Predicting Online Purchase Intention from Real-Time Browsing Behavior: A Comparative Study of Similarity-Based, Interpretable, and Boosted Models

Aleksandra Tatko (648925) | Team 1 - Group 3

January, 2026

1. Introduction

In digital marketing and e-commerce, the ability to identify potential buyers before a purchase is finalized is a critical practice that allows brands to maximize their conversion rates and increase sales. It also helps firms personalize their content, optimize targeting strategies, and allocate resources more effectively. As online shopping environments increasingly operate in real time, predictive models must rely on information available during an ongoing browsing session rather than after the transaction has occurred.

Browsing behavior data offers a rich but challenging source of information. Datasets consisting of page visit counts, time spent on product pages, bounce rates, and navigation patterns often reflect potential consumer interest; at the same time, they are often sparse, highly skewed, and characterized by a substantial number of sessions ending without a purchase. In many datasets, the majority of sessions exhibit little or no engagement, while purchases are relatively rare. Due to this class imbalance, both prediction and interpretation become challenging. The issue is particularly critical when the goal is to identify buyers and avoid generating excessive false positives.

Prior research has applied a wide range of machine learning techniques to address this problem. While researchers often focused on obtaining the highest predictive accuracy through complex models (C. O. Sakar et al. 2019; Jiang 2025), they tend to struggle to produce results that are both interpretable and useful for managerial decision-making. At the same time, creating simpler or more interpretable models may offer clearer insights, but at the cost of reduced performance. This trade-off raises an important methodological question: how do different modeling techniques compare in terms of both predictive power and actionable behavioral insights when applied to real-time browsing data?

Taking these issues into consideration, this paper aims to answer the following research question:

How well can online purchase intention be predicted from real-time browsing behavior, and how do interpretable tree-based models compare to boosting models in terms of predictive performance and actionable behavioral insights?

To address this question, the analysis focuses on three aspects. First, a k-nearest neighbors (KNN) classifier is created as an exploratory benchmark. It aims to assess whether similarity in browsing behavior alone can meaningfully distinguish buyers from non-buyers. Second, a conditional inference tree is estimated as an interpretable rule-based model and compared to a performance-oriented gradient boosting model (XGBoost). Finally, global interpretation techniques, including feature importance measures and partial dependence plots, are used to identify which browsing behaviors most strongly influence purchase intention and to assess the consistency of these effects across models.

2. Data Description

The following analysis is based on the Online Shoppers Purchasing Intention Dataset donated to the UC Irvine Machine Learning Repository database in 2018 (C. Sakar and Kastro 2018). The dataset consists of 12,330 entries collected from an e-commerce website, and includes variables that explain browsing behavior leading to online purchase. Each observation contains a single website entry that results in a binary outcome variable, *Revenue* equal to 1 if the purchase was made and 0 otherwise.

The dataset consists of 18 variables, 4 of which were categorical and the rest numerical. These features capture the real-time purchase intention for visitors in a single session. Numerical variables include the number of visited pages and the total time spent in different content categories (administrative, informational, and product-related pages), as well as session-level metrics such as bounce rate, exit rate, and page value. Page Value is a variable that measures the average revenue contribution of pages visited before a transaction. It reflects how strongly a browsing path is associated with completed purchases. Thus, PageValue cannot be considered a fully real-time variable (the implications of this limitation are examined in a later robustness analysis). Categorical variables describe contextual attributes, including operating system, browser, traffic source, region, visitor type (new or returning), month, and whether the session occurred during a weekend.

The majority of these variables capture real-time user interactions with the websites and can be observed only as a session unfolds. This makes the dataset suitable for studying real-time purchase intention prediction. The main challenge of the dataset is the significant class imbalance in the outcome variable. Approximately 15% of sessions result in a purchase, while the remaining 85% do not. Moreover, many behavioral variables contain a large number of zeros, especially among non-purchasing sessions. As a result, the given dataset is highly sparse. Exploratory analysis reveals that purchasing sessions are systematically more active. They contain fewer features equal to zero and higher engagement across most browsing dimensions. These properties have important implications for model choice, favoring methods that can handle sparse data, asymmetric decision boundaries, and rare positive events.

