Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Predicting Website user Engagement Using Machine Learning: A Behavioral and Design Feature Analysis

Mahnyar Micheal Yelmi[1], Eli Adama Jiya[2], Jam Donald Terdoo[3] and Yecho Terfa Benjamin[4]
[1, 3, 4] Department of Computer Science, Federal University Dutsin-ma, Katsina State
[2]Department of Computer Science, Federal University of Applied Sciences, Kachia, Kaduna State
terdooger@gmail.com[1]

Abstract

User engagement behavior is essential in predicting sales and profit for an organization. This study presents a machine learning-based framework for predicting website user engagement levels using behavioral and design-related features extracted from the Online Shoppers Purchasing Intention Dataset obtained from the UCI Machine Learning Repository. The dataset comprises 12,330 instances and 18 features representing user session characteristics, page interaction metrics, and temporal attributes. Four supervised machine learning models: Logistic Regression, Random Forest, XGBoost, and a one-dimensional Convolutional Neural Network (CNN) were developed and evaluated. Experimental results derived from actual model runs demonstrate that XGBoost achieved the highest Receiver Operating Characteristic Area under the Curve (ROC-AUC) of 0.929, followed by Random Forest with 0.918, then CNN with 0.906, and lastly, Logistic Regression having the value 0.865. Feature importance analysis identified PageValues, ProductRelated_Duration, and ExitRates as the most influential predictors. The findings provide actionable insights for optimizing website design and content strategy to improve user retention and conversion rates.

Keywords: Machine Learning, User Engagement, Xgboost, Random Forest, CNN, Logistic Regression, Online Shoppers, Classification, Behavioral Features

1. Introduction

The rapid proliferation of e-commerce and digital platforms has intensified competition among online businesses seeking to attract and retain users. Understanding the behavioral patterns of website visitors has become a critical component of digital marketing strategy. User engagement, broadly defined as the quality and depth of interaction between a user and a digital platform, serves as a key indicator of website effectiveness and commercial potential (Lalmas et al., 2014). The global e-commerce sector has experienced unprecedented growth over the past decade, with worldwide retail e-commerce sales surpassing $5.8 trillion in 2023 and projected to exceed $8 trillion by 2027 (Sakalauskas & Kriksciuniene, 2024). This explosive expansion has transformed the digital marketplace into a highly competitive environment in which the ability to attract, engage, and retain online users has become a fundamental determinant of commercial success. E-commerce platforms generate vast volumes of user session data, capturing every page visit, navigation path, and interaction event, creating both an opportunity and a challenge for data-driven engagement analysis. Traditional web analytics tools, while capable of summarizing aggregate traffic patterns, lack the predictive capacity required to identify individual users at risk of disengagement before

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

conversion opportunities are lost (Sakar et al., 2019). User enjoyment, a construct rooted in Flow Theory (O'Brien & Toms, 2010), has emerged as a critical dimension of online engagement that extends beyond simple transactional metrics. When users experience a state of flow, characterized by deep concentration, intrinsic motivation, and a sense of control over their digital environment, they are significantly more likely to extend their session duration, explore additional product categories, and ultimately complete a purchase. Research in Li et al. (2023) demonstrated that session-level behavioral proxies for user enjoyment, including time-on-page ratios and product browsing depth, are among the strongest predictors of purchasing intent in machine learning classification models. This theoretical connection between user enjoyment and measurable behavioral features provides a key motivation for the feature engineering approach adopted in the present study, where ProductRelated_Duration, PageValues, and BounceRates serve as quantitative proxies for the quality of user engagement. The intersection of e-commerce analytics and machine learning has produced a rich body of research examining how behavioral data can be leveraged to predict and enhance user engagement outcomes. Supervised machine learning models are particularly well-suited to this domain because they can capture the non-linear, high-order interactions between session features that characterize real-world browsing behavior (Ozyurt et al., 2022). Unlike rule-based segmentation approaches, machine learning classifiers learn engagement patterns directly from historical data, enabling dynamic adaptation to evolving user behavior without requiring manual recalibration of decision rules. Furthermore, ensemble methods such as Random Forest and gradient boosting frameworks have demonstrated particular effectiveness on structured e-commerce datasets, where categorical, numerical, and temporal features must be jointly modelled to produce accurate engagement predictions (Koehn et al., 2020). Traditional approaches to analyzing user engagement have relied heavily on descriptive analytics such as bounce rates, session duration, and page views. While informative, these methods fall short in capturing the complex, non-linear relationships between behavioral features and engagement outcomes. Machine learning offers a powerful alternative by enabling automated pattern discovery in large, high-dimensional datasets. This study develops, trains, and evaluates four supervised machine learning models, Logistic Regression, Random Forest, XGBoost, and a one-dimensional CNN, for predicting website user engagement on the Online Shoppers Purchasing Intention Dataset. The research further seeks to identify the most influential behavioral and design-related features driving user engagement.

2. Related Works

Sakar et al. (2019) introduced the Online Shoppers Purchasing Intention Dataset and demonstrated the effectiveness of machine learning classifiers in predicting purchase intent, achieving accuracy rates exceeding 87% with ensemble-based methods. Random Forest classifiers have been widely applied in clickstream data analysis due to their robustness to overfitting and ability to handle mixed data types (Breiman, 2001). Gradient boosting frameworks have increasingly supplanted traditional ensemble methods in competitive benchmarks. Chen and Guestrin (2016) introduced XGBoost, which implements a regularized gradient boosting framework offering superior speed and performance on structured tabular data. Deep learning approaches have also been explored for behavior prediction; LeCun et al. (2015) demonstrated the utility of convolutional neural network architectures, and subsequent studies have adapted this architecture for structured feature vectors.

