Journal of Management Science Research Review
https://jmsrr.com/index.php/Journal/about

Volume 4 Issue No 3 (2025)

Online ISSN: 3006-2047
Print ISSN: 3006-2039

Explainable Machine Learning for Predicting Online Purchase Intention: A Decision-Support Framework for Digital Marketing

Asfandyar Khan
Institute of Computer Sciences & Information Technology (ICS/IT), The University of Agriculture, Peshawar, Pakistan Email: asfandyar@aup.edu.pk

Awais Khan
IBMS, University of Science & Technology Bannu, Pakistan
Email: awaisuni@gmail.com

Muhammad Jehangir
Institute of Business Studies and Leadership, Abdul Wali Khan University, Mardan, Pakistan Email: salman.ce27@gmail.com

Majid Khan
Department of Mathematics, Statistics and Computer Science, The University of Agriculture, Peshawar – Pakistan Email: majidkhan@aup.edu.pk

Abstract
Predicting the likelihood that a visitor to a website will purchase during a browsing session is a core problem of digital marketing analytics, and it fuels real-time personalization, retargeting and conversion-rate-optimization decisions. In this study, we propose and empirically investigate an explainable machine learning decision-support framework for online purchase-intention prediction based on UCI/Sakar et al. Online Shoppers Intention. dataset (12,330 e-commerce sessions, 17 behavioral and technical attributes). We trained and compared four supervised learning algorithms on accuracy, precision, recall, F1-score and ROC-AUC: Random Forest, Gradient Boosting, a histogram-based gradient boosting model in the same algorithmic family as XGBoost, and an Artificial Neural Network (Multilayer Perceptron). To provide transparency of model reasoning to marketing decision-makers, Shapley Additive exPlanations (SHAP) values were computed for 120 held-out sessions per model using a Monte Carlo permutation-sampling estimator of the exact Shapley value . The SHAP values afford a global feature-importance ranking as well as instance-level, session-specific explanations. Gradient Boosting showed the best discrimination (ROC-AUC = 0.928, recall = 0.828), followed by the histogram-based gradient boosting model (AUC = 0.923), the neural network (AUC = 0.913) and Random Forest (AUC = 0.909). PageValues, a metric that measures the monetary value of the pages a visitor views, was by far the dominant driver of predicted purchase intention across all four models, followed by Month (seasonality), ExitRates and TrafficType. PageValues had a mostly monotonic, non-linear relationship with purchase probability, and high ExitRates suppressed purchase likelihood even after controlling

for on-site engagement, as confirmed by SHAP dependence analysis. We show that the combination of strong ensemble and neural models with careful session-level SHAP explanations produces a decision-support artifact that is both accurate and actionable, allowing marketing teams to focus real-time interventions (e.g., dynamic incentives for high-PageValues, high-ExitRate sessions) on transparent, auditable model reasoning as opposed to opaque black-box scores.

Keywords: Explainable AI, SHAP, Purchase Intention, E-Commerce Analytics, Digital Marketing, Ensemble Learning, Artificial Neural Networks, Conversion Prediction, Decision Support

Introduction

The growth in the digital technologies has allowed to be a part of various applications (Bangash et. al., 2014). E-commerce platforms produce an enormous amount of clickstream data that describes how visitors browse, stay and leave in a browsing session. The ability to accurately predict purchase intent from this behavioral exhaust in real time has become a strategic priority for digital marketing teams seeking to optimize promotional spend, trigger retargeting campaigns and personalize on site experiences (Khattak et al., 2019). Machine learning algorithms have been shown to achieve good predictive performance on this task, in particular ensemble tree methods and neural networks. But marketing and customer-experience teams can’t responsibly act on a prediction they can’t interpret: a black-box model that just spits out a purchase-probability score with no explanation offers limited guidance on which lever to pull for a given customer segment or session.

This work addresses this interpretability gap by developing an explainable machine learning framework that integrates powerful predictive models with transparent, quantitative explanations of their reasoning. We trained and rigorously compared four algorithms in bagging (Random Forest), traditional boosting (Gradient Boosting), modern histogram-based boosting (in the algorithmic family popularized by XGBoost) and deep feed-forward learning (Artificial Neural Network) on the Online Shoppers Intention dataset. We used SHAP for explainability, a game-theory based attribution method that decomposes each individual prediction into additive feature contributions.

The paper is organized around three questions:

RQ1: What class of machine learning model (ensemble tree-based vs. neural) produces the best predictive performance for online purchase-intention classification?

RQ2: What are the main behavioral and contextual drivers of predicted purchase intention and do these drivers differ across structurally different model families?