Prior to model estimation, categorical variables were converted to factors, and numerical variables were retained in their original scale, except where standardization was required for distance-based methods. Due to the class imbalance and to facilitate later comparison across different models, the data were split into training (70%) and test (30%) sets using stratified sampling to preserve the class distribution. This split ensures that enough data is secured to access the generalization performance, which is especially important when the positive outcomes are rare. All preprocessing steps relying on data distributions were performed on the training set only and subsequently applied to the test set to avoid information leakage.

The dataset contained a small number of identical observations, which were identified as sessions with minimal or no recorded activity. These observations were kept as they represent a valid part of the session when the engagement was low. Removing them would distort the empirical distribution of browsing behavior. Lastly, no missing values were present in the dataset, and no imputation was required.

3. Methodology

This study addresses the binary classification problem, where the main objective is to predict a relatively rare positive outcome. In the setting of high-dimensional and potentially sparse behavioral dataset methodological choices such as model selection, evaluation, and validation strategies, are particularly critical. All models are evaluated using a hold-out test set, which is not used during training or tuning. A fixed random seed (123) was used to ensure reproducibility. All models' hyperparameters were chosen through 5-fold cross-validation in the training set. Specifically, repeated k-fold cross-validation was used, where the training data are split into multiple folds, and the model is iteratively trained and validated across different partitions. Repeating this procedure several times lowers performance estimate variance and offers a more reliable foundation for comparing models (Hastie, Tibshirani, and Friedman 2009).

Given the marketing context of the problem, evaluation focuses on metrics that reflect the ability to identify potential buyers rather than overall classification accuracy. Because purchasing sessions are relatively rare, accuracy alone would be misleading, as a model could achieve high accuracy by predicting the majority

class. Instead, recall is emphasized to capture how effectively models identify purchasing sessions, while precision and F1-score are reported to reflect trade-offs between false positives and false negatives. ROC–AUC is used as a threshold-independent measure of ranking performance, allowing comparison across models with different decision mechanisms. Confusion matrices are subsequently used to interpret classification behavior at specific thresholds. These metrics are commonly recommended for rare-event classification problems, where identifying positive cases is often more important than minimizing false positives (Saito and Rehmsmeier 2015).

Similarity-Based Classification: k-Nearest Neighbors. The paper starts by exploring a k-nearest neighbor (KNN) classifier model, which compares new data points to existing ones rather than building a complex model. KNN considers the k most similar neighbors and assigns the new observation based on the majority class among its k. To detect which neighbors are the closest, KNN typically uses Euclidean distance, which provides a simple and intuitive measure of similarity in continuous feature spaces, when variables are standardized, and differences across dimensions are equally important (Cover and Hart 1967). Formally, let Nk (x) denote the set of indices corresponding to the k nearest neighbors of a new observation x. The KNN classification rule can be expressed as:

$$\hat{y}(x) = \arg\max_{c \in \{0,1\}} \sum_{i \in \mathcal{N}_k(x)} \mathbf{1}(y_i = c)$$

KNN creates an exploratory benchmark rather than a performance-optimal model. It provides insights into whether similarity in observed behavioral patterns alone is sufficient to distinguish positive from negative outcomes. Because KNN relies directly on distance calculations, feature scaling is required to prevent variables with large numeric ranges from dominating the similarity measure. At the same time, the standard KNN model does not require class weights or other mechanisms to handle class imbalance. As a result, this method is used purely as a baseline for assessing the usefulness of similarity-based approaches in imbalanced settings (James et al. 2023).

Interpretable Tree-Based Models: Conditional Inference Trees. To incorporate interpretability into the analysis, a conditional inference tree (CIT) is used. CIT is particularly suitable in settings with sparse behavioral data, mixed numeric and categorical predictors, and imbalanced outcomes, as its statistical testing framework avoids biased variable selection and naturally identifies meaningful behavioral thresholds. This makes CITs a well-suited and interpretable contrast to similarity-based approaches such as k-nearest neighbors.

CITs are a form of decision tree that was developed to overcome well-known biases in traditional classification and regression trees, particularly those arising from the way split variables are selected. Unlike CART-based trees, which rely on impurity measures such as the Gini index, conditional inference trees separate the process of choosing a splitting variable from the estimation of the split point itself (Hothorn, Hornik, and Zeileis 2006). This approach relies on formal statistical testing and helps prevent variables with many possible cut points from being selected too frequently.

