Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU38028

Core Feature Screening for Imbalanced Online Purchase Intention Prediction

Kexin Gao

School of Data Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China
125090143@link.cuhk.edu.cn

Abstract. E-commerce browsing data suffers severe class imbalance, making accuracy unreliable for purchase-intention prediction. Using the Online Shoppers Purchasing Intention dataset from the University of California, Irvine (UCI) Machine Learning Repository, this paper compares Logistic Regression, K-Nearest Neighbors, Random Forest and Gradient Boosting via stratified 5-fold cross-validation without synthetic oversampling. Three feature settings are evaluated: full-feature set, Random-Forest-derived top-5 feature subset, and single PageValues. The macro-averaged F1 score (macro-F1) and the area under the precision-recall curve (PR-AUC) serve as core metrics. Class weight and sample weight adjustment are applied to handle imbalance during training. Gradient Boosting obtains the best overall performance. Under imbalance, accuracy is misleading for model comparison. PageValues exhibits strong standalone predictive power, and lightweight models using five core features retain nearly full predictive performance. The results provide practical references for feature screening and lightweight model deployment in imbalanced e-commerce purchase prediction, and demonstrate that a small number of high-signal behavioral indicators is sufficient for effective purchase-intention prediction.

Keywords: e-commerce purchase prediction, stratified 5-fold cross-validation, gradient boosting, feature importance, imbalanced classification

1. Introduction

The rapid growth of e-commerce generates abundant user browsing data, which can be used to predict visitor purchasing intention for precise marketing and layout optimization. Accurately identifying sessions with purchase intention enables timely recommendation, targeted promotion and conversion rate improvement, and has become a core analytical task for online retailers. Real e-commerce datasets generally face severe class imbalance, where purchasing samples account for a tiny fraction of all records. Simple accuracy cannot reflect a model's capacity to identify potential buyers, so targeted evaluation metrics and imbalance processing strategies are required.

Few existing works conduct controlled experiments to verify standalone predictive power of core indicators or build lightweight models on screened features [1]. This study uses the Online Shoppers Purchasing Intention dataset from the University of California, Irvine (UCI) Machine Learning Repository and compares four classification models with stratified 5-fold cross-validation. Class-weight and sample-weight adjustment handle imbalance instead of synthetic oversampling. Three

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU38028

feature schemes are evaluated: full-feature set, top-5 feature subset, and single PageValues. The macro-averaged F1 score (macro-F1) and the area under the precision-recall curve (PR-AUC) are primary metrics; accuracy is only auxiliary.

Results demonstrate Gradient Boosting achieves the best overall performance, and PageValues has strong standalone predictive capability. Lightweight models with five core features preserve nearly full prediction performance. This paper is structured as follows. Section 2 reviews related work on purchase intention prediction and feature engineering. Section 3 introduces dataset, preprocessing, models and evaluation protocols. Section 4 presents descriptive statistics and three groups of comparative experiments. Section 5 concludes this work with limitations and future directions. The three feature-control experiments provide reproducible evidence for feature screening practice in imbalanced purchase prediction.

2. Literature review

Machine-learning-based purchase-intention prediction is widely studied for e-commerce session data [2]. In-session browsing behavior can predict purchase outcomes before a session terminates [3], and deep-learning methods boost performance on large-scale platform datasets [4]. Most existing works, however, employ ensemble tree or linear models on complete feature sets to forecast buying behavior, and they seldom distinguish the independent predictive capacity of individual session indicators. Nevertheless, most existing models are trained on full-feature sets and rarely quantify the independent predictive contribution of individual session features. Common imbalance-mitigation methods such as the Synthetic Minority Over-sampling Technique (SMOTE) may create artificial bias and distort real-world user-behavior patterns.

Feature engineering has been shown to strongly condition classification outcomes on this task [5]. Among candidate algorithms, gradient boosting variants frequently report the most competitive discrimination performance [6], and clickstream-based studies confirm that a small number of behavioral variables carries most of the predictive signal [7]. Although PageValues is widely included in feature pools, few scholars isolate this metric for single-variable validation to quantify its exclusive predictive signal [8]. In addition, most prior papers rely on accuracy as the primary judging standard, ignoring its inherent defects under imbalanced data distribution.