RQ3: How to operationalize SHAP based explanations as a practical decision support tool for digital marketing teams?

Literature Review

Clickstream and session level features, such as page views, time-on-page, bounce and exit rates and traffic source, have been widely used to study online purchase intention.

Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Prior work leveraging tree ensembles and neural networks on the Online Shoppers Intention dataset and related clickstream corpora has generally demonstrated strong discriminative performance (ROC-AUC in the 0.85-0.95 range), with PageValues, a Google-Analytics-derived metric that reflects the average value of pages leading to a transaction, consistently a top single predictor across studies.

Random Forest: Combines many decorrelated decision tree models using bootstrap aggregation. This gives robustness to overfitting and good performance on tabular data with mixed numeric and categorical features. Gradient Boosting builds trees one after another, each correcting the residual errors of its predecessors, and usually achieves higher predictive accuracy than bagging on structured/tabular prediction tasks, but is more sensitive to hyperparameters. Modern histogram-based boosting implementations, most notably XGBoost, LightGBM, and scikit-learn's HistGradientBoosting , discretize continuous features into histograms when training, which greatly speeds up the process of finding splits while maintaining or improving predictive performance compared to classical gradient boosting (Khattak et al., 2018). These methods have become the de facto standard for high-performance tabular machine learning in industry deployments (Mukhtar et al., 2019).

Feed-forward artificial neural networks (multilayer perceptrons) can model complex non-linear feature interactions and have been applied to purchase-intention and customer-churn prediction with competitive results. However, they usually require more careful regularization and hyperparameter tuning than tree ensembles to avoid overfitting on tabular datasets of moderate size, and typically remain less interpretable without post-hoc explanation methods.

Shapley Additive exPlanations SHAP Lundberg and Lee (2017) is a unified framework for a family of additive feature-attribution methods based on the Shapley Value of cooperative game theory. The SHAP value of each feature is the average marginal contribution of the feature to the prediction over all possible coalitions of features, and satisfies desirable theoretical properties such as local accuracy (additivity), consistency, and missingness. SHAP values can be computed exactly for small feature sets, by efficient model-specific algorithms (e.g., TreeSHAP) for tree ensembles, or approximated for arbitrary models by Monte Carlo sampling over feature orderings. The permutation-sampling approach was formalized by Štrumbelj and Kononenko (2010) and is the model-agnostic estimator we use in this study, which can be applied uniformly to all four model families we evaluate (Batool et al., 2016). In digital marketing contexts, SHAP has been applied to explain churn models, recommendation rankings and conversion-propensity scores, translating opaque model output into feature-level narratives that marketing analysts can act upon (Imran et al., 2023; Khattak et al., 2021).

It is common to compare ensemble and neural models on purchase-intention data with an emphasis on accuracy, but few studies have used a uniform, model-agnostic SHAP approach consistently across bagging, boosting, and neural architectures on the same data to compare predictive performance and stability of feature-attribution rankings across model families, a comparison that speaks directly to whether marketing teams can trust a single, consistent explanatory narrative in spite of the model ultimately

Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

deployed (Khan et al., 2018; Khattak et al., 2016). The present study addresses this gap.

Research Methodology

The proposed methodology is discussed in the following sub-sections.

Dataset

Publicly available datasets are important for the variety of today's machine learning applications (Khalil et. al., 2016, Zaman et. al., 2020). In this paper, we use the public Online Shoppers Intention data set (Sakar et al.) which has 12,330 e-commerce sessions collected over a one year period. Each session is described by 17 predictor features and a binary target, Revenue, which is the indicator of whether the session resulted in a transaction. Features include the page count and duration variables for the three page categories (Administrative, Informational, Product-Related), the aggregate behavioral metrics (BounceRates, ExitRates, PageValues), a seasonal indicator (SpecialDay, proximity to a special calendar event), categorical browsing context (Month, OperatingSystems, Browser, Region, TrafficType, VisitorType), and a Weekend flag. The dataset is complete and has no missing values, but the class distribution is highly imbalanced: 1,908 (15.5%) sessions led to a purchase, whereas 10,422 (84.5%) sessions did not (Figure 1).