At each node, the algorithm checks whether the outcome variable is statistically related to any of the available predictors. This results in the following hypothesis:

$$H_0 : \quad Y \perp\!\!\!\perp X_j \quad \text{for all } j \in \{1, \ldots, p\}$$

This hypothesis states that at any given point in the tree, none of the predictors are related to the outcome. If the null hypothesis cannot be rejected, the splitting process stops. If it is rejected, the predictor with the strongest statistical association with the outcome variable is selected. At the same time, its optimal split point is determined, and the algorithm continues. By using statistical tests to find the most significant variables, the CITs provide an unbiased and systematic approach. This makes CIT particularly attractive when interpretability and statistical validity are important. The resulting decision rules are easy to visualize and interpret on the graph, which allows for direct inspection of how combinations of features lead to specific predictions. However, single-tree models are known to be relatively unstable and may sacrifice predictive performance compared to ensemble methods, especially in complex or noisy classification tasks.

Unlike many classification models, conditional inference trees rely on permutation-based independence tests for split selection, which reduces bias toward majority-class predictors. As a result, no explicit resampling or class weighting was applied, since such procedures would compromise interpretability by duplicating observations and distorting the underlying decision structure.

Gradient Boosting for Binary Classification: XGBoost. In contrast to the previous models, XGBoost is not intended primarily for interpretability but for maximizing predictive performance. Its role is to establish an upper bound on achievable accuracy and to assess how much predictive improvement can be gained by modeling nonlinear interactions and asymmetric decision boundaries.

Gradient boosting is an ensemble method that builds a large number of simple decision trees and runs the algorithm sequentially, so that each new tree is trained to correct the errors made by the previous one (Friedman 2001). XGBoost is a widely used gradient boosting algorithm, used especially when it is important to improve predictive performance while controlling for overfitting. This is achieved through regularization and shrinkage (Chen and Guestrin 2016).

XGBoost estimates an additive model by minimizing a regularized objective function,

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) \; + \; \sum_{k=1}^{K} \Omega(f_k)$$

where $\ell(\cdot)$ denotes the loss function measuring prediction error and $\Omega(f_k)$ penalizes the complexity of each tree in the ensemble. This formulation allows the model to learn flexible nonlinear relationships while discouraging overly complex solutions.

Because decision trees can naturally handle variables with many zeros and capture nonlinear threshold effects, tree-based boosting methods are well-suited to sparse datasets with diverse predictors. Moreover, XGBoost can apply class weights when an imbalanced outcome variable is the issue. This technique increases the penalty for misclassifying rare positive outcomes. As a result, the model pays greater attention to correctly identifying purchasing sessions.

Cross-validation was used to tune key hyperparameters and improve predictive performance as well as prevent overfitting. The hyperparameters control features such as the number and depth of trees, the learning rate, the degree of regularization, and randomness introduced during training. Adjusting them allows the model to balance flexibility and generalization. This is particularly important when working with sparse predictors and imbalanced outcome classes.

Predictive performance indicates how well a model can rank sessions by their likelihood of resulting in a purchase, but the practical usefulness of a model depends on how these predictions are converted into decisions. In models such as XGBoost, which produce predicted probabilities rather than fixed class labels, the choice of probability threshold is particularly important. Different thresholds imply different trade-offs between false positives and false negatives. Evaluating model performance across a range of thresholds, therefore, makes it possible to tailor predictions to specific objectives, such as prioritizing the identification of potential buyers or achieving a balanced trade-off between precision and recall. This flexibility, combined with the ability to capture nonlinear patterns and handle sparse, imbalanced data, makes gradient boosting a useful contrast to simpler or more interpretable methods.

Model Interpretability and Global Interpretation Techniques While ensemble models often achieve strong predictive performance, they are frequently criticized for their lack of transparency. As a result, they are often called "black box" models. To help interpret these algorithms, several global interpretation methods are employed to complement predictive evaluation. SHAP (Shapley Additive Explanations) values are used to quantify feature contributions based on cooperative game theory (Lundberg and Lee 2017). SHAP assigns each feature a contribution value based on how much it increases or decreases the predicted probability for a given class. As a result, the plot shows which values of the variables influence the outcome the most, and which direction they follow In addition, partial dependence plots (PDPs) are used to visualize the marginal effect of selected features on the predicted outcome while averaging over all other variables

(Friedman 2001). Although they rely on the assumption of feature independence, the PDPs are particularly useful for identifying nonlinear patterns and threshold effects in tree-based models.