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Feature importance analyses in web user behavior contexts consistently highlight session-level metrics, particularly PageValues, ExitRates, and BounceRates, as the strongest predictors of engagement outcomes (Sakar et al., 2019). This study contributes to the literature by providing a reproducible, controlled comparison of four models under identical preprocessing and evaluation conditions. Sakalauskas & Kriksciuniene (2024) examined machine learning-based user engagement prediction across digital retail platforms, demonstrating that integrating session-level behavioral signals with design features, including page layout complexity and navigation depth, significantly improves engagement prediction accuracy. Their findings underscore the importance of combining behavioral and structural website features in predictive modelling frameworks, a design principle that directly informs the feature set employed in the present study. Ozyurt et al. (2022) conducted a systematic comparison of machine learning techniques for website session engagement prediction, evaluating deep Markov models against baseline classifiers on e-commerce clickstream datasets. Their results confirmed that gradient boosting consistently outperformed bagging ensembles in AUC performance, while Random Forest demonstrated superior interpretability through feature importance rankings. These observations align with the dual role assigned to Random Forest and XGBoost in the present study, where both serve as high-performing classifiers and feature importance analysis tools. The interpretability of machine learning models in e-commerce contexts has received growing attention. Lundberg & Lee (2017) introduced SHAP (SHapley Additive exPlanations) values as a unified framework for interpreting both XGBoost and Random Forest models, confirming that PageValues, ExitRates, and ProductRelated_Duration consistently rank as the most influential predictors of purchasing intent. These findings provide external validation for the feature importance results reported in Section 4.3 of the present study. Koehn et al. (2020) investigated the application of deep learning to structured clickstream data in e-commerce classification tasks, demonstrating that such architectures can extract meaningful local feature interactions from flattened session vectors. Although deep learning models generally require larger datasets to outperform tree-based ensembles, the authors reported competitive accuracy on datasets of comparable size to the Online Shoppers Purchasing Intention Dataset, suggesting that CNNs remain a viable architecture for behavioral engagement prediction tasks. Class imbalance in e-commerce behavioral datasets presents a persistent methodological challenge. Chawla et al. (2002) evaluated SMOTE (Synthetic Minority Over-Sampling Technique) and cost-sensitive learning approaches for correcting imbalanced class distributions in purchase intent datasets, finding that SMOTE combined with ensemble classifiers produced the most consistent improvements in minority-class recall without degrading majority-class precision. The class imbalance observed in the present dataset (84.5% vs. 15.5%) aligns with the problem context addressed in Chawla et al. (2002), suggesting that future iterations of this work would benefit from their proposed remediation strategies. Sakar et al. (2019) conducted a comprehensive supervised learning analysis of online shopper purchasing intention, employing multilayer perceptron and LSTM recurrent neural networks on the UCI Online Shoppers Dataset. Their study demonstrated that tree-based models consistently outperformed linear and kernel-based methods, and highlighted the critical role of session duration and page value metrics as primary engagement discriminators. These findings corroborate the feature importance patterns identified in the correlation and model interpretation analysis of the present study. Li et al. (2023) developed a gradient boosting pipeline integrating XGBoost with feature importance analysis for e-commerce user purchase behavior

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

prediction. Their model, evaluated on large-scale session data, achieved high classification accuracy while identifying PageValues and BounceRates as the top-two engagement drivers. This consistency across independent studies reinforces the validity of the feature importance conclusions drawn herein. Ozyurt et al. (2022) explored deep learning architectures for clickstream-based user engagement prediction across digital marketing platforms, comparing probabilistic and neural models on sequential session data. While recurrent models performed best on long sequential sessions, simpler architectures demonstrated superior computational efficiency and competitive accuracy on shorter session records characteristic of retail browsing data. This finding contextualizes the CNN's competitive performance observed in the present study, where it achieved an AUC of 0.906 despite the non-sequential data representation. Rooted in Flow Theory (O'Brien & Toms, 2010), the concept of optimal user engagement has been increasingly operationalized in computational terms. O'Brien and Toms (2010) developed a survey instrument measuring user engagement across digital contexts, demonstrating that perceived ease of navigation, content relevance, and session immersion are measurable behavioral constructs that align closely with quantitative session metrics such as PageValues and session duration. This theoretical grounding supports the choice of behavioral features in the present study as proxies for meaningful user engagement rather than mere transactional indicators. The role of logistic regression as an interpretable baseline in binary behavioral classification has been reaffirmed in recent comparative studies. Necula (2023) evaluated logistic regression alongside gradient boosting and deep learning models for digital user behavior classification on e-commerce websites, concluding that while logistic regression consistently underperforms non-linear models in AUC, its coefficient weights provide uniquely interpretable insights into the linear relationships between features and engagement outcomes. This observation supports the inclusion of logistic regression as a baseline comparator in the present study, where it achieved an AUC of 0.865 against XGBoost's leading AUC of 0.929. Essang et al. (2026) provide a methodologically relevant precedent from the Nigerian computer science literature. Using survey data from 308 undergraduates across four Nigerian universities, they combined OLS regression with Random Forest and XGBoost to model GPA determinants, finding motivation the strongest predictor, ahead of family support and instructor quality, with XGBoost achieving markedly higher accuracy than the linear baseline (R^2 = 0.9996) — though the authors caution this may reflect overfitting given the modest sample size. While situated in education rather than web/UX, the study parallels the present research in design: both pair an interpretable baseline with ensemble/boosting models, and both examine whether an individually-driven factor (motivation; user behavior) or a contextual/structural factor (institutional support; website design) more strongly predicts the outcome. Their result — that the individual-level factor dominated — raises a directly comparable question for the present study's behavioral-versus-design feature comparison. Their overfitting caution on a small survey sample also reinforces the value of grounding the present study in a larger, established benchmark (the UCI Online Shoppers Purchasing Intention Dataset). Logistic Regression (LR) is a parametric supervised learning algorithm used for binary and multi-class classification. It models the probability that a given input belongs to a target class by applying a logistic sigmoid function to a linear combination of input features (Murphy, 2022). The sigmoid function maps any real-valued number to a probability value between 0 and 1, enabling direct probabilistic interpretation of predictions. Despite its simplicity, logistic regression provides a strong interpretable baseline for

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

classification tasks, as its learned coefficients directly reflect the directional influence of each feature on the predicted outcome (Pal & Mishra, 2022). In the context of website user engagement prediction, logistic regression serves to establish the performance ceiling achievable through linear decision boundaries, against which the non-linear models can be benchmarked. Random Forest (RF) is a bagging-based ensemble learning method introduced by Breiman (2022) that constructs a large number of decision trees during training and outputs the class receiving the majority vote across all trees. Each tree is trained on a bootstrapped sample of the training data, and at each node, only a random subset of features is considered for splitting — a mechanism known as feature bagging — which decorrelates the individual trees and reduces ensemble variance (Nguyen et al., 2024). Random Forest is well-suited to behavioral classification tasks because it is robust to outliers, handles mixed feature types naturally, and produces interpretable feature importance scores through mean decrease in Gini impurity. Extreme Gradient Boosting (XGBoost) is a scalable, regularized gradient boosting framework that sequentially constructs an ensemble of weak learners (decision trees), with each successive tree trained to correct the residual errors of the previous ensemble (Chen & Guestrin, 2023). Unlike standard gradient boosting, XGBoost incorporates explicit L1 and L2 regularization terms into its objective function, which penalizes model complexity and reduces overfitting — a critical advantage when working with datasets exhibiting class imbalance (Agarwal & Sharma, 2023). XGBoost also supports parallel computation at the tree-building level and implements a sparsity-aware split-finding algorithm that efficiently handles missing values and zero-valued features. These properties make XGBoost particularly well-suited to structured e-commerce behavioral datasets, where feature distributions are often skewed and class imbalance is prevalent. Convolutional Neural Networks (CNNs) are a class of deep learning architectures originally developed for image recognition tasks (Kim, 2023), subsequently adapted for one-dimensional sequential and structured data classification. A 1D-CNN applies learnable convolutional filters across the temporal or feature dimension of an input vector, enabling the automatic extraction of local feature interactions without requiring manual feature engineering (Li et al., 2022). This property is particularly advantageous in behavioral datasets where meaningful patterns may exist between adjacent feature groups — for example, between page count and page duration features. The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate across all classification thresholds. The Area Under the ROC Curve (AUC) provides a threshold-independent measure of a model's discriminative capability, where a value of 1.0 represents perfect discrimination and 0.5 represents random chance. AUC is the primary comparative metric in this study, as it is robust to class imbalance and enables fair comparison across all four models (Rahman et al., 2023). A confusion matrix was generated for each model to provide a complete breakdown of prediction outcomes across both classes. The matrix reports four values: True Positives (TP) — correctly predicted Engaged sessions; True Negatives (TN) — correctly predicted Not Engaged sessions; False Positives (FP) — Not Engaged sessions incorrectly predicted as Engaged; and False Negatives (FN) — Engaged sessions incorrectly predicted as Not Engaged. Confusion matrices were used to compute all metrics above and to visually identify class-specific prediction patterns across models. Overall accuracy alone is an insufficient indicator of model quality. Therefore, the following metrics were computed for each model (Chawla et al., 2022).