Addressing these gaps, this paper avoids synthetic oversampling and conducts three comparative experiments: full features, core-feature subset, and single-feature PageValues. Controlled comparisons isolate the contribution of each feature scheme under identical data splits and evaluation protocols. Using macro-F1 and PR-AUC as core metrics, this study tests whether a single high-signal indicator can approach the performance of the full feature set and explores lightweight-model feasibility, and provides reproducible evidence for feature screening practice in imbalanced purchase prediction.

3. Methodology

3.1. Data source

This study adopts the UCI Online Shoppers Purchasing Intention Dataset (ID: 468) [9], one-year real-world e-commerce session data under CC BY 4.0 license. It has 12,330 complete records with no missing values, so data imputation is unnecessary. The binary target Revenue labels purchase completion. Purchasing samples make up only 15.47%, forming a mildly imbalanced classification

DOI: 10.54254/2753-8818/2026.PU38028

task. A stratified 80%-20% train-test split (random_state=42) preserves class proportions to prevent distribution shift.

3.2. Variables used in this research

Five high-cardinality categorical columns (Month, OperatingSystems, Browser, Region, TrafficType) are removed to avoid heavy one-hot encoding and overfitting, leaving 12 predictive features plus target Revenue. All retained variables and their encodings are summarized in Table 1.

3.2.1. Supplementary notes for the key variable pagevalues

PageValues is a Google-Analytics-derived metric computed during browsing without label leakage, motivating the dedicated single-feature control experiment.

3.2.2. Visitor type distribution and grouping rules

The visitor group counts are as follows: Returning_Visitor: 10,551 records, New_Visitor: 1,694 records, and Other: only 85 records. The tiny "Other" subgroup is excluded from subgroup-level analysis, while VisitorType is encoded via one-hot encoding for modeling.

3.2.3. Data encoding and preprocessing pipeline rules

Binary Weekend and Revenue are encoded as 0/1. VisitorType uses one-hot encoding to avoid artificial ordinal bias. Standardization is applied only for Logistic Regression and K-Nearest Neighbors (KNN) within scikit-learn Pipeline [10]; scalers are fitted solely on training folds to prevent information leakage. Tree-based models require no standardization.

3.3. Four selected classification models

Four models are compared with corresponding imbalance-handling strategies:

Logistic Regression: baseline linear model with class_weight='balanced'.

KNN: distance-based lazy learner; imbalance addressed via standardization.

Random Forest: ensemble tree with class_weight='balanced', outputs feature importance [11].

Gradient Boosting: iterative boosting tree, applies sample weight for imbalanced labels [12].

SMOTE and all synthetic-oversampling approaches are excluded from all experiments.

3.4. Evaluation methodology

With merely 15.47% positive purchasing samples, a trivial all-negative classifier can achieve an accuracy over 84%, making accuracy unsuitable as the primary metric. Macro-F1 and PR-AUC serve as core imbalanced-classification metrics: macro-F1 equally averages F1-scores for both classes, and PR-AUC prioritizes performance on rare positive purchasing samples. ROC-AUC and accuracy are only auxiliary references.

Stratified 5-fold cross-validation preserves class ratios per fold. Imbalance mitigation depends only on class/sample weights and standardized pipelines; no synthetic resampling is used.

All cross-validation experiments adopt StratifiedKFold with n_splits=5, shuffle=True, random_state=42. Every evaluation metric within each fold is calculated from identical predicted labels and predicted probabilities to guarantee metric consistency.