By combining these interpretation techniques, the analysis balances predictive accuracy with interpretability, enabling both methodological comparison and meaningful behavioral insights.

4. Analysis and Results

The following section aims to explain the distinct roles of three separate models. Each algorithm addresses a different analytical question. Firstly, the KNN classifier represents an exploratory benchmark and tries to assess whether the sole similarity in browsing behavior holds sufficient predictive power. Secondly, the conditional inference tree is introduced as an interpretable classification model. It focuses on transparent decision rules and behavioral thresholds. Finally, XGBoost is applied as a performance-oriented model created to maximize predictive accuracy and capture complex nonlinear interactions. This framework allows for a clear comparison between exploratory, interpretable, and predictive performance. *Figure 1* summarizes the performance of all models on the 30% held-out test set. The presented results clearly show performance differences across all the models.

Model Performance Comparison

Test set results across classification models

Model	ROC–AUC	Sensitivity	Specificity	Precision	F1-score	Balanced Accuracy
KNN	0.885	0.483	0.970	0.748	0.587	0.726
Conditional Inference Tree	0.916	0.551	0.958	0.706	0.619	0.754
XGBoost (High Recall, t = 0.20)	0.936	0.967	0.648	0.334	0.497	0.807
XGBoost (Balanced, t = 0.50)	0.936	0.851	0.867	0.539	0.660	0.859
XGBoost (High Precision, t = 0.70)	0.936	0.731	0.932	0.663	0.696	0.831
XGBoost (No PageValues, t = 0.50)	0.776	0.762	0.663	0.292	0.423	0.712

Figure 1: Model performance comparison

4.1 Exploratory Similarity Analysis: k-Nearest Neighbors

The KNN model is used as a baseline method that makes minimal assumptions about the data. It aims to answer the preliminary question: "*Does similarity in observed browsing behavior alone allow buyers to be distinguished from non-buyers?*", rather than maximize the predictive performance.

Hyperparameters were selected via cross-validation, with the number of neighbors set to k=25. Larger neighborhoods were preferred because smaller values of k led to highly unstable predictions in this sparse feature space. Since KNN relies on distance calculations, all numerical variables were standardized before model estimation so that no single feature dominated the Euclidean distance measure. As shown in *Figure 1* the KNN model results represent the weakest performance across all the models. However, ROC- AUC of approximately 0.87 reflects the model's ability to capture some meaningful predictive information. At the same time, sensitivity remains below 0.50, indicating that the model failed to capture more than half of the purchase sessions. This pattern is not uncommon in an imbalance setting. Many purchasing sessions are located in neighborhoods dominated by non-buyers, causing the majority voting to overwhelm rare positive sessions. While increasing the value of k improves the stability of the ranking, it further reduces sensitivity by strengthening the majority-class influence. These findings confirm that similarity alone is not sufficient in detecting positive events. This is especially true in sparse, imbalanced datasets. As a result, the analysis continues with a tree-based model, which can capture non-linear relationships and rule-based decision boundaries.

4.2 Interpretable Classification: Conditional Inference Trees

The following conditional inference tree model represents a clear, statistically significant approach that creates easy-to-interpret results. CIT was tuned with the following hyperparameters: the minimum split criterion and number of observations were set to a significance level of 0.95 and 100, respectively. This approach ensures that only statistically meaningful splits are created and the model avoids noise-driven decisions. Furthermore, the max depth of the tree was constrained to 3 to prioritise interpretability.