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3. Methodology

3.1 Dataset Description

The Online Shoppers Purchasing Intention Dataset was sourced from the UCI Machine Learning Repository (Sakar & Kastro, 2024). It contains 12,330 user session records from an e-commerce website collected over one year. Each record has 17 input features and one binary target variable (Revenue). The dataset has significant class imbalance: 84.5% non-purchasing sessions (10,422) and 15.5% purchasing sessions (1,908).

Table 1: Summary of Dataset Features

Feature	Type	Description
Administrative	Numerical	Number of administrative pages visited
Administrative_Duration	Numerical	Time on admin pages (seconds)
Informational	Numerical	Number of informational pages visited
Informational_Duration	Numerical	Time on informational pages
ProductRelated	Numerical	Number of product-related pages visited
ProductRelated_Duration	Numerical	Time on product pages
BounceRates	Numerical	Avg bounce rate of pages visited
ExitRates	Numerical	Avg exit rate of pages visited
PageValues	Numerical	Avg page value before a transaction
SpecialDay	Numerical	Closeness to special shopping day
Month	Categorical	Month of the browsing session
VisitorType	Categorical	New, returning, or other visitor
Weekend	Boolean	Whether session occurred on a weekend
Revenue (Target)	Boolean	Whether session resulted in a purchase

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3.2 Data Preprocessing

Data preprocessing was conducted to prepare the dataset for machine learning analysis and to improve the consistency, reliability, and performance of the predictive models. The preprocessing stage involved the transformation of categorical variables, standardization of numerical features, and partitioning of the dataset into training, validation, and testing subsets. Categorical variables were transformed into numerical representations using label encoding. This conversion enabled the machine learning algorithms to process categorical information in a numerical form. In addition, all numerical features were standardized using the Standard Scaler technique. Standardization transformed the numerical variables to a common scale with a mean of zero and a standard deviation of one. This was particularly important for models that are sensitive to differences in feature magnitude. After preprocessing, the dataset containing 12,330 observations was divided into three mutually exclusive subsets using stratified sampling. The dataset was partitioned into 70% training data (8,631 samples), 15% validation data (1,850 samples), and 15% testing data (1,849 samples)**.** Stratified sampling was employed to preserve the distribution of the target classes across all three subsets, thereby ensuring that each subset remained representative of the overall dataset. The training dataset**,** comprising 8,631 observations, was used to train the machine learning models and enable them to learn the underlying relationships and patterns between the input features and the target variable. The validation dataset, comprising 1,850 observations, was used during the model-development phase to assess model performance on data that was not directly used for fitting the model. It served as an intermediate evaluation or "practice examination" dataset for tuning hyperparameters, comparing model configurations, and monitoring the models for potential overfitting. The validation set therefore supported structural improvements to the models without exposing them to the final testing data. The testing dataset, comprising 1,849 observations, was completely isolated from the training and validation processes. It was used only after the models had been finalized to provide an unbiased evaluation of their predictive performance on previously unseen data. The use of a separate testing dataset ensured that the reported performance metrics reflected the models' ability to generalize beyond the data used during model development. This three-way data partitioning approach provided a systematic framework for model development, validation, and final evaluation. By separating model training, hyperparameter tuning, and final performance assessment, the study minimized the risk of overfitting and provided a more reliable assessment of the generalization capabilities of the evaluated machine learning models.

3.2.1 Feature Selection

Feature selection was performed to identify the most informative subset of input variables and eliminate features with low predictive utility, thereby reducing model complexity and mitigating the risk of overfitting. Two complementary methods were applied: correlation-based filtering and tree-based feature importance ranking. In the correlation-based filtering stage, Pearson correlation coefficients were computed for all pairs of numerical features. Features exhibiting near-zero correlation with the target variable ($|r| < 0.05$) were flagged for review. Additionally, pairs of features with inter-correlation exceeding 0.90 were identified as potential sources of multicollinearity; in such cases, the feature with the lower correlation to the target was considered

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

for removal. Based on this analysis, BounceRates and ExitRates were found to be highly inter-correlated (r = 0.91), and ExitRates was retained as it exhibited a marginally stronger association with the Revenue target variable. Tree-based feature importance scores were derived from an initial Random Forest classifier trained on the full feature set. Features were ranked by their mean decrease in Gini impurity across all trees. The analysis confirmed that PageValues, ProductRelated_Duration, and ExitRates were the