DOI: 10.54254/2753-8818/2026.PU38028

Table 1. Description of all retained variables

Column Name	Meaning	Data Type	Supplementary Note
Administrative	Count of administrative pages browsed in one session	Numeric Feature	Numerical session behavior indicator
Administrative_Duration	Total stay time on administrative pages (seconds)	Numeric Feature	Numerical session behavior indicator
Informational	Count of informational pages browsed in one session	Numeric Feature	Numerical session behavior indicator
Informational_Duration	Total stay time on informational pages (seconds)	Numeric Feature	Numerical session behavior indicator
ProductRelated	Count of product pages browsed in one session	Numeric Feature	Numerical session behavior indicator
ProductRelated_Duration	Total stay time on product pages (seconds)	Numeric Feature	Numerical session behavior indicator
BounceRates	Bounce rate of the corresponding page group	Numeric Feature	Numerical session behavior indicator
ExitRates	Exit rate of the corresponding page group	Numeric Feature	Numerical session behavior indicator
PageValues	Page value score calculated by Google Analytics	Numeric Feature	Core indicator for follow-up single-feature test; strong predictive signal
SpecialDay	Weighted proximity to special holidays	Numeric Feature	Numerical session behavior indicator
Weekend	Whether the session occurs on weekend	Categorical	Binary encoded into 0 (weekday) and 1 (weekend)
VisitorType	Identity category of website visitor	Categorical	Three groups, transformed via one-hot encoding
Revenue	Whether the visitor generates a transaction (target label)	Target Label	Binary encoded into 0 (no purchase) and 1 (purchase)

4. Results

Stratified 5-fold cross-validation is adopted for all classification experiments to eliminate bias caused by imbalanced transaction labels, with macro-F1, PR-AUC and ROC-AUC selected as core

DOI: 10.54254/2753-8818/2026.PU38028

evaluation metrics and accuracy listed only for side comparison. Three groups of comparative experiments are conducted, namely the full feature set, the top-5 important feature subset and the single PageValues feature, to analyze model performance and feature predictive power comprehensively.

4.1. Descriptive statistical analysis

Basic exploratory data analysis is implemented to intuitively observe variable distribution, correlation and feature differences between purchasing and non-purchasing visitors. The imbalanced sample distribution of the target transaction label is presented in Figure 1. Figure 2 extends the correlation analysis to include the target variable Revenue, displaying all pairwise correlation coefficients of numeric indicators with readable numerical values. Four representative session behavior indicators are compared between two user groups via boxplots in Figure 3.

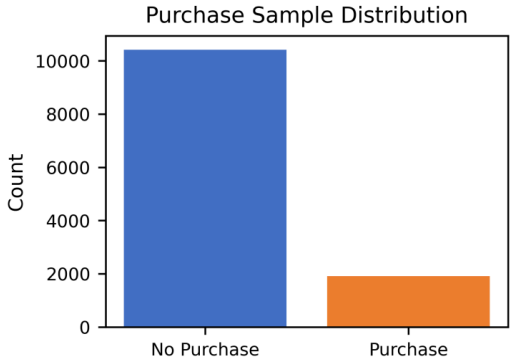

Figure 1. Bar chart of purchase sample distribution

(Data from: UCI Online Shoppers Purchasing Intention dataset [9])

This bar chart visualizes the quantity gap between purchasing and non-purchasing sessions. A severe class imbalance is clearly observed: non-purchasing visitors account for the vast majority of samples, while purchasing cases only occupy approximately 15.47% of the total dataset. This uneven sample distribution explains why overall accuracy cannot act as a reliable evaluation metric; a naive classifier that predicts all visitors as non-buyers can achieve misleadingly high accuracy yet fails to identify potential paying customers.

This 11×11 heatmap contains all pairwise correlation coefficients between the ten numeric behavior features and the target Revenue, with exact correlation values printed inside each grid cell. Two pairs of indicators show extremely strong positive correlation: ProductRelated and ProductRelated_Duration reach a correlation coefficient of 0.86, while BounceRates and ExitRates reach 0.91. Such high correlation means the two variables within each pair carry heavy overlapping information. By contrast, PageValues only maintains weak correlation with the other nine behavioral features, yet it has the highest positive correlation coefficient with the target Revenue among all input predictors.

Four critical session indicators including PageValues, ExitRates, BounceRates and ProductRelated_Duration are divided into purchasing and non-purchasing subgroups for visual comparison in a 2×2 subplot layout. For PageValues, the median value and the full range of the box for purchasing visitors are distinctly higher than those of non-purchasing visitors, and the box ranges of the two groups barely overlap. For ExitRates and BounceRates, non-purchasing visitors generally

DOI: 10.54254/2753-8818/2026.PU38028

have larger values, though significant overlap exists between the two subgroups. The gap of ProductRelated_Duration between purchasing and non-purchasing users is relatively moderate.