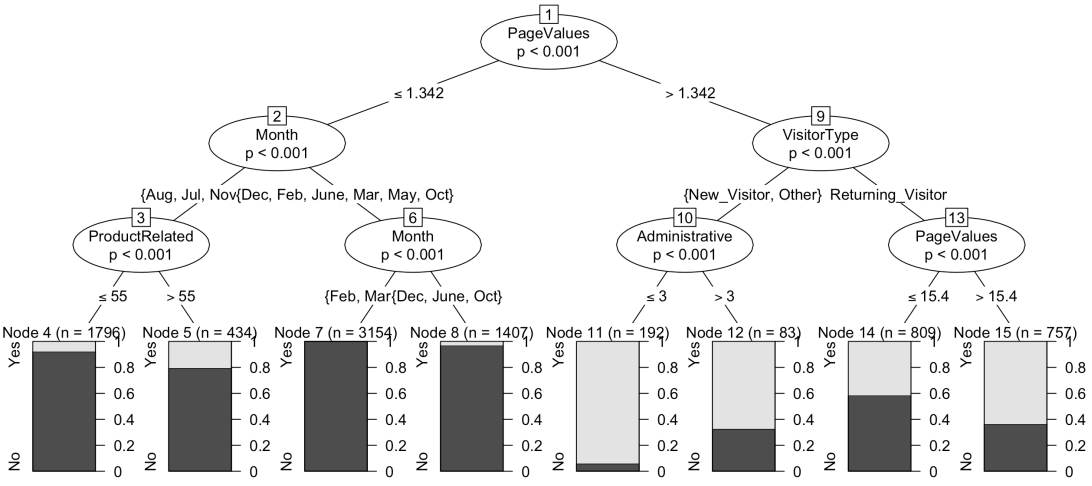

Figure 2: Pruned conditional inference tree used for interpretability

The CIT model achieves better results than KNN (*Figure 1*), especially in recall, which increased to around 0.55. At the same time, specificity remained high (approximately 0.96) and balanced accuracy increased slightly by 0.028. Beyond improved performance metrics, the CIT model provides valuable behavioral insights. As shown in *Figure 2*, early splits, hence statistically the most important, are dominated by Page Value, which represents the average revenue contribution of a page. Moreover, the model shows product-related activity and temporal factors, such as Month, as other significant variables. Sessions with very low product engagement are consistently classified as non-purchasing. At the same time, sessions with a higher activity level increase purchase probability, especially for returning visitors.

Exit rates and administrative page activity appear in later splits, which suggests that disengagement and navigation behavior further clarify predictions once basic engagement levels have already been established. This confirms that explicit behavioral thresholds are more effective than local similarity in identifying buyers.

4.3 Performance-Oriented Modeling: XGBoost

The final and main model of the analysis is the performance-oriented XGBoost algorithm. This model serves as an ensemble method that can capture complex, non-linear relationships and aims to reach high predictive accuracy. The model was tuned using repeated 3 times 5-fold cross-validation. The grid for the hyperparameters (and final chosen numbers) was as follows: the number of trees was set between 100 and 200 (final = 100); its depth was restricted to a maximum of 5 (5), to avoid overfitting; the most suitable learning rate was searched between 0.05 and 0.1 (0.05) to ensure gradual updates. To introduce randomness and improve generalization, subsampling and column sampling were applied. To handle class imbalance among the outcome variable, class weights were introduced, which increased the penalty for misclassifying purchasing sessions. As shown in *Figure 1*, XGBoost scores the highest across the majority of the chosen metrics. ROC–AUC increases substantially to 0.94 when compared with 0.88 and 0.92 for KNN and CIT, respectively. Recall also improves at a set threshold of 0.5 and reaches 0.85. These improvements show the model's ability to capture nonlinear interactions between engagement, exit behavior, and visitor characteristics.

Since XGBoost produces predicted probabilities rather than direct class labels, its performance depends on the set decision threshold. To take that into account, the model was analysed across three threshold scenarios - each representing valuable information depending on the final business problem.

When the decision threshold was equal to 0.2, the model prioritized recall and correctly identified the majority of successful sessions (sensitivity = 0.97), outperforming other models in that metric. However, at the same time, the precision drops, which indicates a higher number of false positives. This model might be suitable in situations where missing a potential buyer is particularly costly, even if this means targeting more users who ultimately do not purchase.

At the default threshold of 0.5, the model achieves a more balanced trade-off between sensitivity and precision. Balanced accuracy and F1-score are maximized at this threshold, and reach 0.85 and 0.66, respectively. This represents a stable compromise between identifying buyers and, at the same time, avoiding excessive false alarms.

When the threshold is set to 0.7, the precision improves further to 0.66, but at the cost of sensitivity, which declines to 0.73. While the F1 score reaches its highest value at 0.70, balanced accuracy drops to 0.83. This model becomes more conservative, identifying fewer buyers but with higher confidence. As a result, this threshold is preferred in situations when the budget is limited, and the firm prefers to target only more certain customers.