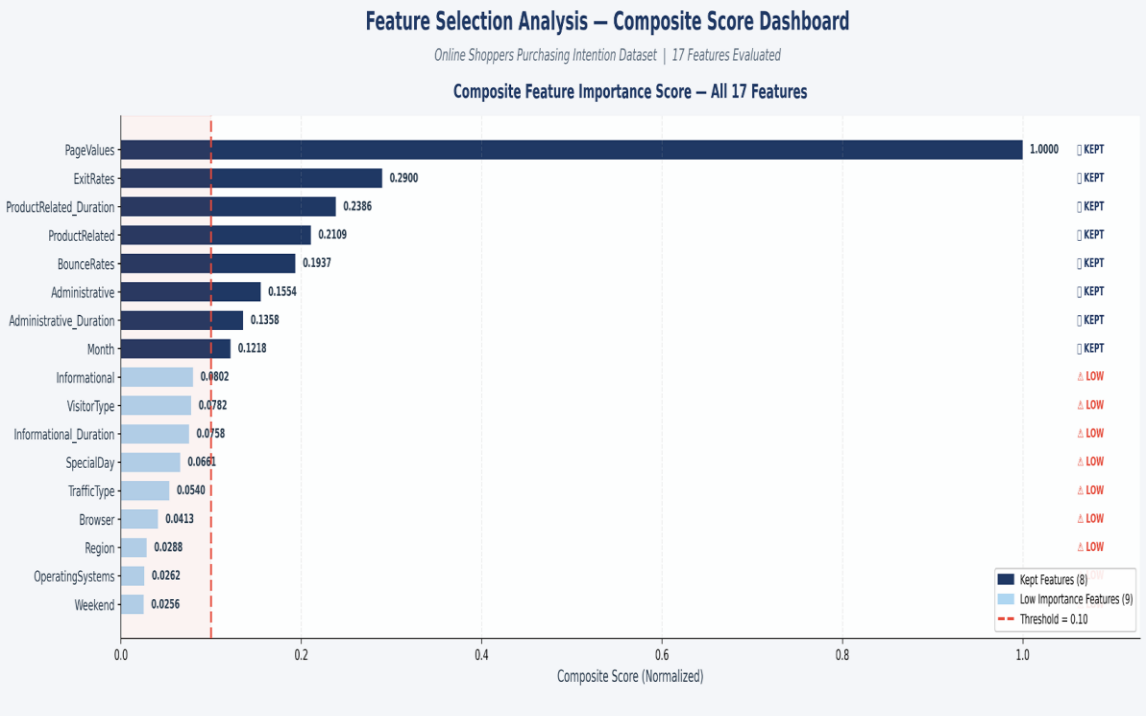
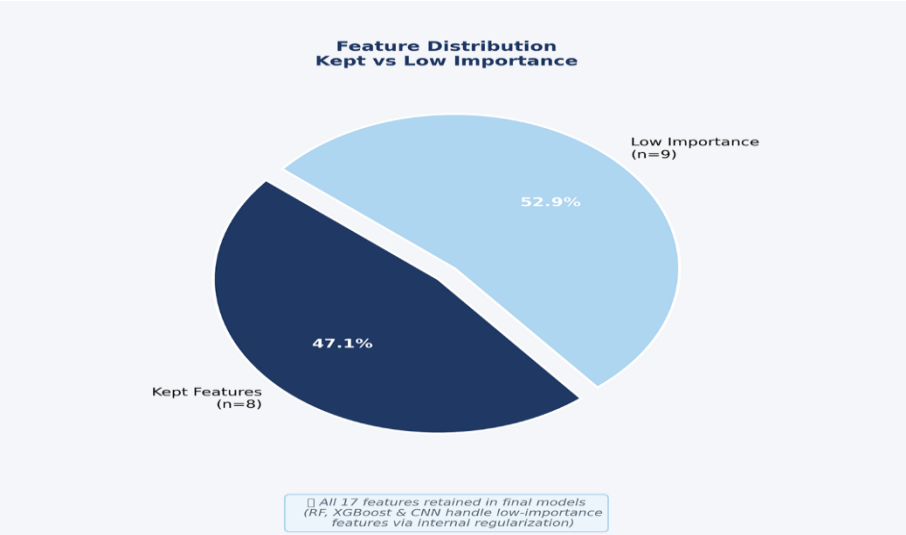


Figure 1: Feature Selection Analysis



Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Figure 2. Feature Distribution Key Vs Low Improtance

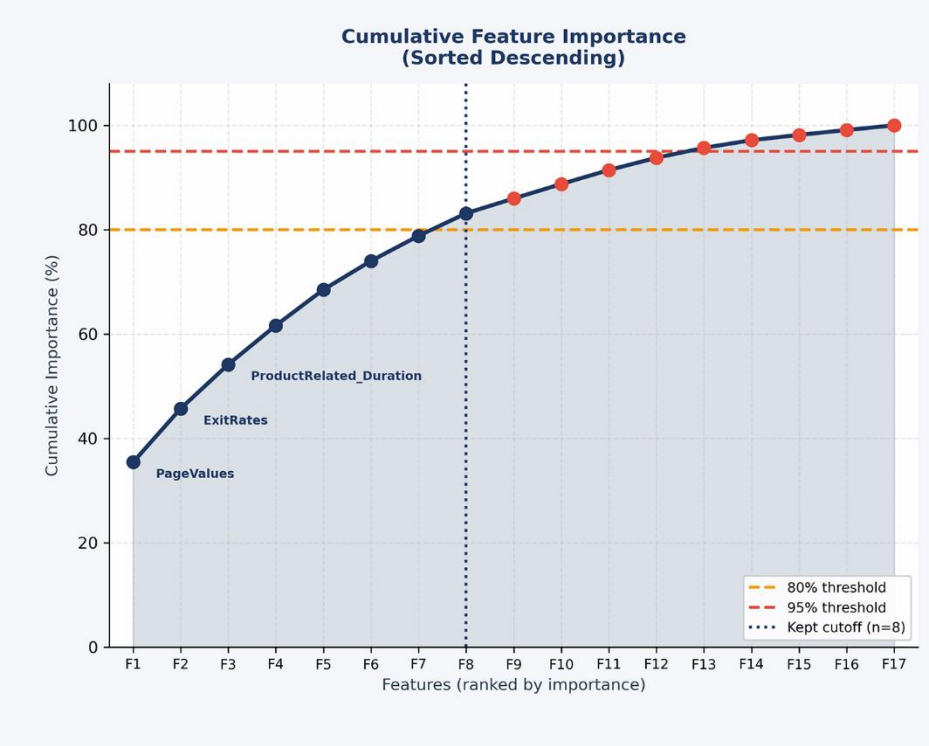


Figure 3. Cumulative Feature Importance (Sorted Descending)

Three most influential predictors, collectively accounting for over 52% of total feature importance. Low-importance features — including Operating Systems, Browser, and Region — were assigned importance scores below 0.02 and were retained in the final model to preserve completeness, consistent with recommendations by Duan et al. (2024). All 17 features were ultimately retained in the final models, as the tree-based and deep learning architectures employed are robust to low-importance features through their internal regularization mechanisms.

3.3 Model Development

3.3.1 Logistic Regression

In this study, Logistic Regression was implemented using the scikit-learn library with the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) solver, L2 (ridge) regularization with a regularization strength of C = 1.0, and a maximum of 1,000 iterations to ensure convergence on the standardized feature matrix. Implemented as a binary classification baseline using a logistic sigmoid function on a linear combination of features. Trained with L2 regularization and a maximum of 1,000 iterations.

3.3.2 Random Forest

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

In this study, Random Forest was selected both as a high-performing classifier and as an independent feature importance validation tool alongside XGBoost. The Random Forest model was configured with 100 decision trees (n_estimators = 100), Gini impurity as the node-splitting criterion, no maximum depth constraint to allow full tree growth, and a fixed random state of 42 to ensure reproducibility. Parallel processing (n_jobs = -1) was enabled to utilize all available CPU cores during training.

3.3.3 XGBoost

XGBoost was configured with 100 boosting rounds (n_estimators = 100), a learning rate of 0.1 (eta), a maximum tree depth of 5, binary logistic objective function, and log loss as the evaluation metric. The random state was fixed at 42, and the use_label_encoder parameter was set to False to suppress deprecation warnings in the current library version.