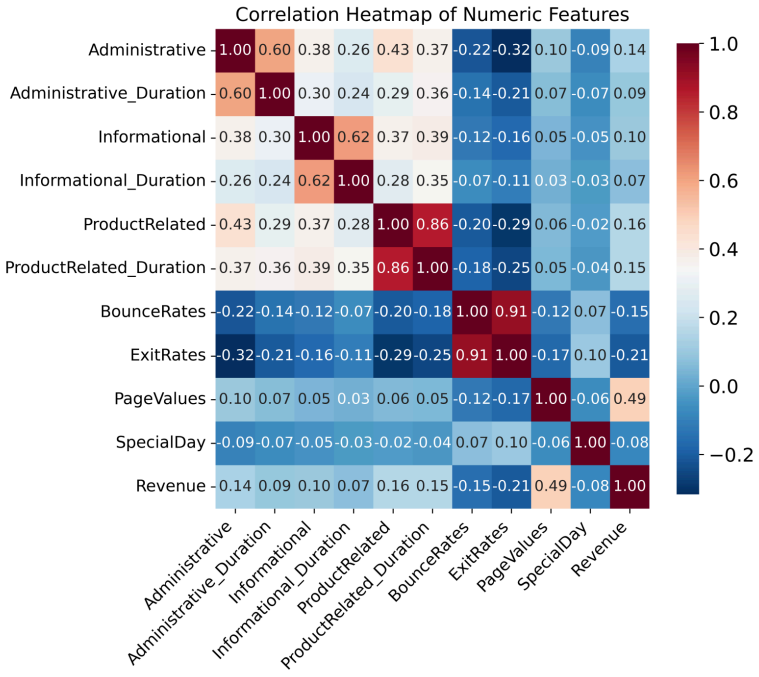

Figure 2. Correlation heatmap of ten numeric session features and the target Revenue

(Data from: UCI Online Shoppers Purchasing Intention dataset [9])

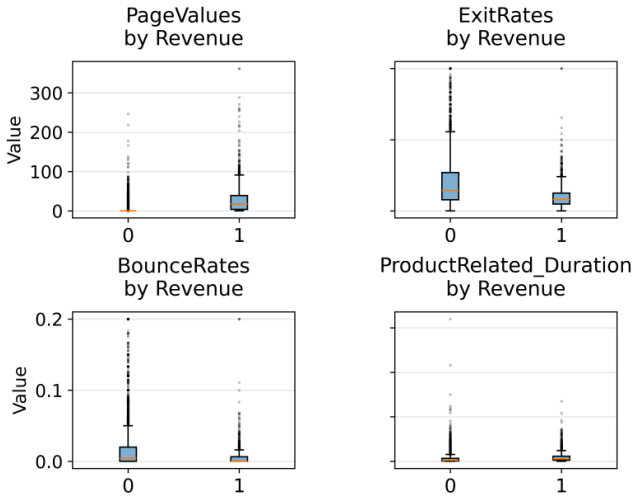

Figure 3. Boxplots of four core features grouped by purchase result

(Data from: UCI Online Shoppers Purchasing Intention dataset [9])

4.2. Classification model performance with the full feature set

Four classic machine learning models, Logistic Regression, KNN, Random Forest and Gradient Boosting, are trained on the complete encoded feature set under stratified 5-fold cross-validation. Table 2 summarizes the averaged four evaluation metrics of all models.

DOI: 10.54254/2753-8818/2026.PU38028

Table 2. Average classification metrics of the four models under 5-fold stratified cross-validation (full feature set)

Model	macro-F1	PR-AUC	ROC-AUC	Accuracy
Logistic Regression	0.7806	0.6393	0.8896	0.8710
KNN	0.7418	0.6230	0.8300	0.8827
Random Forest	0.7729	0.7114	0.9033	0.8961
Gradient Boosting	0.7819	0.7200	0.9084	0.8959

As shown in Table 2, Gradient Boosting achieves the optimal overall performance in terms of macro-F1, PR-AUC and ROC-AUC. Although Random Forest yields marginally higher accuracy, this metric is less informative under class imbalance. All models reach accuracy above 87% due to heavy class imbalance, yet this metric cannot reliably reflect the ability to identify potential buyers. Since PR-AUC is the core metric for imbalanced classification tasks and focuses on recognizing rare purchasing samples, Gradient Boosting is selected as the optimal model for further visualization analysis. The KNN model delivers the weakest predictive capacity in macro-F1, PR-AUC and ROC-AUC, which can be attributed to its distance-based lazy learning mechanism being disturbed by heterogeneous behavioral features. Logistic Regression serves as a stable baseline, while Random Forest exhibits competitive performance close to Gradient Boosting.