Given the primary objective of the paper, which is identifying purchase intention in real time, a threshold of 0.5 was chosen as the final model. This follows the real-life logic, when companies often face a trade-off between missing the potential buyers and allocating resources efficiently. This threshold provides a stable balance between the marketing objectives and serves as a suitable benchmark for comparing models and interpreting behavioral patterns.

4.4 Robustness Analysis: Excluding PageValues

To evaluate whether model performance depends on information that may not be fully available in real time, the analysis followed with an XGBoost model without the PageValues variable. This feature can bias real-time prediction because it is computed using information related to completed transactions. Excluding PageValues results in a significant decline in performance. ROC–AUC and precision decrease by 0.16 and 0.25, respectively. Moreover, although sensitivity remains relatively high (=0.76), the F1 score drops by 0.24, and the model creates substantially more false positives. Despite this decline, the model without PageValues still performs comparably to the CIT model and outperforms KNN in terms of recall. Overall, this indicates that despite the Page Value variable that holds a lot of information, core engagement variables continue to provide a meaningful predictive signal. This robustness check supports the interpretation of the remaining features as meaningful real-time predictors.

4.5 Global Feature Importance and SHAP Analysis

The XGBoost model is commonly known as Black Box, which means its direct interpretation is limited. As a result, this analysis used multiple global interpretation methods to better understand the model. *Figure 3* presents the SHAP summary plots for XGBoost with and without the PageValues variable. Feature values were rescaled within each variable to a common scale (1-10) for better interpretability, as original variables differ greatly in scales (for example, time-based variables versus rates). This transformation affects only the color gradient in the plot and does not change the SHAP values or the model results.

Both figures show a similar pattern in variable effects with slightly different ordering. When PageValues is present, it dominates the SHAP because it contains near-direct information about transaction value and is therefore strongly correlated with the purchase outcome. As a result, it overshadows all other behavioral variables. Furthermore, in both plots, Exit Rates clearly dominate as one of the most influential predictors. Higher values strongly reduce the likelihood of purchase, reflecting the user's disengagement. Bounce rates also tend to reduce the probability of a successful session, although their effect is nonlinear and more dispersed. On the other hand, the variable such as ProductRelated_Duration shows that the more time the user spends on the product website, the higher the probability of a successful purchase session. The variable

