**IJCCS (Indonesian Journal of Computing and Cybernetics Systems)**
Vol.x, No.x, July xxxx, pp. 1~5
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.xxxx


# Comparison of Random Forest and XGBoost for Impulsive Buying Prediction

**Adinda Bintang Oktavia*[1], Andy Prasetyo Utomo[2], Soni Adiyono[3]**
[1,2,3] Department of Information System, Faculty of Engineering, Muria Kudus University, Kudus, Indonesia
e-mail: ***[1]adindabintang1023@gmail.com**, [2]andy.prasetyo@umk.ac.id, [3]soni.adiyono@umk.ac.id

***Abstrak***

*Penelitian ini bertujuan membandingkan kinerja algoritma Random Forest dan XGBoost dalam memprediksi impulsive buying pada pengguna mobile marketplace menggunakan data historis 2022 dan data kuesioner 2026. Tahapan penelitian meliputi prapemrosesan, pembentukan fitur dan target, pembagian data secara stratified, penanganan ketidakseimbangan kelas menggunakan SMOTE dalam pipeline, serta pemodelan dan evaluasi menggunakan Accuracy, Precision, Recall, F1-Score, dan Area Under the Curve (AUC). GridSearchCV dengan five-fold cross-validation digunakan untuk pemilihan model, sedangkan SHAP digunakan untuk menginterpretasikan kontribusi fitur terhadap prediksi. Pada data historis 2022, Random Forest memperoleh F1-Score sebesar 0,8129 dan AUC sebesar 0,8260, lebih tinggi daripada XGBoost dengan F1-Score 0,7922 dan AUC 0,7933. Pada data kuesioner 2026, kedua algoritma menghasilkan F1-Score sebesar 0,9279, sementara AUC Random Forest sebesar 0,8724 dan XGBoost sebesar 0,8432. Analisis SHAP menunjukkan bahwa Hedonic Browsing merupakan fitur dengan kontribusi prediktif terbesar pada kedua dataset. Secara keseluruhan, Random Forest menunjukkan kinerja validasi yang lebih tinggi berdasarkan F1-Score pada kedua dataset.*

***Kata kunci****— Impulsive Buying, Mobile Marketplace, Random Forest, XGBoost, SHAP*

***Abstract***

*This study compares Random Forest and XGBoost for predicting impulsive buying using historical data from 2022 and questionnaire data from 2026. The datasets were analyzed independently using an 80:20 stratified split, with SMOTE applied only within training folds. Model selection used GridSearchCV with five-fold cross-validation, and test performance was assessed using Accuracy, Precision, Recall, F1-score, and AUC. On the 2022 test set, Random Forest outperformed XGBoost across all reported metrics. On the 2026 test set, both models obtained the same Accuracy, Precision, Recall, and F1-score, while Random Forest achieved a higher AUC. Five-fold cross-validation showed a higher mean F1-score and lower F1-score standard deviation for Random Forest on both datasets. SHAP identified Hedonic Browsing as the feature with the greatest relative predictive contribution in both datasets; this finding does not imply causality.*

***Keywords—*** *Impulsive Buying, Mobile Marketplace, Random Forest, XGBoost, SHAP*

*Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012*

## 1. INTRODUCTION

The development of mobile marketplaces has made online shopping more accessible and convenient. However, easy access to products, promotional stimuli, and engaging browsing experiences may encourage impulsive buying, which refers to purchasing decisions made without prior planning. Previous studies have associated online impulsive buying with personality traits, emotions, and time pressure [1], as well as hedonic and utilitarian shopping values [2]. These behavioral factors provide a basis for developing predictive models that can identify patterns associated with impulsive buying.

Machine learning has increasingly been applied to consumer behavior prediction. Chen [3] investigated machine learning approaches for predicting online shopping behavior, while Budiman and Wijaya [4] introduced a dataset of mobile app impulsive buying that includes Physical Environment, Social Environment, Time Perspective, Hedonic Browsing, Utilitarian Browsing, and Impulsive Buying. More directly, Villdina et al. [5] applied Random Forest to predict impulsive buying among Indonesian marketplace users, and Nurfadhilah and Azhari [6] used Random Forest to classify impulsive purchasing tendencies among Generation Z consumers in e-commerce. These studies demonstrate the potential of machine learning for impulsive buying prediction, but they do not establish whether Random Forest or Extreme Gradient Boosting (XGBoost) performs better when evaluated using the same behavioral constructs across datasets collected in different periods.