To further interpret the predictive behavior of the optimal Gradient Boosting model, confusion matrix, ROC curve and precision-recall curve are plotted on the held-out test set. Note that Figures 4-6 are generated on the held-out 20% test split, whereas Tables 2-4 report averaged metrics from stratified 5-fold cross-validation.

As visualized in Figure 4, the confusion matrix demonstrates that the model balances the identification of purchasing and non-purchasing visitors. Although a small number of potential buyers are misclassified as non-purchasing users, the model avoids the trivial all-negative prediction issue caused by data imbalance, which verifies the effectiveness of the sample weight imbalance mitigation strategy adopted in the Methodology section.

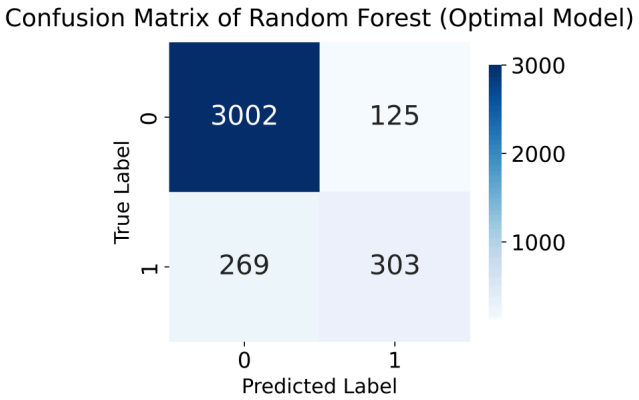

Figure 4. Confusion matrix of the Gradient Boosting classifier

(Data from: UCI Online Shoppers Purchasing Intention dataset [9])

As shown in Figure 5, the ROC curve is far above the random-guess diagonal baseline, with a test-set ROC-AUC consistent with cross-validation results, indicating stable discriminative ability

DOI: 10.54254/2753-8818/2026.PU38028

across all classification thresholds.

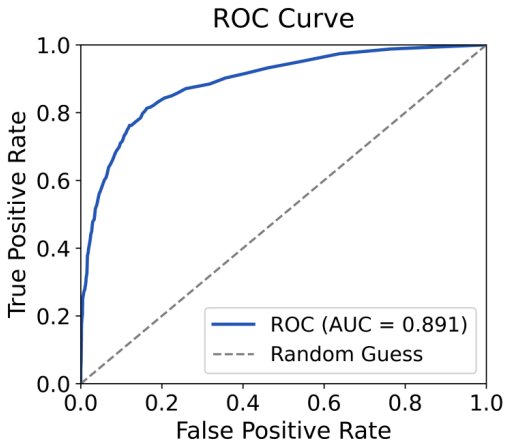

Figure 5. ROC curve of the Gradient Boosting classifier

(Data from: UCI Online Shoppers Purchasing Intention dataset [9])

As displayed in Figure 6, the precision-recall curve significantly outperforms the random baseline determined by the proportion of purchasing samples, confirming that the model can effectively capture positive transaction signals instead of relying on majority-class bias.

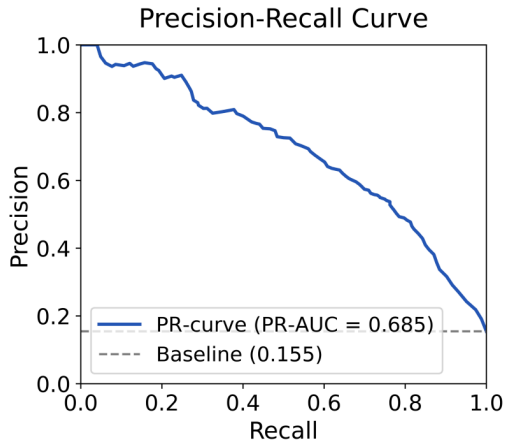

Figure 6. Precision-recall curve of the Gradient Boosting classifier

(Data from: UCI Online Shoppers Purchasing Intention dataset [9])

4.3. Feature importance and lightweight subset experiment

A balanced Random Forest model is trained on the full training set to quantify the predictive contribution of each input feature [13]. The resulting feature importance ranking is displayed in Figure 7.