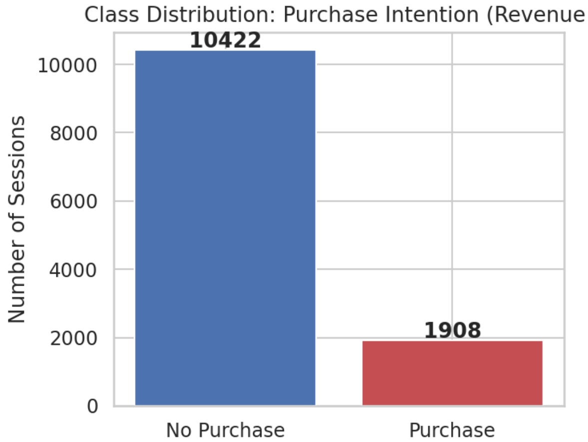


Figure 1: *Class distribution of the target variable*

Data Pre-processing

The data set was split randomly into a stratified 75% training set (n = 9,247) and 25% held-out test set (n = 3,083), maintaining the 15.5% purchase base rate in both partitions. The ten sequential behavioral variables (e.g. PageValues, BounceRates, ProductRelated_Duration) were normalized to zero mean and unit variance . Seven categorical/nominal features (Month, VisitorType, Weekend, OperatingSystems, Browser, Region, TrafficType) were one-hot encoded. OperatingSystems, Browser, Region and TrafficType are kept as integer codes in the raw data but were treated as

Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

nominal categories as these codes do not have any meaningful ordinal relationship. To prevent data leakage, all pre-processing transformations were fitted on the training partition only and applied unchanged to the test partition. To handle the class imbalance, we employed balanced sample weighting (inverse-frequency weighting) during fitting a model instead of resampling, which preserves the natural test-set distribution for an unbiased performance evaluation.

Machine Learning Models

Four supervised classifiers were implemented using scikit-learn 1.8:

Random Forest: ensemble of 400 trees with balanced class weights (max depth = 12) and prediction aggregation by majority vote over bootstrap samples.

Gradient Boosting: a classical sequential boosting ensemble of 250 shallow trees (max depth = 3, learning rate = 0.05), trained with balanced sample weights.

Histogram Gradient Boosting: a histogram-binned gradient boosting ensemble (300 boosting iterations, max depth = 6, learning rate = 0.06, L2 regularization = 0.5), implemented via scikit-learn's HistGradientBoostingClassifier.

Artificial Neural Network (Multilayer Perceptron) – feed-forward network with two hidden layers (64 and 32 units respectively), L2 regularization (alpha = 1e-3), early stopping trained with balanced sample weights.

All models were run with a fixed random seed (random_state=42) to ensure reproducibility.

Predictive Performance Metrics

The model performance was evaluated on the held-out test set using accuracy, precision, recall, F1-score, and ROC-AUC . Given the 15.5% base rate of the positive (purchase) class, we chose ROC-AUC, recall, and F1-score as the main metrics for comparison, because raw accuracy can be inflated by the majority-class prediction in imbalanced settings.

Explainability: SHAP via Monte Carlo Permutation Sampling

The explainability was established using Shapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) calculated using a Monte Carlo permutation-sampling estimator of the exact Shapley value (Štrumbelj & Kononenko, 2010) and applied in the same way and in a model-agnostic fashion to all four classifiers. For each of 120 randomly selected held out test sessions, and for each of 12 random orderings (permutations) of the 17 original features, a background session was gradually transformed into the target session by replacing one feature at a time in the order of the permutation. The change in each model’s predicted purchase probability at each step was assigned to the feature just replaced. We repeated this procedure 10 times for independently drawn background sessions per permutation and averaged these values, providing an unbiased Monte Carlo estimate of the Shapley value for each feature for each explained session. This estimator enjoys the defining efficiency property of SHAP, that the sum of the feature attributions for a given session is equal to the difference between the model prediction for that session and the average prediction

Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

over the background sample. The sum of the absolute SHAP values for the 120 explained sessions provides a global feature-importance ranking for each model, while the signed, session-level SHAP values allow for local, case-by-case explanation of individual purchase-probability predictions.

Results

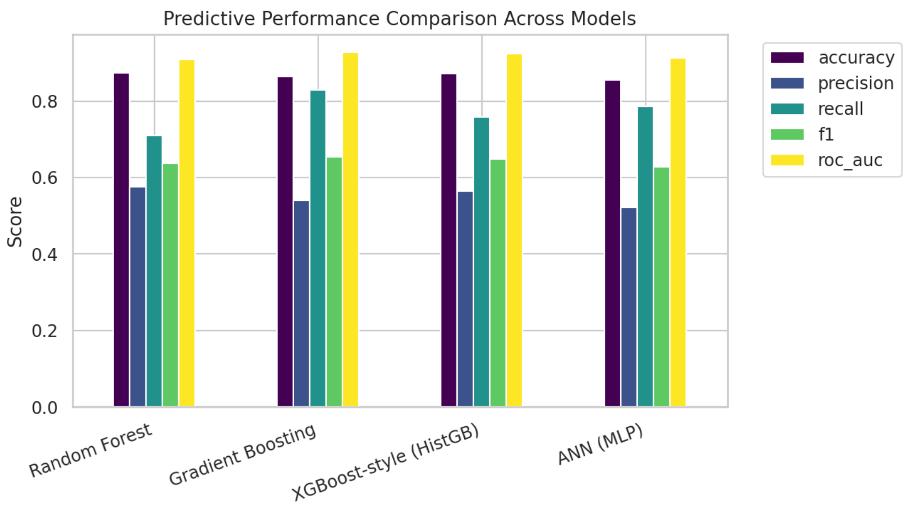
The following sub-sections discuss the simulation results of the proposed paper.

