Mia, R.; Sattar, A. Customer Data Prediction and Analysis in E-Commerce Using Machine Learning. *Bull. Electr. Eng. Inf.* **2024**, *13*, 2624–2633. [CrossRef]
36. Li, S. Machine Learning-Based Prediction Mechanism of Repeated Purchase Behavior of e-Commerce Customers. In Proceedings of the 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), New Taipei City, China, 19–21 April 2024.
37. Knobbe, A.J.; Haas, M.; Siebes, A. Propositionalisation and Aggregates. In Proceedings of the Principles of Data Mining and Knowledge Discovery (PKDD 2001), Freiburg, Germany, 3–5 September 2001.
38. Kanter, J.M.; Veeramachaneni, K. Deep Feature Synthesis: Towards Automating Data Science Endeavors. In Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, 19–21 October 2015.
39. Katz, G.; Shin, E.C.R.; Song, D. ExploreKit: Automatic Feature Generation and Selection. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016.
40. Elkan, C. The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, WA, USA, 4–10 August 2001.
41. Zadrozny, B.; Langford, J.; Abe, N. Cost-Sensitive Learning by Cost-Proportionate Example Weighting. In Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003), Melbourne, FL, USA, 19–22 November 2003.
42. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
43. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
44. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
45. Cox, D.R. The Regression Analysis of Binary Sequences. *J. R. Stat. Soc. Ser. B Methodol.* **1958**, *20*, 215–232. [CrossRef]
46. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]

47. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Francisco, CA, USA, 1993.
48. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth: Belmont, CA, USA, 1984.
49. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
50. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Volume KDD'16.
51. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2017.
52. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A.C. Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*; Neural Information Processing Systems Foundation: La Jolla, CA, USA, 2018.
53. Chiou, J.S. Service Quality, Trust, Specific Asset Investment, and Expertise: Direct and Indirect Effects in a Satisfaction-Loyalty Framework. *J. Acad. Mark. Sci.* **2006**, *34*, 613–627. [CrossRef]
54. Ngobo, P.V. The Trajectory of Customer Loyalty: An Empirical Test of Dick and Basu's Loyalty Framework. *J. Acad. Mark. Sci.* **2016**, *45*, 229–250. [CrossRef]
55. Yuen, K.F.; Koh, L.Y.; Wong, Y.Q. Sustainable Crowdsourced Delivery: A Study of Technological, Health, Value, and Trust Antecedents of Consumer Loyalty. *J. Clean. Prod.* **2023**, *405*, 137010. [CrossRef]
56. Tamaddoni, A.; Stakhovych, S.; Ewing, M.T. Comparing Churn Prediction Techniques and Assessing Their Performance. *J. Serv. Res.* **2015**, *19*, 123–141. [CrossRef]
57. Kim, S.; Shin, W.; Kim, H.W. Predicting Online Customer Purchase: The Integration of Customer Characteristics and Browsing Patterns. *Decis. Support Syst.* **2023**, *177*, 114105. [CrossRef]
58. Zhang, L. Personalized and Contextualized Data Analysis for E-Commerce Customer Retention Improvement with Bi-LSTM Churn Prediction. *IEEE Trans. Consum. Electron.* **2024**, *71*, 4406–4414. [CrossRef]
59. Chen, J.J.; Tsai, C.A.; Moon, H. Decision Threshold Adjustment in Class Prediction. *SAR QSAR Environ. Res.* **2006**, *17*, 337–352. [CrossRef] [PubMed]

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