Comparing predictive algorithms is important because their performance may differ depending on data characteristics and class distributions. Insani and Denny [7] compared Random Forest, XGBoost, and LightGBM for predicting potential buyers in Indonesian e-commerce, although potential-buyer prediction differs from questionnaire-based impulsive buying classification. Furthermore, Pathak et al. [8] applied explainable machine learning and SHapley Additive exPlanations (SHAP) to examine predictive features associated with online impulse-buying tendency. Cui [9] also investigated machine learning models for distinguishing impulsive and planned purchasing behavior in e-commerce. Nevertheless, further evidence is needed on the comparative performance of Random Forest and XGBoost for impulsive buying classification using consistent behavioral constructs, alongside feature-level interpretation through SHAP.

To address this gap, this study compares Random Forest and XGBoost using historical data collected in 2022 and newly collected questionnaire data from 2026. Both datasets use the same measurement constructs and indicators, but are analyzed independently to evaluate each algorithm under the respective data distribution. The primary objective is to determine and compare the predictive performance of the two algorithms within each dataset, rather than to test temporal changes in consumer behavior or cross-period model generalization. SHAP is additionally used to identify the features that contribute most to model predictions in each dataset. This study therefore contributes empirical evidence on algorithm performance and predictive feature contributions across two independently analyzed datasets collected in different periods, without interpreting the results as proof of causal relationships or behavioral changes over time.

Model performance is evaluated using Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC). Five-fold cross-validation is used to assess performance across training folds, while a held-out test set provides the final evaluation. SHAP analysis is then conducted to interpret the contribution of individual features to the predictions. Through this approach, the study provides a comparative assessment of Random Forest and XGBoost for impulsive buying classification and examines the consistency of predictive feature contributions across the two datasets.

## 2. METHODS

### *2.1 Research Data*

This study employed a quantitative predictive approach to compare the performance of Random Forest and Extreme Gradient Boosting (XGBoost) in predicting impulsive buying behavior among mobile marketplace users. Two independent datasets were analyzed: a historical dataset collected in 2022 and a primary dataset collected in 2026. Both datasets used the same measurement constructs, questionnaire indicators, and five-point Likert scale. However, because the datasets were collected in different periods using different sampling techniques, they were analyzed separately rather than treated as a single longitudinal dataset.

The historical dataset was obtained from the open-access dataset published by Budiman and Wijaya [4], entitled Mobile App Impulsive Buying: A Situational Factors Dataset Analysis. The original data comprised 628 Indonesian consumers collected between April and October 2022 using purposive sampling. Eligible respondents were aged 18–40 years, used mobile devices, and had regularly purchased products through online marketplace applications during the preceding six months. The dataset included 369 female respondents (58.8%) and 259 male respondents (41.2%). Regarding age, 154 respondents (24.5%) were aged 26–30 years, 386 (61.5%) were aged 31–35 years, and 88 (14.0%) were aged 36–40 years. The questionnaire measured Physical Environment (PE), Social Environment (SE), Time Perspective (TP), Hedonic Browsing (HB), Utilitarian Browsing (UB), and Impulsive Buying (IB) using a five-point Likert scale.

The primary dataset was collected in September 2026 through an online questionnaire distributed using Google Forms and shared through WhatsApp and other social media platforms. Convenience sampling was used to recruit respondents who met the study's eligibility criteria: being aged 18–40 years and having purchased products through a marketplace within the preceding six months. Participation was voluntary, and informed consent was obtained before respondents completed the questionnaire. The questionnaire also collected demographic information, including gender, age, shopping frequency, and marketplace usage. After data preparation, 628 responses were retained for analysis.