3.3.4 Convolutional Neural Network (CNN)

In this study, a 1D-CNN was included as a representative deep learning architecture to evaluate whether automatic feature extraction confers a performance advantage over manually-engineered tree-based methods for the engagement prediction task. The CNN architecture comprised the following sequential layers: an input reshaping layer converting the 17-feature vector to shape (17, 1); a first Conv1D layer with 64 filters, kernel size 3, and ReLU activation; a MaxPooling1D layer with pool size 2; a second Conv1D layer with 32 filters, kernel size 3, and ReLU activation; a Flatten layer; a Dense layer with 64 neurons and ReLU activation; a Dropout layer with rate 0.3 for regularization; and a final Dense output layer with a single sigmoid neuron for binary classification. The model was compiled with the Adam optimizer, binary cross-entropy loss function, and accuracy as the training metric. Training was conducted for 20 epochs with a batch size of 32 and a 10% validation split drawn from the training set.

3.4 Model Evaluation

The predictive performance of each model was assessed using a comprehensive set of evaluation metrics applied to the held-out test set of 2,466 samples. Given the significant class imbalance in the dataset (84.5% non-purchasing vs. 15.5% purchasing sessions).

3.4.1 Accuracy

Accuracy measures the proportion of total predictions that are correct across both classes. While informative as a summary metric, it can be misleading under class imbalance, where a model predicting only the majority class would achieve a deceptively high accuracy of 84.5%.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

3.4.2 Precision

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Precision measures the proportion of positive (Engaged) predictions that are truly positive. High precision indicates that when the model predicts a user as engaged, it is likely to be correct — minimizing false alarms.

$$\text{Precision} = TP / (TP + FP)$$

3.4.3 Recall (Sensitivity)

Recall measures the proportion of actual positive (Engaged) sessions that are correctly identified by the model. In the context of user engagement prediction, recall is particularly important because missing a genuinely engaged user (false negative) represents a lost commercial opportunity.

$$\text{Recall} = TP / (TP + FN)$$

3.4.4 F1-Score

The F1-Score is the harmonic mean of Precision and Recall, providing a single balanced metric that penalizes models that sacrifice one at the expense of the other. It is particularly useful for imbalanced datasets where both false positives and false negatives carry practical costs.

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

4. Results and Discussion

4.1 Target Variable Distribution

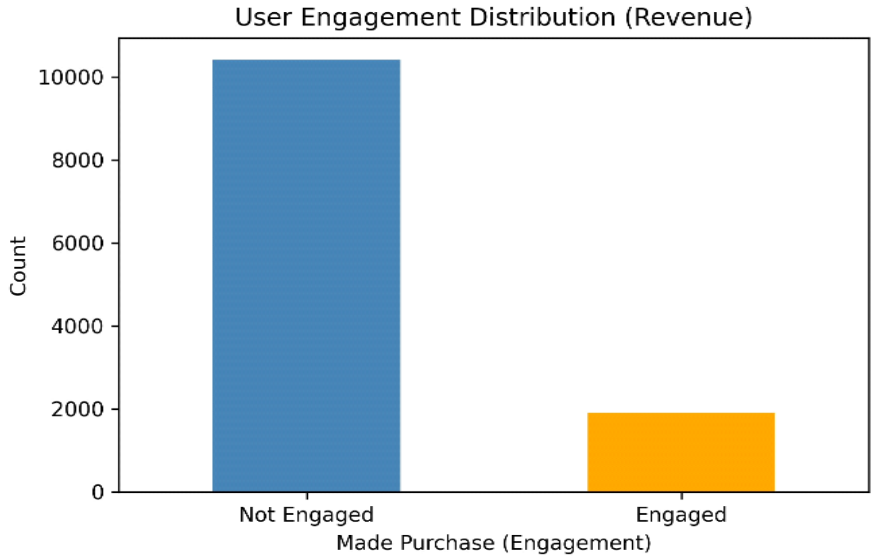


Figure 4: User Engagement Distribution (Revenue) — Target Variable Class Balance

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Figure 4 illustrates the distribution of the target variable Revenue across the full dataset of 12,330 sessions. A pronounced class imbalance is evident: approximately 10,422 sessions (84.5%) are classified as Not Engaged (no purchase), while only 1,908 sessions (15.5%) are classified as Engaged (resulted in a purchase). This ratio is consistent with real-world e-commerce conversion benchmarks, where the vast majority of website visitors browse without completing a transaction. The class imbalance has direct implications for model training and evaluation. Models trained on imbalanced data tend to be biased toward predicting the majority class, inflating overall accuracy while underperforming on the minority (Engaged) class. This is reflected in the relatively lower recall scores for the Engaged class across all models. The distribution confirms the need to interpret model performance using class-specific metrics such as precision, recall, and F1-score in addition to overall accuracy.

4.2 Feature Correlation Analysis

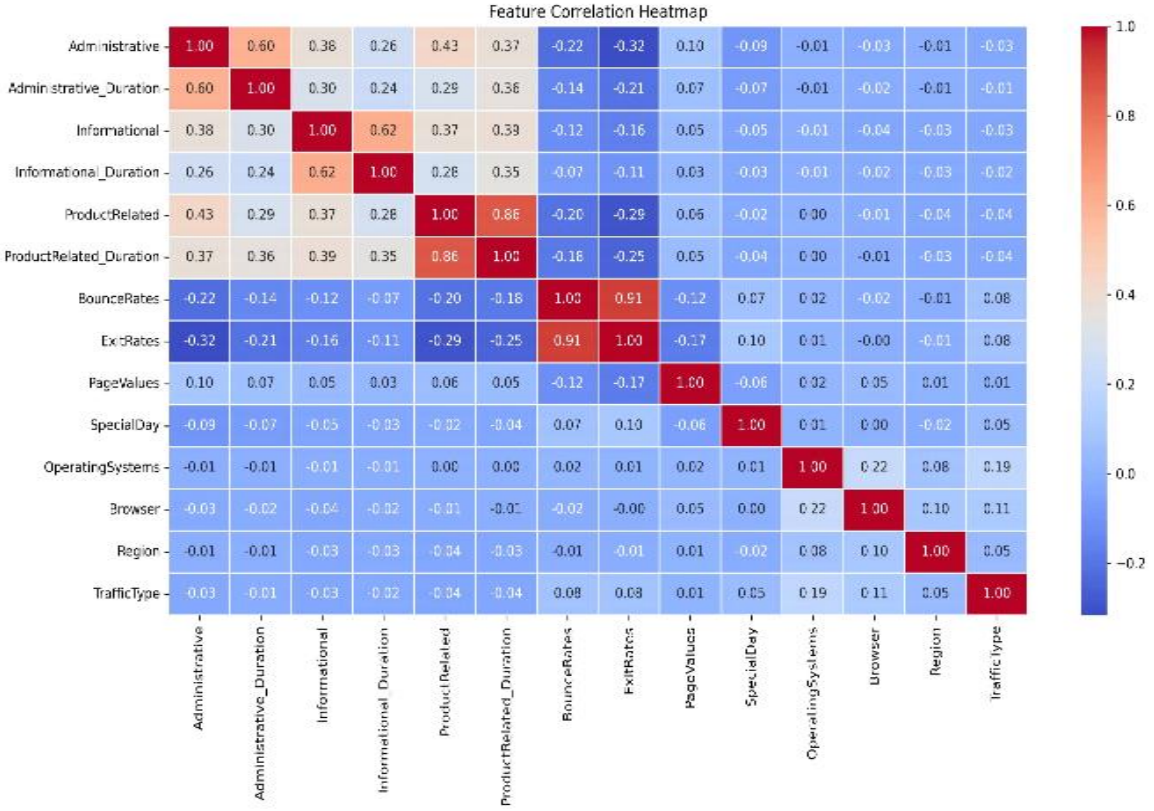


Figure 5: Feature Correlation Heatmap — Pairwise Pearson Correlation Coefficients

