Feature Dominance and Explanation Stability in Imbalanced Purchase Intention Prediction

Yiming Ding

School of Economics, Shandong University, Jinan, Shandong, 250100, China
m13863658686_1@163.com

Abstract. Purchase intention prediction supports conversion management in e-commerce, but evaluation can be misleading when purchase sessions are imbalanced and conversion-related features dominate prediction. This study examines session-level purchase intention prediction using XGBoost, four imbalance-handling strategies, PageValues ablation, and SHAP-based explanation stability analysis. The Online Shoppers Purchasing Intention Dataset is used as the empirical setting. Among all evaluated models, the complete-feature XGBoost model exhibited the best overall ranking performance, yielding an F1 score of 0.6785, an area under the receiver operating characteristic curve of 0.9330, an area under the precision-recall curve of 0.7446, and a Matthews correlation coefficient of 0.6183. Class weighting and random oversampling did not provide consistent improvement, while SMOTE reduced several metrics. Removing PageValues and its zero-value indicator reduced F1-score to 0.4242 and PR-AUC to 0.3703, showing that performance depended heavily on conversion-derived information. SHAP results further show that SMOTE changed top-ranked explanations more than class weighting and random oversampling. Purchase intention models should therefore be judged by predictive performance, real-time feature availability, and explanation stability in both offline experiments and practical deployment.

Keywords: Purchase intention; Class imbalance; Feature dominance; SHAP; E-commerce analytics

1 Introduction

Purchase intention prediction is an important task in e-commerce analytics because online platforms need to identify sessions that are likely to end with a transaction. A session records the behavior of an online shopper, including page visits, browsing duration, bounce rate, exit rate, traffic source, visitor type, and purchase outcome. These variables can support promotional allocation and customer journey optimization. Sakar et al. used multilayer perceptron and recurrent neural networks to predict online purchase intention in real time [1]. Satu and Islam found that feature engineering can strengthen customer

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_148

intention modeling [2]. Gomes et al. represented shopping sessions with embedding methods for purchase prediction [3]. Trivedi et al. applied machine learning to analyze behavioral factors that prompt online consumers to purchase [4].

Two methodological challenges remain important in this task. First, purchase sessions usually form the minority class, so a model may learn the non-purchase pattern more easily than the purchase pattern. Class weighting, random over sampling, and synthetic sampling are often used to address this issue, but imbalance correction is not automatically beneficial. Van den Goorbergh et al. showed that such corrections may harm risk prediction when they change the original data-generating structure [5]. Second, high-performing tree-based models can be difficult to interpret. SHAP provides an additive feature attribution framework for explaining individual and global predictions [6]. Chen et al. further showed that imbalance handling can affect the stability of post-hoc explanations in credit scoring [7].

These studies show that purchase intention prediction is not only a binary classification problem but also a basis for interpreting consumer behavior in digital retail. However, model usefulness depends on whether the learned signals can be collected before a marketing decision is made. The present study evaluates purchase intention prediction from three perspectives: predictive performance, feature availability, and explanation stability. XGBoost is used as the main classifier, and four imbalance strategies are compared under complete and PageValues-removed feature settings. SHAP rankings and rank-stability indicators are used to examine whether sampling strategies change the explanatory structure of the model.

2 Methodology

2.1 Dataset and Feature Engineering

The study uses the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository [8]. The dataset contains 12,330 online shopping sessions collected from a real e-commerce website over a one-year period. Each session belongs to a different user, which reduces repeated-user bias. The target variable is Revenue, where 1 indicates a purchase and 0 indicates no purchase. The dataset contains 10,422 non-purchase sessions and 1,908 purchase sessions, so the positive class accounts for 15.47 percent of all observations. Original variables include page counts, page durations, BounceRates, ExitRates, PageValues, SpecialDay, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, and Weekend.

Preprocessing converted categorical variables into one-hot indicators and transformed the target into numerical labels. Numerical variables were imputed with median values. Several behavioral features were constructed to describe session intensity and attention distribution. Total duration sums administrative, informational, and product-related page duration. Total page visits sums the corresponding page counts. Average page duration

measures time per visited page. Product view ratio, information view ratio, and administrative view ratio measure the share of page attention assigned to each page type. A PageValues zero indicator was also constructed.

Two feature settings were compared. The complete-feature setting retained all original and engineered variables. The no-PageValues setting removed both PageValues and the PageValues zero indicator. This ablation design quantified dependence on conversion-derived information and tested whether traditional behavioral variables can support prediction when this strong signal is unavailable. The comparison also separates post-session diagnostic features from features that are closer to real-time behavioral observation.

2.2 Model and Imbalance Strategies

XGBoost was selected as the main classifier because gradient boosting decision trees perform well on tabular data, handle nonlinear relationships, and capture feature interactions through sequential tree construction [9]. The model was evaluated under four imbalance-handling strategies. The original strategy trained the classifier on the observed class distribution. The class weight strategy increased the loss contribution of minority purchase sessions. Random oversampling duplicated minority-class observations. SMOTE created synthetic minority-class observations through interpolation among neighboring minority samples [10].