DOI: 10.54254/2753-8818/2026.PU38028

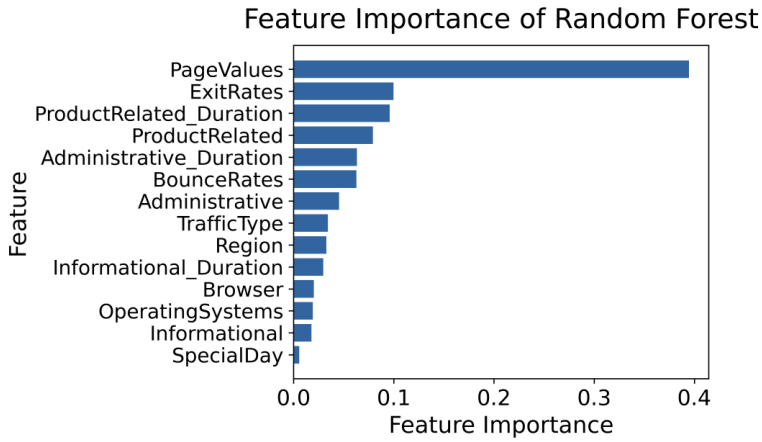

Figure 7. Feature importance ranking derived from Random Forest (full feature set)

(Data from: UCI Online Shoppers Purchasing Intention dataset [9])

As illustrated in Figure 7, PageValues ranks first in feature importance, which aligns with the correlation heatmap finding that PageValues holds the strongest linear association with the target Revenue. The top five most predictive features are PageValues, ExitRates, ProductRelated_Duration, ProductRelated and BounceRates. These five indicators constitute the core behavioral signals distinguishing purchasing intentions, while the remaining features contribute marginal predictive gains.

To verify the feasibility of lightweight prediction, all four models are retrained on the top-5 feature subset with identical stratified 5-fold cross-validation settings. The averaged metrics are presented in Table 3.

Table 3. Average classification metrics of the four models (top-5 important feature subset)

Model	macro-F1	PR-AUC	ROC-AUC	Accuracy
Logistic Regression	0.7842	0.6410	0.8853	0.8738
KNN	0.7647	0.6377	0.8507	0.8868
Random Forest	0.7624	0.6824	0.8896	0.8892
Gradient Boosting	0.7826	0.7066	0.9002	0.8962

Compared with the full-feature-set results in Table 2, removing low-contribution features brings only very limited performance changes for most models. For Gradient Boosting, PR-AUC decreases slightly from 0.7200 (full features) to 0.7066 (top-5 subset), while accuracy remains nearly unchanged at 0.8962 versus the full-feature value of 0.8959. This result demonstrates that the top-five core features preserve almost all critical predictive information, which supports lightweight model deployment using fewer input variables.

4.4. Single-feature control experiment of pagevalues

To validate the strong independent predictive signal of PageValues emphasized in the Methodology section, a control experiment is conducted using only the PageValues feature for model training. Table 4 records the cross-validation performance of the four models under single-feature input.

DOI: 10.54254/2753-8818/2026.PU38028

Table 4. Average classification metrics of the four models (only the PageValues feature)

Model	macro-F1	PR-AUC	ROC-AUC	Accuracy
Logistic Regression	0.7924	0.6602	0.8633	0.8891
KNN	0.7642	0.6261	0.8460	0.8852
Random Forest	0.7368	0.5485	0.7289	0.8723
Gradient Boosting	0.7697	0.6562	0.8617	0.8874

A control experiment uses only PageValues for model training (Table 4). PageValues exhibits considerable standalone predictive power: Logistic Regression and Gradient Boosting reach PR-AUC above 0.65, ROC-AUC around 0.86 and accuracy above 0.88, ruling out label-leakage risks.

Ensemble trees show divergent sensitivity to feature quantity. Random Forest degrades severely with only one input feature since it relies heavily on multi-feature interactions. In contrast, Gradient Boosting and Logistic Regression maintain stable single-feature performance. Notably, Logistic Regression achieves competitive macro-F1 and PR-AUC using PageValues alone, implying extra behavioral features may introduce noise for linear decision boundaries.