To maintain measurement consistency, the primary dataset used the same constructs, questionnaire indicators, and five-point Likert scale as the historical dataset. Nevertheless, the two datasets differ in their collection periods and sampling techniques. Therefore, identical sample sizes and measurement instruments do not establish that the datasets represent statistically equivalent populations. Each dataset was consequently processed and evaluated independently. The comparison focuses on model performance and predictive feature contributions within each dataset, rather than on longitudinal changes or direct generalization between the two respondent groups.

### *2.2 Research Variables and Feature Construction*

The study used five predictor variables and one target variable. The predictors represented situational and browsing-related factors associated with impulsive buying behavior. The target variable represented the respondent's impulsive buying classification. The indicators are presented in Table 1.

Table 1 Research indicators

| Construct | Indicators | Indicator Summary |
|---|---|---|
| Physical Environment | PE1-PE4 | Ease of use, attractive appearance, visual comfort, and application feature accessibility. |
| Social Environment | SE1-SE3 | Views, encouragement, and recommendations from the social environment regarding marketplace use. |
| Time Perspective | TP1-TP3 | Ease of use based on time and place, and convenience of use. |
| Utilitarian Browsing | UB1-UB5 | Price and quality search, product information, comparison, and shopping efficiency. |
| Hedonic Browsing | HB1-HB3 | Browsing for entertainment, leisure, and enjoyment. |
| Impulsive Buying | IB1-IB3 | Urge to purchase additional items beyond the planned purchase. |

For each construct, a composite score was calculated using the arithmetic mean of its corresponding questionnaire indicators. The mean score of the Impulsive Buying construct was calculated using Equation (1).

$$IB_{mean}\,\frac{IB_1+IB_2+IB_3}{3} \tag{1}$$

The same mean-based approach was applied to the indicators of the five predictor constructs. The resulting composite scores were used as model inputs, while the Impulsive Buying composite score was converted into a binary target for classification.

*2.3 Target Variable Definition and Class Distribution*

The target variable was constructed from the mean score of the three Impulsive Buying indicators (IB1, IB2, and IB3). Considering that the questionnaire used a five-point Likert scale, with a score of 3 representing the midpoint, this study adopted a threshold of 3 to classify respondents into two categories. Specifically, respondents with an Impulsive Buying mean score greater than 3 were assigned to Target 1, whereas respondents with a mean score equal to or below 3 were assigned to Target 0. The target classification rule is expressed in Equation (2).

$$y = \begin{cases} 1, \text{ if } IB_{mean} > 3 \\ 0, \; if \; IB_{mean} \leq 3 \end{cases} \tag{2}$$

This threshold was used as an operational rule for the binary classification task. It should not be interpreted as a universally established psychological boundary between individuals who do and do not engage in impulsive buying.

Applying this rule resulted in different class distributions across the two datasets. In the historical dataset collected in 2022, Target 0 comprised 237 respondents (37.7%), whereas Target 1 comprised 391 respondents (62.3%). In the primary dataset collected in 2026, Target 0 comprised 81 respondents (12.9%), while Target 1 comprised 547 respondents (87.1%).

These distributions indicate that the primary dataset had a substantially higher proportion of respondents classified as Target 1 than the historical dataset. This difference was retained rather than artificially adjusted during target construction because the classification rule was applied consistently to both datasets. The resulting imbalance was addressed during model training using Synthetic Minority Over-sampling Technique (SMOTE), as described in Section 2.5. The original class distributions were retained for evaluating the models on their respective test sets.

*2.4 Data Preparation and Training–Testing Procedure.*

The two datasets were prepared and modeled independently. For each dataset, the predictor matrix consisted of the five composite predictor scores (PE, SE, TP, HB, and UB), while the binary Impulsive Buying classification was used as the target variable.

Each dataset was divided into training and testing subsets using an 80:20 ratio. Stratified sampling was applied during the split to preserve the original target-class proportions as closely as possible in both subsets. A random state of 42 was used to support reproducibility. With 628 observations in each dataset, the split produced 502 training observations and 126 testing observations per dataset.

The training subset was used for hyperparameter optimization, cross-validation, and model development. The testing subset was held out from these training procedures and used for the final evaluation of predictive performance. This separation ensured that the reported test-set metrics were calculated on observations not used to fit the models during training.