For each random seed, stratified sampling split the data into training, validation, and independent test sets. The validation set selected the probability threshold that maximized F1-score, and the test set was used for final evaluation. Twenty random seeds were used to reduce dependence on a single split. Precision, recall, F1-score, ROC-AUC, PR-AUC, balanced accuracy, and Matthews correlation coefficient were recorded. Because the task is imbalanced, F1-score, PR-AUC, recall, and Matthews correlation coefficient received greater emphasis than ordinary accuracy.

2.3 Explanation Stability

SHAP values were computed for selected XGBoost models to examine global feature attribution. Mean absolute SHAP values were used to rank variables. Top ten overlap measures the proportion of shared features between the top ten SHAP features of each imbalance strategy and those of the original strategy. Spearman rank correlation measures the consistency of global feature rankings across strategies. Paired Wilcoxon signed-rank tests across 20 random seeds were used to assess whether performance differences were statistically reliable.

3 Results

Table 1 reports the mean performance of XGBoost over 20 random seeds. Under the complete-feature setting, the original model achieved the strongest overall ranking metrics. Its F1-score was 0.6785, ROC-AUC was 0.9330, PR-AUC was 0.7446, and the Matthews correlation coefficient was 0.6183. Random oversampling produced the same rounded F1-score and slightly higher recall, but it did not improve ROC-AUC, PR-AUC, or Matthews correlation coefficient. Class weighting was close to the original model but did not provide an overall advantage. SMOTE produced the lowest F1-score among the complete-feature variants. These results indicate that imbalance handling did not consistently improve prediction when informative session variables were available.

Table 1. XGBoost performance over 20 random seeds.

Feature Set	Strategy	Precision	Recall	F1	ROC-AUC	PR-AUC	MCC
Complete features	Original	0.6465	0.7177	0.6785	0.9330	0.7446	0.6183
Complete features	Random over sampling	0.6387	0.7263	0.6785	0.9320	0.7420	0.6176
Complete features	Class weighting	0.6388	0.7207	0.6760	0.9320	0.7414	0.6148
Complete features	SMOTE	0.6277	0.7263	0.6723	0.9311	0.7370	0.6101
No PageValues	Original	0.3198	0.6407	0.4242	0.7815	0.3703	0.3047
No PageValues	Class weighting	0.3200	0.6310	0.4233	0.7801	0.3696	0.3018

The PageValues ablation produced the largest performance change. After PageValues and its zero-value indicator were removed, F1-score declined from 0.6785 to 0.4242, PR-AUC declined from 0.7446 to 0.3703, and Matthews correlation coefficient declined from 0.6183 to 0.3047. Class weighting under the same feature setting showed nearly the same result. The decline suggests that the complete-feature model relied heavily on PageValues-related variables.

The ablation also changed the feature ranking. In the complete-feature setting, PageValues and the PageValues zero indicator dominated the importance structure. In the no-PageValues setting, ExitRates, Average page duration, Product related duration, BounceRates, Product view ratio, Administrative view ratio, and Month November became more influential. November ranked highly, which is consistent with stronger seasonal shopping behavior around November. These variables describe browsing engagement and exit tendency, and they remain useful when direct conversion-derived information is removed.

Table 2 reports the Wilcoxon signed-rank test results. Differences between the original model and class weighting were not significant for F1-score, although the original model had a higher ROC-AUC. Random oversampling did not significantly improve the F1-score. SMOTE produced significantly lower F1-score and lower ranking metrics. The no-PageValues comparison showed a large and statistically reliable decrease in F1-score, PR-AUC, and Matthews correlation coefficient.

Table 2. Wilcoxon signed-rank test results.

Comparison	Metric	Difference	p-value	Interpretation
Original vs class weighting	F1	+0.0025	0.3488	Not significant
Original vs random over sampling	F1	0.0000	0.8983	Not significant
Original vs SMOTE	F1	+0.0062	0.0441	Original higher
Original vs class weighting	ROC-AUC	+0.0010	0.0027	Original higher
Original vs random over sampling	ROC-AUC	+0.0010	0.0240	Original higher
Original vs SMOTE	ROC-AUC	+0.0019	0.0002	Original higher
Complete vs no PageValues	F1	+0.2543	0.000002	Large decrease
Complete vs no PageValues	PR-AUC	+0.3743	0.000002	Large decrease
Complete vs no PageValues	MCC	+0.3136	0.000002	Large decrease

Table 3 presents the SHAP stability results. Class weighting and random over sampling maintained high top ten overlap with the original strategy under both feature settings. SMOTE produced lower overlap, especially after PageValues-related features were removed. In the complete-feature setting, the SMOTE top ten overlap was 0.68. In the no-PageValues setting, it declined to 0.58. Spearman correlations remained relatively high, but overlap results show that SMOTE altered the leading explanatory variables more clearly than the other imbalance strategies.

Table 3. SHAP explanation stability results.