Predictive Performance

Table 1 shows the test-set performance of all four classifiers (n = 3,083). Gradient Boosting had the best overall discrimination (ROC-AUC = 0.928) and the highest recall (0.828), correctly flagging the largest share of true purchasing sessions, an important property for a marketing decision-support tool, where failing to identify a genuine buyer (a false negative) typically carries a higher opportunity cost than a false positive retargeting impression. The histogram-based gradient boosting model (XGBoost-style) came in a very close second (AUC = 0.923, recall = 0.759). The Artificial Neural Network had a strong AUC of 0.913 with the second highest recall (0.786). Out of the four models, Random Forest achieved the highest precision (0.577) and the highest overall accuracy (0.874), but the lowest recall (0.711), which implies a more conservative decision boundary. All four models outperform a naive majority-class baseline (84.5% accuracy but 0.0 recall), showing that true discriminative learning is present across all architectures.

Table 1: *Performance Evaluation of the four classifiers using testing subset*

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.874	0.577	0.711	0.637	0.909
Gradient Boosting	0.864	0.540	0.828	0.653	0.928
XGBoost-style (HistGB)	0.872	0.565	0.759	0.648	0.923
ANN (MLP)	0.856	0.522	0.786	0.628	0.913



Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Figure 2: *Comparative Analysis of the four models in terms of accuracy, precision, recall, F1-score, and ROC-AUC*