To address class imbalance during model fitting, SMOTE was incorporated into an imbalanced-learn pipeline together with the classification algorithm. Oversampling was performed only within the training data available to each fitting operation. During cross-validation, SMOTE was applied separately to the training portion of each fold, without resampling that fold's validation portion. The held-out test set was not oversampled and retained its original class distribution.

This procedure prevented synthetic observations from being generated using information from the test set and reduced the risk of data leakage. It also ensured that model evaluation reflected performance on the original, unseen observations rather than on an artificially balanced test set.

*2.5 Classification Algorithms and Hyperparameter Optimization*

Two tree-based ensemble learning algorithms, Random Forest and Extreme Gradient Boosting (XGBoost), were employed to predict the binary Impulsive Buying target. Both algorithms were trained and optimized independently for each dataset.

Hyperparameter optimization was conducted using GridSearchCV with five-fold cross-validation, using the F1-score as the selection metric. The F1-score was selected because it balances precision and recall, which are relevant to classification tasks involving imbalanced target classes. The candidate hyperparameter values evaluated for each algorithm are presented in Table 2.

Table 2 Hyperparameter Search Space

| Algorithm | Hyperparameter | Candidate Values |
|---|---|---|
| Random Forest | n_estimators | 50,100 |
| | max_depth | None, 5, 10 |
| | min_samples_split | 2, 5 |
| XGBoost | n_estimators | 50, 100 |
| | max_depth | 3, 5 |
| | learning_rate | 0.01, 0.1 |

For each dataset, GridSearchCV evaluated the candidate configurations using five-fold cross-validation on the training subset. The configuration with the highest mean cross-validation F1-score was selected as the best configuration for each algorithm. The resulting best hyperparameter configurations for both datasets are summarized in Table 3.

Table 3 Best Hyperparameter Configurations

| Dataset | Algorithm | Best Hyperparameter Configuration |
|---|---|---|
| 2022 | Random Forest | max_depth=10, min_samples_split=2, n_estimators=100 |
| | XGBoost | learning_rate=0.1, max_depth=5, n_estimators=100 |
| 2026 | Random Forest | max_depth=None, min_samples_split=5, n_estimators=50 |
| | XGBoost | learning_rate=0.1, max_depth=3, n_estimators=100 |

To address class imbalance during model training, SMOTE was incorporated into an imbalanced-learn pipeline with each classification algorithm. Oversampling was performed only on the training portion of each cross-validation fold. Neither the validation portion of each fold nor the held-out test set was oversampled.

In addition to the cross-validation conducted during GridSearchCV, a separate five-fold cross-validation procedure was performed on the training subset using each selected estimator. The mean and standard deviation of the fold-level scores were used to summarize performance across folds. These additional cross-validation results were reported separately from the best mean F1-score obtained through GridSearchCV and the final evaluation metrics calculated on the held-out test set.

The final models were evaluated on the held-out test subset using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). For SHAP interpretation, the selected algorithm was determined by comparing the best mean cross-validation F1-scores obtained through GridSearchCV. The model with the higher score was selected for each dataset, with Random Forest selected in the event of a tie.

### *2.6 Model Evaluation*

The optimized Random Forest and XGBoost models were evaluated on the held-out test subsets using five metrics: accuracy, precision, recall, F1-score, and ROC-AUC [10]. Accuracy measures the proportion of correct predictions and is calculated using Equation (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Precision measures the proportion of correct positive predictions, as shown in Equation (4).

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

Recall measures the proportion of actual positive observations correctly identified, as presented in Equation (5).

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

The F1-score represents the harmonic mean of precision and recall, as expressed in Equation (6).

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

ROC-AUC was calculated using prediction scores for the positive class to assess the model's ability to distinguish between the two target classes. In Equations (3)–(6), (TP), (TN), (FP), and (FN) denote true positives, true negatives, false positives, and false negatives, respectively. All metrics were calculated on the original held-out test subsets and reported separately from the cross-validation results.

### *2.7 Explainable Machine Learning Using SHAP*

SHapley Additive exPlanations (SHAP) was used to interpret feature contributions to the selected model's predictions. SHAP assigns feature attribution values based on Shapley values to explain model predictions. For each dataset, the model with the higher best mean five-fold cross-validation F1-score from GridSearchCV was selected for interpretation; Random Forest was selected in the event of a tie.