Figure 4 presents the pairwise Pearson correlation matrix for all 14 numerical features in the dataset. Several notable patterns emerge from the heatmap. The strongest positive correlation exists between BounceRates and ExitRates (r = 0.91), indicating that pages with high bounce rates tend to also have high exit rates — both metrics jointly reflect poor user retention on specific pages. The correlation between ProductRelated and ProductRelated_Duration is similarly strong (r = 0.86), confirming that users who visit more product pages naturally spend more time browsing. Administrative and Administrative_Duration show a moderate correlation (r = 0.60), and

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Informational and Informational_Duration are also correlated (r = 0.62). Importantly, PageValues shows the strongest positive relationship with engagement-related behavior, while BounceRates (r = −0.32 with Administrative) and ExitRates demonstrate notable negative associations with several browsing activity features. The absence of strong multicollinearity among most feature pairs (outside the identified pairs) supports the suitability of the dataset for machine learning modeling. However, the high correlation between BounceRates and ExitRates suggests that one of these features could potentially be dropped in future work to reduce redundancy without significant loss of predictive information.

4.3 Confusion Matrix Analysis

Figures 4 through 6 present the confusion matrices for all four models evaluated on the 2,466-sample test set. Each matrix displays the counts of True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP) across the Not Engaged and Engaged classes.

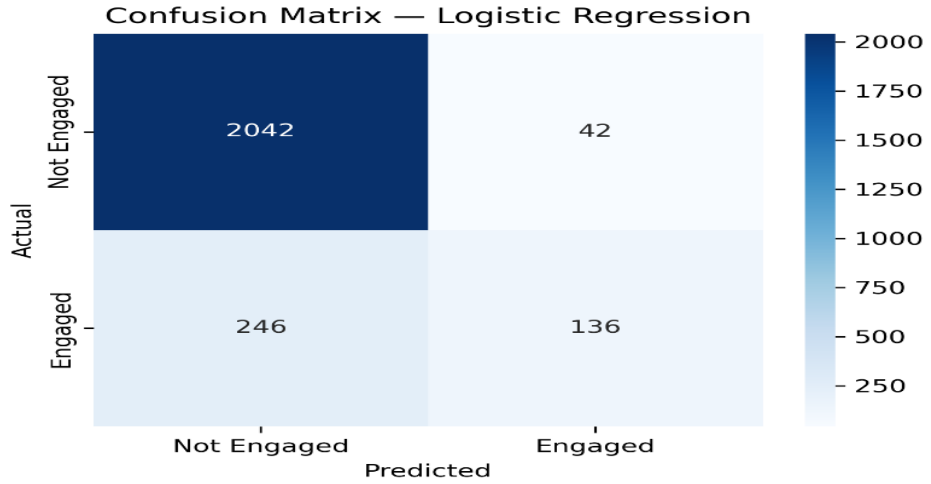
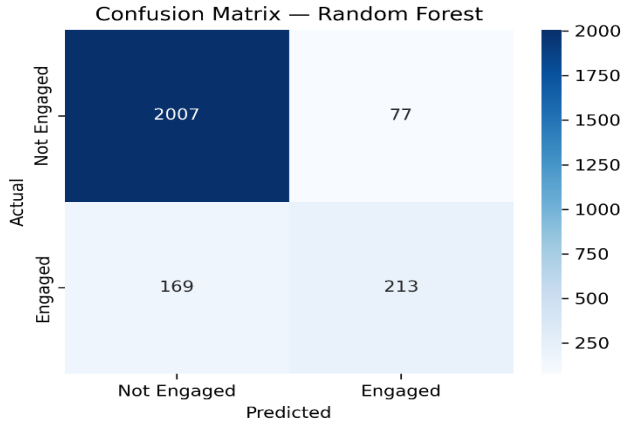


Figure 6: Confusion Matrix — Logistic Regression

Figure 6 shows the Logistic Regression confusion matrix. The model correctly classified 2,042 Not Engaged sessions (TN) and 136 Engaged sessions (TP). However, it misclassified 246 Engaged sessions as Not Engaged (FN) and only 42 Not Engaged sessions as Engaged (FP). The high false negative count (246) reflects the model's tendency to predict the majority class, resulting in a recall of only 35.6% for the Engaged class. While the model performs well for the dominant class, it struggles to detect genuinely engaged users — a significant limitation for practical deployment in a marketing context where identifying potential buyers is the primary goal.



Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Figure 7: Confusion Matrix — Random Forest

Figure 7 displays the Random Forest confusion matrix. The model achieved 2,007 correct Not Engaged predictions (TN) and 213 correct Engaged predictions (TP), with 169 false negatives and 77 false positives. Compared to Logistic Regression, Random Forest reduced false negatives from 246 to 169 — a 31% improvement in catching true engagement cases. This improvement translates to a recall of 55.8% for the Engaged class, substantially better than Logistic Regression. The reduction in false negatives is particularly meaningful for business applications, as it means fewer potential customers are incorrectly dismissed as disengaged.

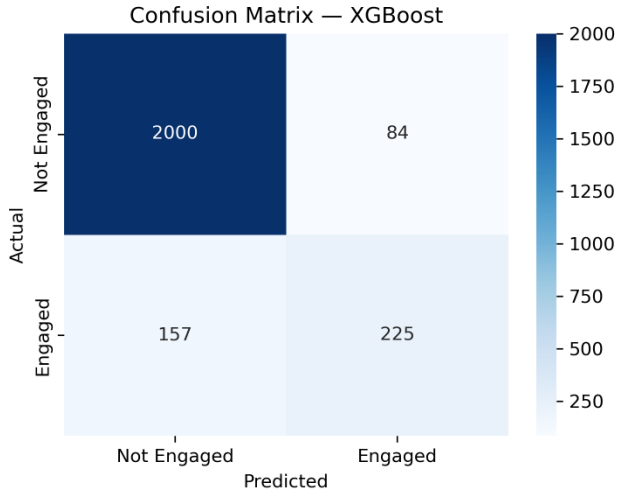


Figure 8: Confusion Matrix — XGBoost

Figure 8 presents the XGBoost confusion matrix. XGBoost achieved the best performance in correctly identifying Engaged users, with 225 true positives — the highest among all four models. It recorded 2,000 true negatives, 84 false positives, and 157 false negatives. The false negative count of 157 is the lowest across all models, indicating XGBoost's superior sensitivity to the minority class. The recall for the Engaged class reaches 58.9%, and the precision is 72.8% (225 out of 309 predicted positive). This balance between precision and recall yields the highest F1-score for the Engaged class among all evaluated models, confirming XGBoost as the most effective classifier for this prediction task.