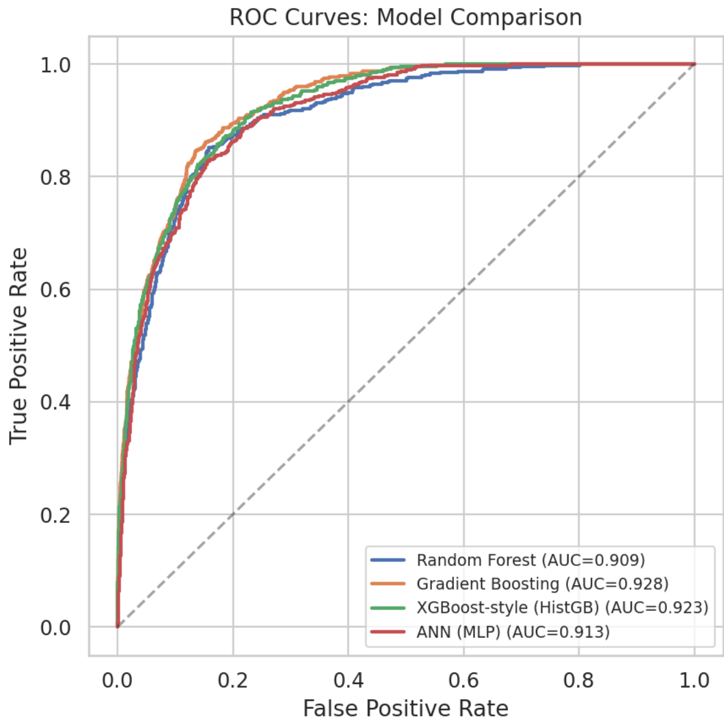


Figure 3. ROC curves of the four classifiers using testing subset

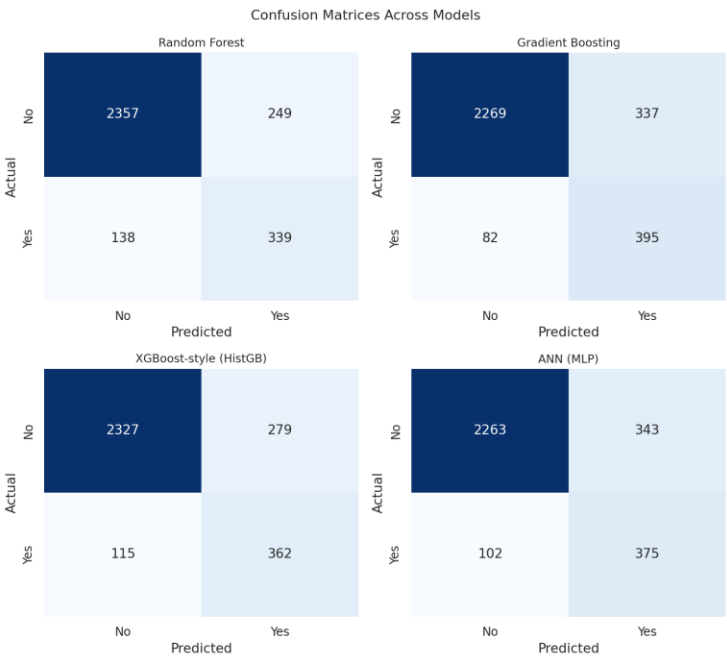


Figure 4. Confusion matrices for all four classifiers on the held-out test set.

Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Global Feature Importance via SHAP

Figure 5 shows the top-10 features ranked by mean absolute SHAP value for each of the four models on 120 randomly sampled held-out sessions. PageValues is by far the most influential feature across all models with mean absolute SHAP contributions of 0.123 (Random Forest) to 0.204 (Gradient Boosting), which is approximately three to five times larger than the second-ranked feature in each model. Month is consistently second (mean |SHAP| between 0.037 and 0.082 across models), tracking seasonal browsing patterns. ExitRates, TrafficType and the ProductRelated engagement variables come next. This convergence across four structurally different algorithms – two bagging/boosting tree ensembles, one classical boosting ensemble, and one neural network – provides strong triangulated evidence that PageValues, seasonality and exit behavior are real, model-independent behavioral signals of purchase intent rather than artifacts of any single algorithm's inductive bias.