SHAP values were calculated using TreeExplainer on the held-out test features. For binary classification, the SHAP values for Target 1 were used when class-specific outputs were available. Global feature importance was calculated as the mean absolute SHAP value, as expressed in Equation (7).

$$I_j = \frac{1}{n} \sum_{i=1}^{n} |\emptyset_{ij}| \quad (7)$$

Where (I_j) denotes the global importance of feature (j), (n) is the number of test observations, and (\phi_{ij}) represents the SHAP value of feature (j) for observation (i). The resulting values were used to rank the five predictors in each dataset. SHAP results were interpreted as contributions to model predictions, not as evidence of causal relationships.

## 3. RESULTS AND DISCUSSION

### *3.1 Target Distribution*

The target variable was constructed using the same threshold for both datasets, where Target 1 represents an Impulsive Buying mean score greater than 3, while Target 0 represents a score less than or equal to 3. The resulting class distributions are presented in Table 4.

Table 4 Original target distribution

| Dataset | Target 0 | Target 1 |
|---|---|---|
| 2022 Historical Data | 237(37.7%) | 391 (62.3%) |
| 2026 Questionnaire Data | 81(12.9%) | 547 (87.1%) |

The target distributions differed between the 2022 historical dataset and the 2026 questionnaire dataset. Target 1 accounted for 62.3% of the historical dataset and 87.1% of the questionnaire dataset. However, these differences should not be interpreted as direct evidence of changes in consumer behavior over time, because the datasets were collected in different periods and involved different sampling techniques. The class distributions are therefore important contextual factors when interpreting model performance, particularly for the 2026 dataset, in which Target 1 was dominant. Accordingly, model evaluation considered Precision, Recall, F1-Score, and AUC in addition to Accuracy.

### *3.2 Model Evaluation on the 2022 Historical Dataset*

The performance of Random Forest (RF) and XGBoost on the 2022 historical dataset was evaluated using the held-out testing set. The evaluation results are presented in Table 5.

Table 5 Test-set performance on the 2022 historical dataset

| Metric | Random Forest | XGBoost |
|---|---|---|
| Accuracy | 0.7698 | 0.7460 |
| Precision | 0.8182 | 0.8026 |
| Recall | 0.8077 | 0.7821 |
| F1-Score | 0.8129 | 0.7922 |
| AUC | 0.8260 | 0.7933 |

Random Forest achieved higher values than XGBoost across all five test metrics. RF obtained an F1-score of 0.8129 and an AUC of 0.8260, compared with 0.7922 and 0.7933 for XGBoost. The comparison is shown in Figure 1.

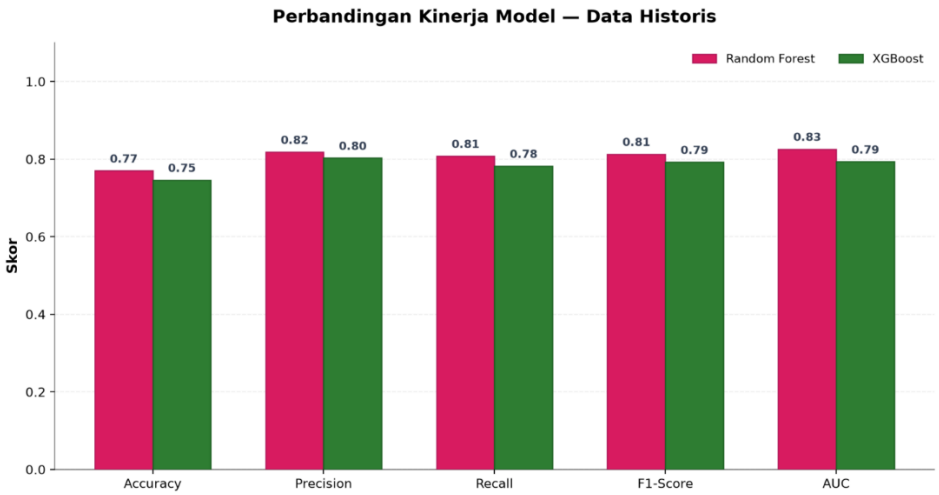



Figure 1 Model performance comparison on the 2022 historical data

The ROC curves show each model's ability to discriminate between the two classes across classification thresholds. RF achieved a higher AUC (0.8260) than XGBoost (0.7933). The model selected for SHAP interpretation was determined using the best mean cross-validation F1-score from GridSearchCV; RF was selected for the 2022 dataset. The ROC curves are shown in Figure 2.