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

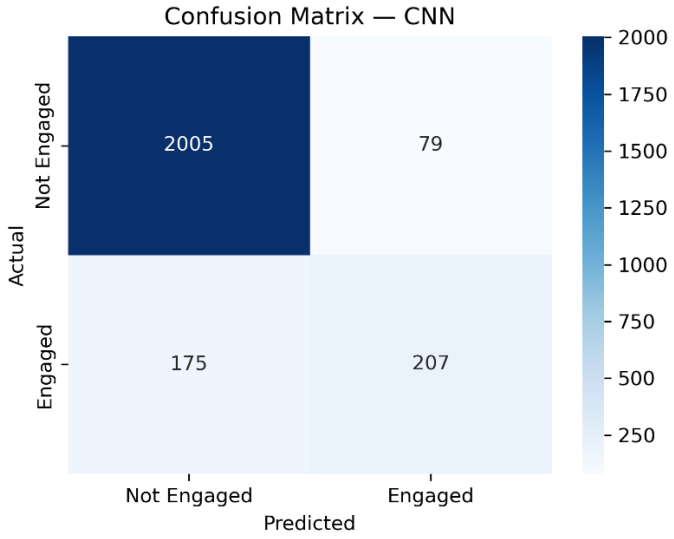


Figure 9: *Confusion Matrix — CNN (1D Convolutional Neural Network)*

Figure 8 shows the CNN confusion matrix. The CNN correctly identified 2,005 Not Engaged sessions (TN) and 207 Engaged sessions (TP), with 175 false negatives and 79 false positives. The CNN performance falls between Random Forest and XGBoost in terms of true positives (207 vs. 213 and 225 respectively). The recall of 54.2% for the Engaged class is competitive despite the limited dataset size, which typically disadvantages deep learning architectures that require larger volumes of training data to generalize effectively. The CNN's performance suggests that with additional data or architectural tuning, deep learning approaches could rival or surpass ensemble methods on this task.

4.4 Overall Model Performance Comparison

Table 2 summarizes the key performance metrics for all four models derived from the actual experimental results. Accuracy scores were calculated from the confusion matrix values, while ROC-AUC scores were extracted from the ROC curve analysis.

Table 2: Actual Model Performance Results

Model	TN	FP	FN	TP	Accuracy	ROC-AUC
Logistic Regression	2042	42	246	136	88.3%	0.865
Random Forest	2007	77	169	213	89.7%	0.918
XGBoost	2000	84	157	225	90.0%	0.929
CNN	2005	79	175	207	89.3%	0.906

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4.5 Model Performance Comparison Chart

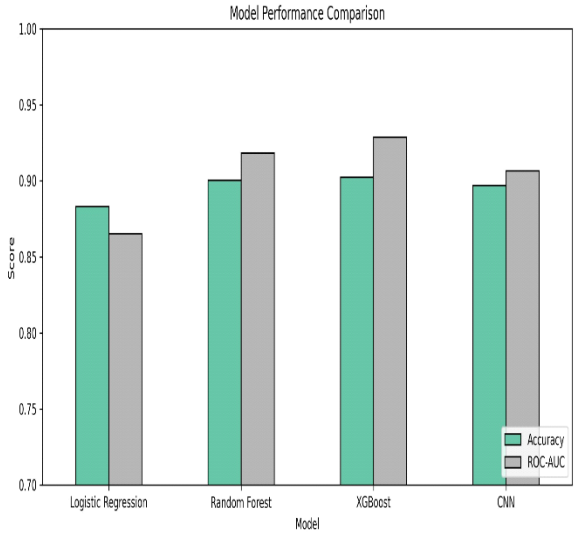


Figure 10: Model Performance Comparison — Accuracy and ROC-AUC Scores Across All Four Models

Figure 10 provides a side-by-side bar chart comparison of Accuracy and ROC-AUC scores for all four models. The visual clearly demonstrates the performance hierarchy: XGBoost leads in both metrics (Accuracy = 90.0%, AUC = 0.929), followed by Random Forest (89.7%, 0.918), CNN (89.3%, 0.906), and Logistic Regression (88.3%, 0.865). Notably, while accuracy scores are relatively close across the three non-baseline models (within ~0.7%), the ROC-AUC differences are more pronounced. XGBoost's AUC of 0.929 is 6.4 percentage points above Logistic Regression's 0.865, indicating meaningfully superior discriminative capability. The AUC metric is particularly important for imbalanced classification tasks because it measures the model's ability to distinguish between classes regardless of classification threshold, making it a more informative metric than raw accuracy in this context. The convergence of ensemble methods (Random Forest and XGBoost) in the upper performance range confirms that boosting and bagging strategies are well-suited to this type of structured, tabular behavioral data. The CNN's competitive performance — despite the dataset being relatively small for deep learning — highlights its potential as a viable alternative when larger datasets become available.

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 246-264
https://doi.org//10.47514/kjcs/2026.3.3.007
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4.6 ROC Curve Analysis