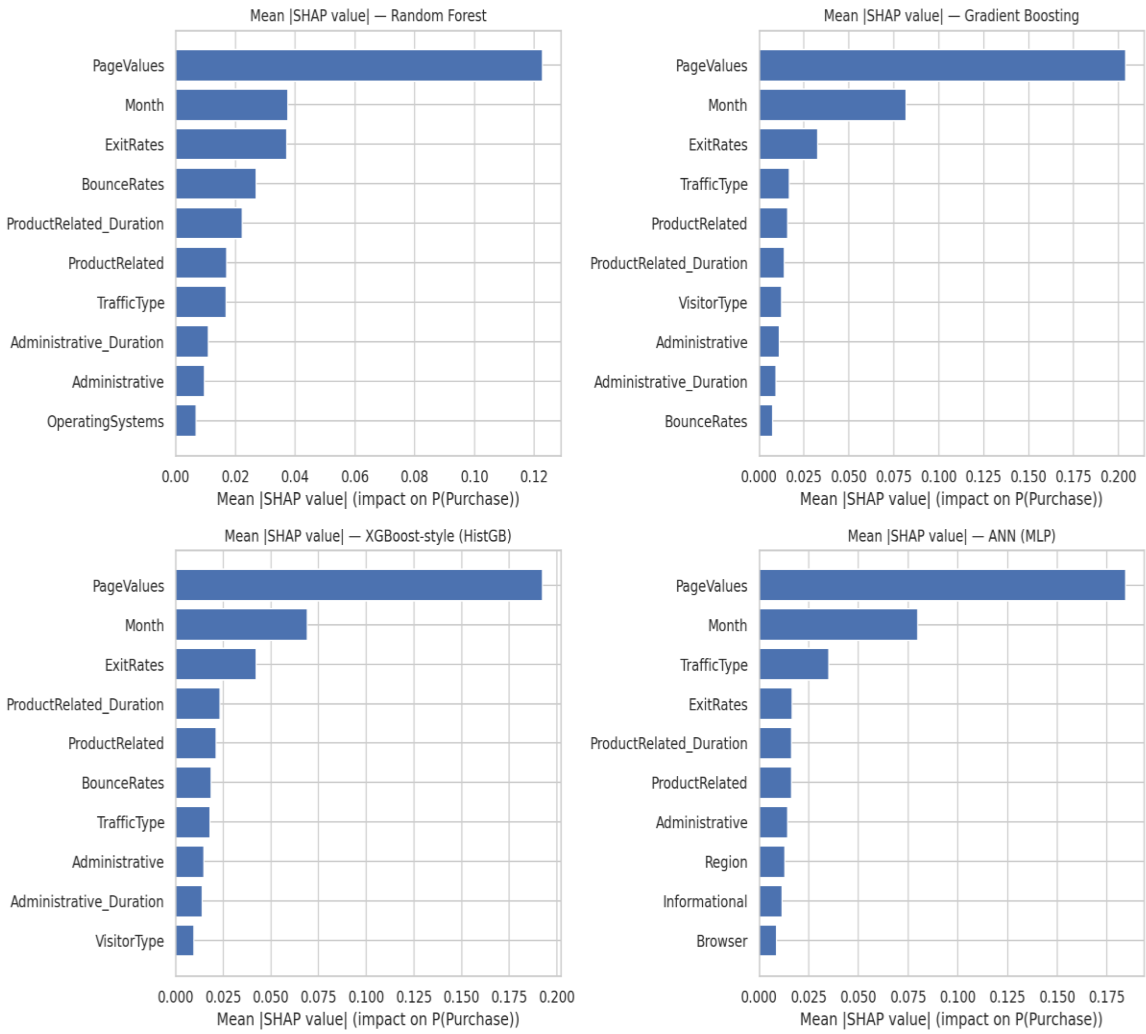


Figure 5: *Top-10 features by mean absolute SHAP value for each of the four models*

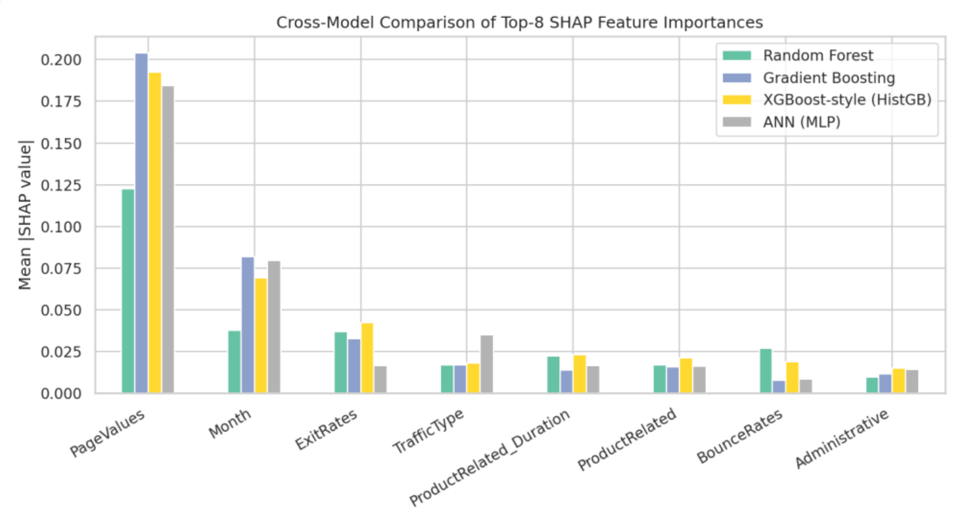


***Figure 6:** Cross-model comparison of the top-8 SHAP-ranked features*

Local Explanation: SHAP Summary and Dependence Plots

Figure 7 shows a SHAP summary (beeswarm-style) plot for the best performing model, Gradient Boosting, visualizing the magnitude and direction of the effect of each top-10 feature across the 120 explained sessions. Sessions with high PageValues (warm-colored points) are clustered at strongly positive SHAP values and thus increase the predicted purchase probability. Sessions with low PageValues (cool-colored points) are clustered around or below zero. ExitRates shows the opposite. Points with high exit rates and warm colors have negative SHAP values, which aligns with our intuitive understanding that a visitor who leaves the site quickly after viewing a page is less likely to convert to a purchase, other things being equal.