Feature Set	Compared Strategy	Top ten overlap	Spearman rho
Complete features	Class weighting	0.90	0.9600
Complete features	Random over sampling	0.88	0.9486
Complete features	SMOTE	0.68	0.9124
No PageValues	Class weighting	0.86	0.9712

No PageValues	Random over sampling	0.88	0.9610
No PageValues	SMOTE	0.58	0.8979

4 Discussion

The results show that imbalance handling should be evaluated empirically rather than treated as a routine improvement. Although purchase sessions form a clear minority class, the original XGBoost model learned a strong decision boundary from the observed distribution. Random over sampling increased recall slightly, but it did not improve ranking metrics or the Matthews correlation coefficient. Class weighting also failed to improve the overall result. SMOTE performed worse, probably because interpolation among heterogeneous purchase sessions changed the local minority-class structure and introduced samples that did not match the real behavioral distribution. This pattern supports the view that imbalance correction can distort prediction when the original features already contain strong information [5].

A more important issue is feature availability. PageValues measures the value that a page contributes to transaction revenue and goal value in Google Analytics [11]. For that reason, it is closely connected with conversion outcomes. If this variable is calculated after a transaction or after the session ends, using it for real-time prediction may create a deployment gap. In an offline experiment, PageValues can increase performance sharply; in a production environment, the same feature may not be available when the platform decides whether to provide a coupon, show a recommendation, or trigger a retention intervention.

The no-PageValues experiment provides a more conservative estimate of real-time behavioral prediction. ExitRates, BounceRates, average page duration, product-related duration, and page-type ratios are more directly connected with observable browsing behavior during the session. Their importance suggests that conversion intention can still be inferred from depth of product interaction, browsing persistence, and reduced exit tendency. However, the lower PR-AUC and F1-score show that these behavioral indicators cannot fully replace PageValues. E-commerce platforms should therefore build both full-feature analytical settings and deployment-feasible settings. The first setting can help explain past conversions, while the second setting is more suitable for real-time targeting and intervention.

The explanation results add another practical implication. Predictive models are often translated into business rules, and such rules require stable evidence about why certain sessions receive higher purchase probabilities. Class weighting and random over sampling produced relatively stable SHAP rankings, while SMOTE changed the leading features more noticeably. In marketing applications, explanations are often used to justify intervention rules or guide website redesign. If explanations change substantially under a sampling strategy without improving performance, managerial trust may decline.

Evaluation of purchase intention models should therefore include explanation stability, especially when model outputs support business decisions.

5 Conclusion

This study examined purchase intention prediction under class imbalance, feature dominance, and explanation stability. XGBoost was evaluated with the original class distribution, class weighting, random oversampling, and SMOTE. The complete-feature original model achieved the strongest overall ranking performance, while imbalance-handling strategies did not provide consistent improvement. Removing PageValues and its zero-value indicator caused a large decline in F1-score, PR-AUC, and Matthews correlation coefficient, confirming that the complete-feature model depended strongly on conversion-derived variables.

The main contribution of the study is a three-part evaluation perspective for purchase intention prediction. A model should be assessed not only by predictive performance but also by the real-time availability of the features that drive its predictions and by the stability of its explanations. Page values can be valuable for post-session diagnosis, but it may be unsuitable for real-time intervention if it is unavailable before conversion. In that setting, behavioral variables such as exit rate, bounce rate, page duration, product attention, and seasonal indicators offer a more deployable but less powerful alternative.

Several limitations remain and should be considered when interpreting the empirical scope of the findings. The analysis uses one public dataset, so results may vary across platforms, product categories, and traffic sources. The data are session-level rather than longitudinal, so repeat-purchase history is not considered. Future research can test additional datasets, compare other sampling methods, and evaluate local explanation stability for individual shoppers.

References

1. Sakar, C.O., Polat, S.O., Katircioglu, M., Kastro, Y.: Real-time prediction of online shoppers purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications 31, 6893–6908 (2019)
2. Satu, M.S., Islam, S.F.: Modeling online customer purchase intention behavior applying different feature engineering and classification techniques. Discover Artificial Intelligence 3, 36 (2023)
3. Alves Gomes, M., Meyes, R., Meisen, P., Meisen, T.: Will this online shopping session succeed? Predicting customers purchase intention using embeddings. In: Proceedings of the 31st ACM International Conference on Information and Knowledge Management, pp. 2873–2882 (2022)
4. Trivedi, S.K., Patra, P., Srivastava, P.R., Zhang, J.Z., Zheng, L.J.: What prompts consumers to purchase online? A machine learning approach. Electronic Commerce Research 24(4), 2953–2989 (2024)

5. van den Goorbergh, R., van Smeden, M., Timmerman, D., Van Calster, B.: The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. Journal of the American Medical Informatics Association 29(9), 1525–1534 (2022)
6. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
7. Chen, Y., Calabrese, R., Martin-Barragan, B.: Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research 312(1), 357–372 (2024)
8. UCI Machine Learning Repository: Online Shoppers Purchasing Intention Dataset, https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset, last accessed 2026/06/08
9. 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, pp. 785–794 (2016)
10. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
11. Cutroni, J.: Understanding and Using Page Value. Google Analytics Blog, https://analytics.googleblog.com/2012/07/understanding-and-using-page-value.html, last accessed 2026/06/08

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