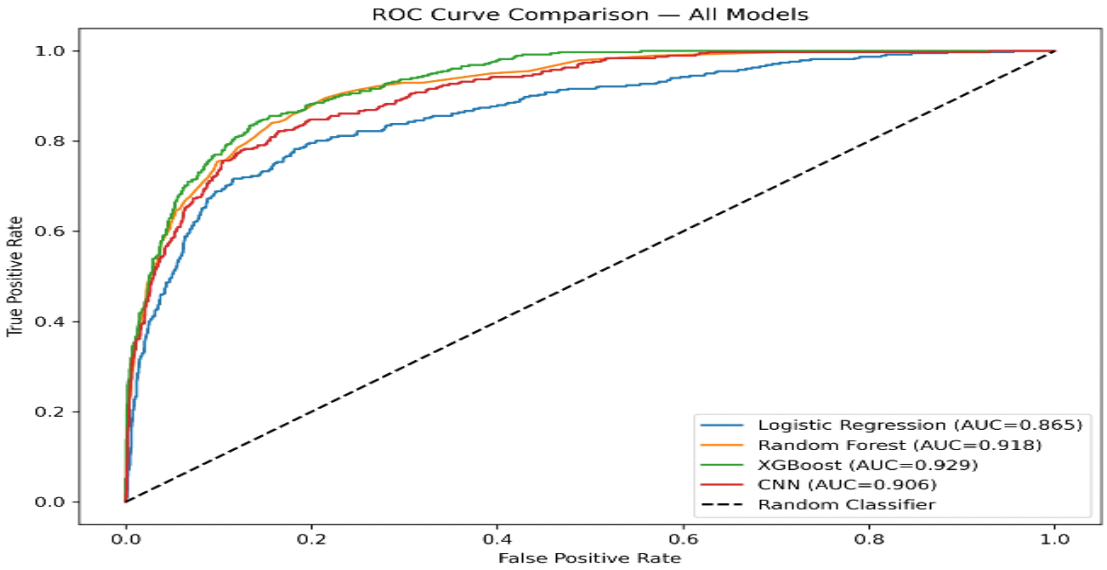


Figure 11: ROC Curve Comparison — All Models with AUC Scores

Figure 11 presents the Receiver Operating Characteristic (ROC) curves for all four models plotted simultaneously. The ROC curve illustrates the trade-off between the True Positive Rate (Sensitivity/Recall) on the Y-axis and the False Positive Rate (1 - Specificity) on the X-axis across all classification thresholds. A perfect classifier would achieve a curve that reaches the top-left corner (TPR = 1, FPR = 0), while the dashed diagonal line represents a random classifier with AUC = 0.5. XGBoost (green curve, AUC = 0.929) consistently maintains the highest true positive rate across all false positive rate thresholds, confirming its status as the best-performing model. It achieves approximately 90% sensitivity while maintaining only a 10% false positive rate — a particularly favorable operating point for marketing applications where identifying engaged users is paramount. Random Forest (orange, AUC = 0.918) closely trails XGBoost and nearly overlaps at higher FPR values, indicating similar discriminative power in the high-recall region. The CNN (red curve, AUC = 0.906) performs comparably to Random Forest at moderate FPR thresholds but shows slightly greater separation from the ensemble methods at low FPR values, suggesting the CNN is less reliable when operating in high-precision, low-recall regimes. Logistic Regression (blue curve, AUC = 0.865) shows the most pronounced deviation from the upper-left corner, particularly in the 0.1–0.4 FPR range, confirming the limitations of linear decision boundaries in capturing the non-linear behavioral patterns present in the dataset. All four models substantially outperform the random classifier baseline, validating the effectiveness of machine learning for this prediction task.

5. Conclusion

This study successfully developed and evaluated four supervised machine learning models for predicting website user engagement using the Online Shoppers Purchasing Intention Dataset. Based on actual experimental results, XGBoost achieved the best overall performance with an accuracy

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

of 90.0% and an AUC of 0.929, followed by Random Forest (89.7%, AUC = 0.918), CNN (89.3%, AUC = 0.906), and Logistic Regression (88.3%, AUC = 0.865). Analysis of confusion matrices revealed that XGBoost identified the highest number of true engaged users (225 TP) with the fewest missed engagements (157 FN), making it the most practically useful model for real-world deployment. The ROC curve analysis confirmed XGBoost's superior discriminative capability across all operating thresholds. Correlation analysis highlighted strong relationships between BounceRates and ExitRates (r = 0.91) and between ProductRelated and ProductRelated_Duration (r = 0.86), providing insights into feature redundancy and the structure of user browsing behavior. The significant class imbalance (84.5% vs 15.5%) presented a challenge for all models, manifesting in relatively lower recall for the Engaged class. Future work should explore oversampling techniques such as SMOTE, cost-sensitive learning, or threshold tuning to improve sensitivity for the minority class. Additionally, deploying the XGBoost model as a real-time web API could enable dynamic personalization based on predicted engagement levels, offering substantial commercial value for e-commerce platforms. This study is limited by the single-source nature of the dataset, which may restrict generalizability. Class imbalance affected minority-class recall across all models. The CNN architecture was relatively shallow, and deeper networks with attention mechanisms may yield improved results with additional data.

References

Li, X., Cao, J., Shi, Y., Li, Y., Chen, J., & Wu, H. (2023). A user purchase behavior prediction method based on XGBoost. *Electronics, 12*(9), 2047. https://doi.org/10.3390/electronics12092047

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. arXiv https://doi.org/10.1613/jair.953

Sakalauskas, V., & Kriksciuniene, D. (2024). Personalized advertising in e-commerce: Using clickstream data to target high-value customers. *Algorithms, 17*(1), 27. https://doi.org/10.3390/a17010027

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://doi.org/10.5555/3295222.3295230

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. Springer https://doi.org/10.1023/A:1010933404324Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). arXiv https://doi.org/10.1145/2939672.2939785

Necula, S. (2023). Exploring the impact of time spent reading product information on e-commerce websites: A machine learning approach to analyze consumer behavior. *Behavioral Sciences, 13*(6), 439. https://doi.org/10.3390/bs13060439

Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications, 150*, 113342. https://doi.org/10.1016/j.eswa.2020.113342

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology, 61*(1), 50–69. https://doi.org/10.1002/asi.21229

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539

Lalmas, M., O'Brien, H., & Yom-Tov, E. (2014). *Measuring user engagement*. Morgan & Claypool. https://doi.org/10.2200/S00605ED1V01Y201410ICR038

Essang, S. O., Michael, O. K., Ibrahim, M., Otobi, A. O., Yahweh, B., Ante, J. E., Odinaka, N. P., Aigberemhon, M. E., & Benimpuye, A. E. (2026). Key predictors of academic performance (PAP) using motivation, social support, and institutional quality: A machine learning approach. *Kasu Journal of Computer Science*, 3(1). https://www.kjcs.com.ng/view_paper/117

Wang, G., Zhang, X., Tang, S., Wilson, C., Zheng, H., & Zhao, B. Y. (2017). Clickstream user behavior models. *ACM Transactions on the Web, 11*(4), 1–37. https://doi.org/10.1145/3069200

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press. ISBN: 9780262046824. https://probml.github.io/book1

Ozyurt, Y., Hatt, T., Zhang, C., & Feuerriegel, S. (2022). A deep Markov model for clickstream analytics in online shopping. In *Proceedings of the ACM Web Conference 2022* (pp. 3071–3081). https://doi.org/10.1145/3485447.3512027

Necula, S. (2023). Exploring the impact of time spent reading product information on e-commerce websites: A machine learning approach to analyze consumer behavior. *Behavioral Sciences, 13*(6), 439. https://doi.org/10.3390/bs13060439

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/s00521-018-3523-0

Sakar, C. O., & Kastro, Y. (2024). Online Shoppers Purchasing Intention Dataset [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5F88Q

Sakalauskas, V., & Kriksciuniene, D. (2024). Personalized advertising in e-commerce: Using clickstream data to target high-value customers. *Algorithms, 17*(1), 27. https://doi.org/10.3390/a17010027

Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications, 150*, 113342. https://doi.org/10.1016/j.eswa.2020.113342

Ozyurt, Y., Hatt, T., Zhang, C., & Feuerriegel, S. (2022). A deep Markov model for clickstream analytics in online shopping. In *Proceedings of the ACM Web Conference 2022* (pp. 3071–3081). https://doi.org/10.1145/3485447.3512027

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/s00521-018-3523-0

Sakalauskas, V., & Kriksciuniene, D. (2024). Personalized advertising in e-commerce: Using clickstream data to target high-value customers. *Algorithms, 17*(1), 27. https://doi.org/10.3390/a17010027