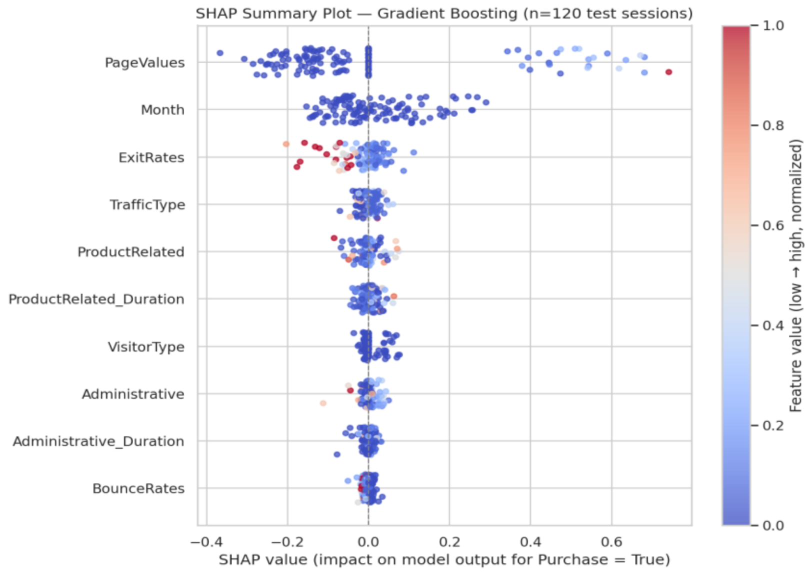


Figure 7. SHAP summary plot for the Gradient Boosting model

Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Figure 8 shows the SHAP dependence plot of PageValues and its own SHAP contribution for the Gradient Boosting model. The relationship is highly non-linear around zero – sessions with PageValues around zero or zero have consistently negative (but individually small) SHAP contributions, ranging from about -0.37 to 0 – and is strongly positive and steeply increasing once PageValues get above about 2-3, with the highest-value sessions in the sample getting SHAP contributions over 0.6. This threshold-like pattern indicates that even small amounts of high-value page engagement evidence significantly increase predicted purchase probability, a pattern which is directly actionable for real-time marketing triggers.

Discussion

Following the aforementioned research questions, the results of the simulation presented in the previous section are discussed in the following sections.

Model Performance and the Precision-Recall Trade-off

The two gradient boosting variants (classical and histogram-based) performed better than both Random Forest and the neural network on ROC-AUC and recall, as expected, confirming the general finding in tabular-data literature that sequential boosting typically outperforms bagging and standard feed-forward neural architectures on structured feature sets of this size and dimensionality. However, the performance gap across all four models is modest (ROC-AUC range: 0.909-0.928), suggesting that for this decision-support use case, model choice may reasonably be driven as much by deployment considerations (inference latency, retraining cadence, interpretability infrastructure) as by marginal accuracy gains. Random Forest has a higher precision and lower recall, making it better suited for situations where marketing actions are expensive and false positives must be minimized (e.g., high-value personal outreach). Gradient Boosting has a higher recall, favoring low-cost, high-volume interventions (e.g., automated on-site nudges) where the cost of missing a real buyer is greater than the cost of a wasted impression.

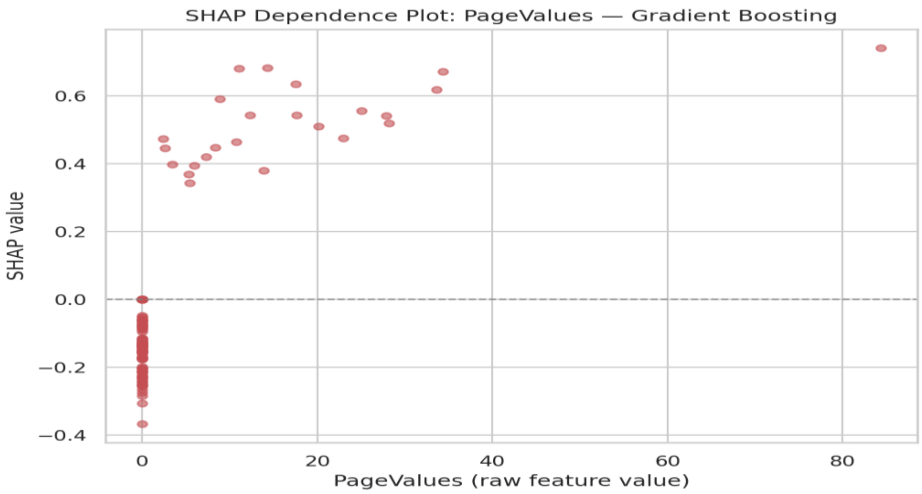


Figure 8. SHAP dependence plot for PageValues of Gradient Boosting model

Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Convergent Explanatory Drivers Across Model Families (RQ2)

The most actionable finding of the study is the consistent dominance of PageValues, Month and ExitRates across four structurally different algorithms. This ranking is stable regardless of what model a firm ultimately deploys, allowing marketing teams to develop one, sustainable explanation around these drivers, rather than re-deriving interpretation each time the underlying model is retrained or replaced. The steep, threshold-like SHAP dependence of PageValues (Figure 8) also suggests that this metric can be used as a simple near-real-time triage signal before a full model score is available, reserving model-based scoring for refining borderline cases.