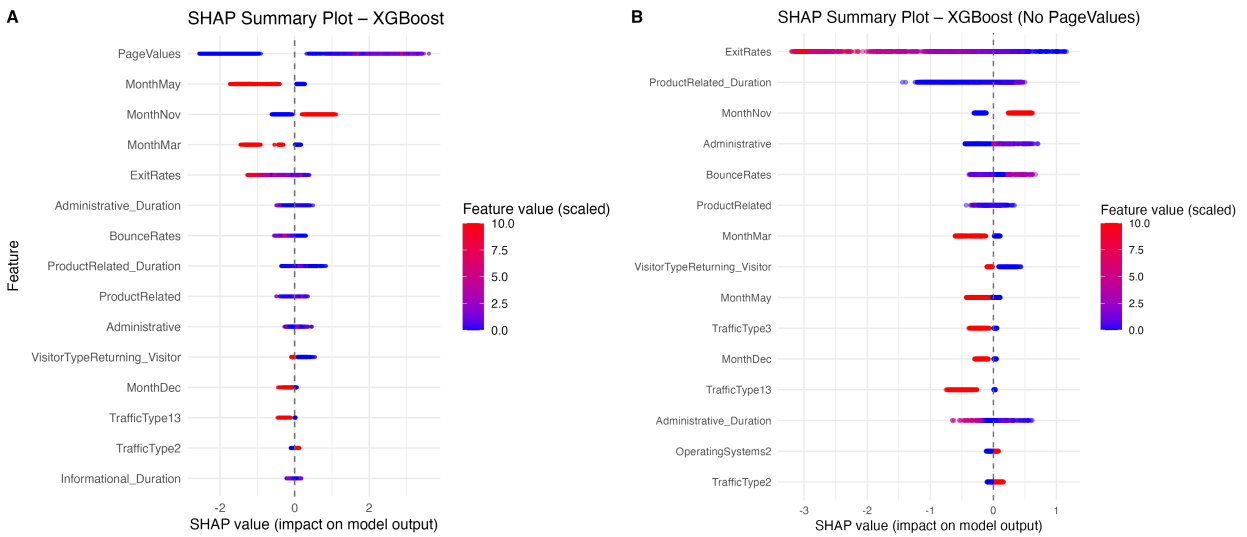

Figure 3: SHAP summary comparison

Month reflects the seasonal impact on the purchasing behavior. Sessions that occur in November have a higher probability of success, while May and March lower the likelihood of purchase. Returning visitors generally exhibit slightly positive SHAP values, which suggest a higher likelihood of conversion. Other variables, such as traffic source and administrative activity, have smaller contributions and place closely around zero. This indicates a limited marginal impact once engagement and exit behavior are accounted for. Overall, the two SHAP plots provide consistent evidence that, in the absence of PageValues, the model relies primarily on interpretable real-time browsing signals. Engagement with product content increases purchase probability, while indicators of disengagement sharply reduce it. The similarity between the two figures confirms the robustness of these findings and supports the interpretation of these behavioral variables as key drivers of online purchase intention.

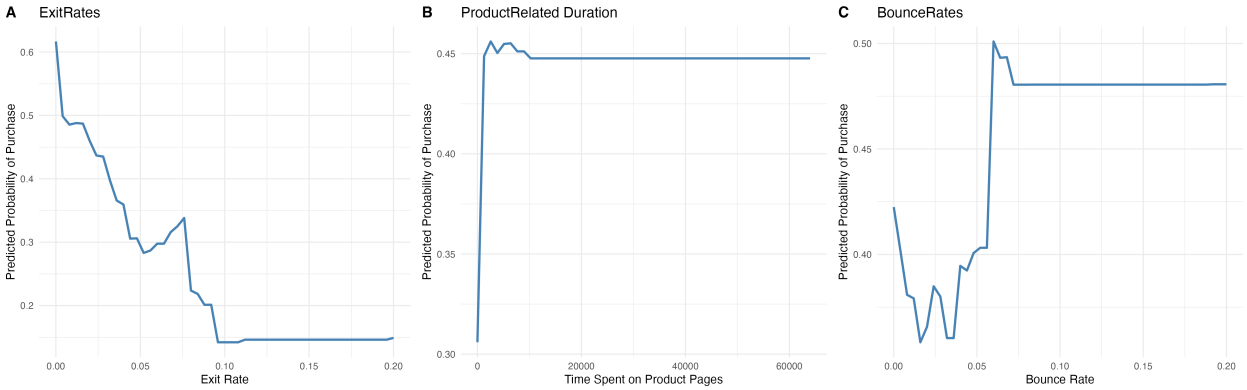

Figure 4: Partial dependence plots for key behavioral variables.

Figure 4 represents the outcome of further analysis for partial dependence plots for the model without PageValues.

Exit rates show the strongest negative impact (*Figure 4A*). At low values, the likelihood of purchase remains relatively high. However, when the number increases, the purchase probability drops sharply. The threshold of around 0.10 creates a point after which the likelihood of conversion becomes very low (around 0.15). From a practical perspective, this suggests that identifying the increasing exit behavior early in the session is important, as there might still be time to intervene through, for example, personalised offers or navigation

support.

The plot for product-related browsing time (ProductRelatedDuration; *Figure 4B*) reveals the initial positive relation with the purchase likelihood, which stabilises at around 0.45 after a certain threshold. This suggests that early engagement is critical, but extended browsing adds limited additional information. This saturation effect is consistent with consumer decision-making theory- when enough information is provided, the additional browsing does not change the purchase decision. Noticeably, this non-linear effect validates the use of tree-based models, as linear methods would most likely fail to capture this relation. From a practical standpoint, the results imply that identifying users who cross the initial engagement threshold is more valuable than targeting those who browse websites for a prolonged period.

Figure 4C shows the PDP for BounceRates, and it confirms the nonlinear effect of the variable. When the values are low, even a slight increase decreases the likelihood of purchase. This suggests that the sole bouncing does not guarantee a successful purchase. As rates increase to a moderate level, purchase probability rises and reaches its highest point. Beyond this threshold, the likelihood of purchase decreases slightly and then remains stable. This indicates that moderate bounce behavior connects with a higher purchase likelihood. When combined with the SHAP analysis, the result implies that BounceRates contribute most strongly when identifying sessions that are unlikely to convert, rather than distinguishing among high-intent users. The model, therefore, uses bounce behavior mostly as a filtering mechanism to eliminate low-quality sessions from further consideration.