4.5. Comprehensive result summary

First, for imbalanced e-commerce purchase prediction, Gradient Boosting with sample weight adjustment achieves the best overall performance among the four mainstream algorithms.

Second, PageValues is the dominant predictive feature, and a lightweight model built on the top five core features retains nearly all of the predictive capacity of the full feature set.

Third, PageValues independently provides sufficient transaction prediction signals, which justifies the dedicated single-feature control experiment designed in the Methodology section.

Fourth, accuracy is misleading on this imbalanced dataset, while macro-F1 and PR-AUC reliably reflect model capacity to identify potential purchasing visitors.

5. Conclusion

This study conducts a series of comparative classification experiments on imbalanced e-commerce browsing session data to predict visitor purchasing intention. First, descriptive statistical analysis confirms serious sample imbalance in the dataset, proving that accuracy is an unreliable evaluation metric, while macro-F1 and PR-AUC can better measure a model's ability to identify potential buyers. Among the four tested algorithms, Gradient Boosting achieves the optimal comprehensive performance after sample weight adjustment, outperforming Logistic Regression, KNN and Random Forest on all core indicators.

Second, feature importance analysis identifies PageValues as the most powerful predictive variable. Controlled experiments further verify that the top five core behavioral features retain nearly all predictive capability of the full feature set, supporting lightweight model deployment with fewer input dimensions. Separate single-feature tests demonstrate that PageValues alone delivers stable prediction results, confirming its independent strong predictive signal without data leakage risks.

The main contributions of this work lie in the three groups of feature control experiments and the avoidance of synthetic oversampling methods that may distort real data distribution. Several limitations should be acknowledged as well: the dataset is collected from a single e-commerce

DOI: 10.54254/2753-8818/2026.PU38028

website within one year, session features are aggregated indicators rather than raw clickstreams, and potential heterogeneity between returning and new visitors has not been examined at the subgroup level. This research provides clear feature screening and model selection references for similar imbalanced user prediction tasks in online retail. Future work can incorporate finer-grained temporal behavior sequences, compare the transferability of behavioral signals across visitor subgroups, test deep learning classifiers, and explore dynamic threshold adjustment strategies to further improve the recognition of low-volume purchasing samples.

References

[1] Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications, 31*(10), 6893–6908. https://doi.org/10.1007/s00521-018-3640-7

[2] Kumar, D. A., Bhadrarao, M. V., Priya, P. S., Ramana, K. V., & Jyothi, M. (2024). Predicting online shopper behavior: Machine learning approaches for enhanced e-commerce insights. *International Journal of Scientific Development and Research, 9*(12), 21–27.

[3] Esmeli, R., Bader-El-Den, M., & Abdullahi, H. (2021). Towards early purchase intention prediction in online session based retailing systems. *Electronic Markets, 31*(3), 697–715. https://doi.org/10.1007/s12525-020-00448-x

[4] Chaudhuri, N., Gupta, G., Vamsi, V., & Bose, I. (2021). On the platform but will they buy? Predicting customers' purchase behavior using deep learning. *Decision Support Systems, 149*, 113622.

[5] Satu, M. S., & Islam, S. F. (2023). Modeling online customer purchase intention behavior applying different feature engineering and classification techniques. *Discover Artificial Intelligence, 3*(1), 36. https://doi.org/10.1007/s44163-023-00086-0

[6] Abdullah-All-Tanvir, Khandokar, I. A., Islam, A. K. M. M., Islam, S., & Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. *Heliyon, 9*(4), e15163. https://doi.org/10.1016/j.heliyon.2023.e15163

[7] Wen, Z., Lin, W., & Liu, H. (2023). Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data. *Systems, 11*(5), 255. https://doi.org/10.3390/systems11050255

[8] Yadav, S. (2023). Predicting consumer purchase intentions using machine learning. *Preprints*. https://doi.org/10.20944/preprints202311.0708.v1

[9] Sakar, C. O., & Kastro, Y. (2018). Online shoppers purchasing intention dataset [dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5F88Q

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[11] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC Press.

[12] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[13] Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25. https://doi.org/10.1186/1471-2105-8-25