Operationalizing SHAP as a Marketing Decision-Support Tool (RQ3)

Beyond global feature rankings, the session-level SHAP values computed in this work can also form the basis of personalized decision support: for any browsing session flagged as high-purchase-probability but with a negative ExitRates contribution, a marketing system could trigger a targeted incentive (e.g., a limited-time discount banner) that directly targets the factor suppressing conversion, rather than applying an undifferentiated, one-size-fits-all intervention. This local, session-level interpretability is exactly the thing a bare probability score from a black-box model cannot provide, and is the core practical contribution of pairing of high-performing ensemble/neural models with SHAP-based explanation in a production marketing analytics pipeline.

Theoretical Contribution

We contribute to the explainable-AI and digital-marketing-analytics literatures by showing, in a single controlled empirical pipeline, that feature-attribution rankings generated via a model-agnostic SHAP estimator are surprisingly stable across bagging, boosting, and neural architectures applied to the same prediction task—evidence that model-agnostic explainability does not have to come at the expense of consistency, and that firms can standardize on a single explanatory framework even as they experiment with or upgrade their underlying predictive model.

Conclusion

This study proposes and empirically tests an explainable machine learning framework for predicting online purchase intention by combining four supervised learning architectures with a robust, model agnostic SHAP explainability layer. Gradient boosting methods achieved the best predictive performance (ROC-AUC up to 0.928) whereas the analysis based on SHAP uncovered a surprisingly stable explanatory structure across all model families, with PageValues, seasonality (Month) and ExitRates being the most important, cross-model drivers of predicted purchase behavior. The proposed decision-support framework delivers a highly accurate predictive engine to a digital marketing team, with a transparent session-level explanatory layer capable of justifying and targeting specific interventions. This directly meets the core Responsible-AI mandate that predictive power in consequential customer-facing decision contexts must be paired with genuine interpretability, rather than used as an unexplained black box.

Volume 4 Issue No 3 (2025)

Online ISSN: 3006-2047
Print ISSN: 3006-2039

References

Bangash, J. I., Abdullah, A., & Khan, A. (2014). Issues and challenges in localization of wireless sensor networks. *Sci. Int (Lahore)*, 595-603.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.

Khalil, T., Bangash, J. I., & Adnan, A. (2016, August). Low level visio-temporal features for violence detection in cartoon videos. In *2016 sixth international conference on innovative computing technology (INTECH)* (pp. 320-325). IEEE.

Khattak, S. R., Saeed, I., & Tariq, B. (2018). Corporate sustainability practices and organizational economic performance. *Global Social Sciences Review*, *3*(4), 343-355.

Khattak, S. R., Batool, S., Saleem, Z., & Takrim, K. (2016). Effects of social media on teachers' performance: Evidence from Pakistan. *The dialogue*, *11*(1).

Khan, T. I., Saeed, I., & Khattak, S. R. (2018). *Impact of Time Pressure on Organizational Citizenship Behavior: Moderating Role of Conscientiousness. Global Social Sciences Review, 3 (3), 317 331*.

Khattak, S. R., Fayaz, M., Rahman, S. U., & Ullah, A. (2021). ORGANIZATIONAL AMBIDEXTERITY, ORGANIZATIONAL LEARNING CAPACITY, AND MARKET ORIENTATION: APOSSIBLE INDICATORS OF ORGANIZATIONAL PERFORMANCE. *Ilkogretim online*, *20*(4), 1953.

Batool, S., Khattak, S. R., & Saleem, Z. (2016). Professionalism: A Key Quality of Effective Manager. *Journal of Managerial Sciences*, *10*(1), 105.

Imran, M., Liu, X., Arif, M., Rahman, S. U., Manan, F., Khattak, S. R., & Wang, R. (2023). Sustainable corporate governance mediates between firm performance and corporate social responsibility using structural equation modelling. *Frontiers in Energy Research*, *11*, 1121853.

Mukhtar, M. A., Baloch, N. A., & Khattak, S. R. (2019). Dynamic capability & firm performance: Mediating role of learning orientation, organizational culture & corporate entrepreneurship: A case study of SME's of Pakistan. *J Manag Sci*, *13*, 119-128.

Volume 4 Issue No 3 (2025)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Khattak, S. R., Saeed, I., & Batool, S. (2019). The mediating effect of CSR on the relationship between authentic leadership and organization citizenship behavior. *Global Social Sciences Review*, *4*(2), 60-66.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

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.

Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11, 1–18.

Wu, X., & Zhao, J. (2020). Online purchase behavior prediction and analysis using ensemble learning. IEEE International Conference on Big Data Analytics.

Zaman, K., Bangash, J. I., Maghdid, S. S., Hassan, S., Afridi, H., & Zohaib, M. (2020, October). Analysis and classification of skin cancer images using convolutional neural network approach. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 1-8). IEEE.