5. Discussion and Conclusion

This paper analyzed how well online shopping intentions can be predicted using real-time browsing data and compared different modelling approaches in terms of both accuracy and interpretability. The results revealed that even with sparse, highly imbalanced data, machine learning (ML) models can achieve strong predictive accuracy while remaining interpretable and applicable in real-time decision-making. Three models applied in the paper represent different attributes of ML models, and each contributed to the results in their own way. The KNN model serves as a useful baseline, proving that the sole similarity is not sufficient to predict rare purchasing events. On the other hand, the CIT model improves the predictive performance by introducing a highly interpretable tree-based model, with transparent decision rules and meaningful behavioral thresholds. However, the limited predictive accuracy led the analysis towards the third, final model - XGBoost. This method achieves the strongest predictive accuracy by capturing nonlinear relationships between engagement, exit behavior, and visitor characteristics. With the help of global interpretation models such as SHAP and PDPs, the black box model was transformed into an interpretable and behaviorally meaningful algorithm with strong predictive results. Overall, these findings show that online purchase intention can be predicted reliably from real-time browsing behavior, with gradient boosting achieving strong performance while interpretable tree-based models provide actionable behavioral insights.

The key finding of the study shows that not only does algorithm selection influence the final usefulness of the model, but also the decision threshold. By adjusting the probability threshold, the same model can support different business objectives, such as prioritizing recall to avoid missing potential buyers or prioritizing precision to control targeting costs. In the case of real-time purchase intention prediction, a threshold of 0.5 was chosen as it provides a realistic trade-off between identifying buyers and avoiding excessive false positives. This highlights that predictive models should be evaluated not only by their ranking performance but also by how their outputs are translated into actions.

Moreover, the analysis emphasized the importance of the availability of real-time features. PageValues, although highly influential, are closely linked to completed transactions and might not be fully observable during an active session. The robustness analysis with the XGBoost model excluding this variable proved that strong predictive performance can still be achieved using mainly real-time behavioral signals such as product engagement and exit-related metrics, although with some loss in precision. This finding suggests that when designing real-time prediction systems, the timing of available information can be as important as model selection.

From a practical, managerial perspective, these findings suggest that companies can use real-time browsing data to intervene early and, at the same time, increase the likelihood of a successful session right at the

beginning of the journey. By identifying users who are either likely to convert or leave the website while the session is still active, marketing teams can create targeted campaigns such as personalized recommendations, dynamic content, or timely promotional messages. Importantly, the finding also suggests that adjusting the decision thresholds allows firms to align these interventions with budget constraints and campaign objectives.

Finally, several limitations of the paper should be acknowledged. Firstly, the analysis is based on a single e-commerce dataset on a session-level, which might limit the generalizability. Moreover, interpretation methods rely on simplifying assumptions. Future research could extend this work by introducing multiple datasets incorporating time-dependent or sequential models and exploring adaptive threshold selection. Despite these limitations, the results show that real-time browsing behavior contains valuable predictive information and that combining high-performing models with interpretability tools is an essential practice for decision-making in digital marketing.

6. References

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *CoRR* abs/1603.02754. http://arxiv.org/abs/1603.02754.

Cover, Thomas M., and Peter E. Hart. 1967. "Nearest Neighbor Pattern Classification." *IEEE Transactions on Information Theory* 13 (1): 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15 (3): 651–74. https://doi.org/10.1198/106186006X133933.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. 2023. *An Introduction to Statistical Learning: With Applications in r*. 3rd ed. New York: Springer. https://www.statlearning.com/.

Jiang, Q. 2025. "Purchase Intention Analysis of Online Shoppers Based on Machine Learning and k-Means Clustering." In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE)*, 259–64. SciTePress. https://doi.org/10.5220/0013686100004670.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1705.07874.

Saito, Takaya, and Marc Rehmsmeier. 2015. "The Precision-Recall Plot Is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLoS ONE* 10 (3): e0118432. https://doi.org/10.1371/journal.pone.0118432.

Sakar, Cemal Okan, Suleyman Olcay Polat, Mete Katircioglu, and Yomi Kastro. 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/s00521-018-3523-0.

Sakar, C., and Yomi Kastro. 2018. "Online Shoppers Purchasing Intention Dataset." UCI Machine Learning Repository.