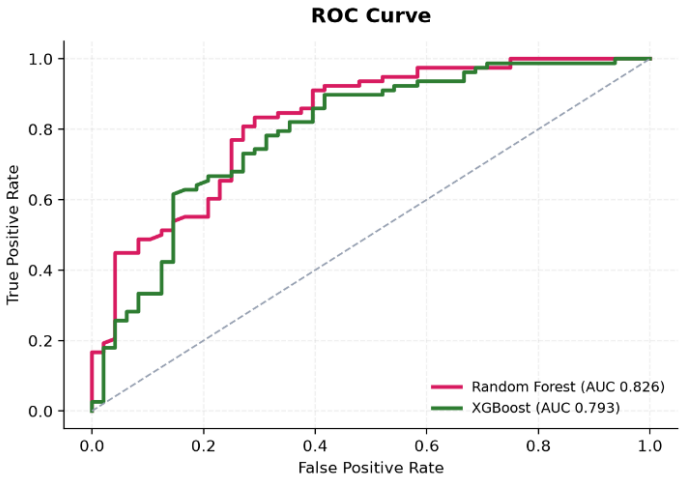



Figure 2 ROC curves of Random Forest and XGBoost on the 2022 historical data

*3.3 Model Evaluation on the 2026 Questionnaire Dataset*

The test-set performance of Random Forest and XGBoost on the 2026 questionnaire dataset is presented in Table 6.

Table 6 Model performance on the 2026 questionnaire dataset

| Metric | Random Forest | XGBoost |
|---|---|---|
| Accuracy | 0.8730 | 0.8730 |
| Precision | 0.9196 | 0.9196 |
| Recall | 0.9364 | 0.9364 |
| F1-Score | 0.9279 | 0.9279 |
| AUC | 0.8724 | 0.8432 |

On the 2026 test set, RF and XGBoost obtained identical Accuracy (0.8730), Precision (0.9196), Recall (0.9364), and F1-score (0.9279). RF achieved a higher AUC (0.8724) than XGBoost (0.8432), indicating better discrimination based on predicted probabilities in this evaluation. The model performance comparison is shown in Figure 3.

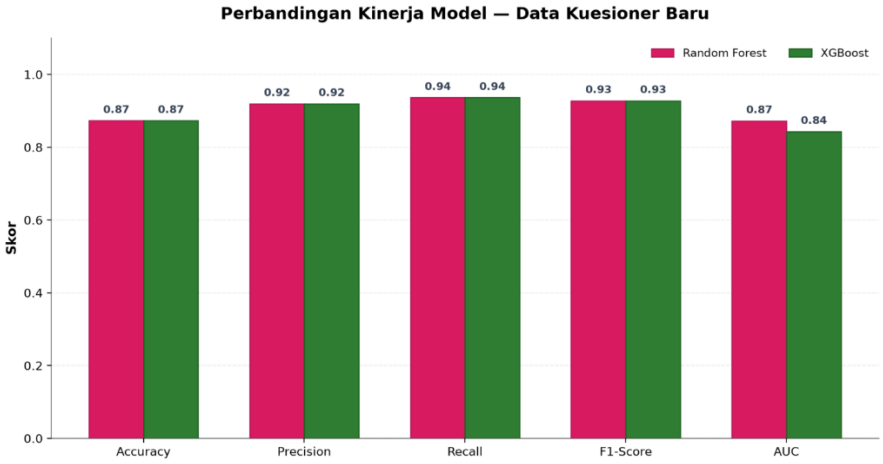



Figure 3 Model performance comparison on the 2026 questionnaire data

The ROC curves of both models are presented in Figure 4.

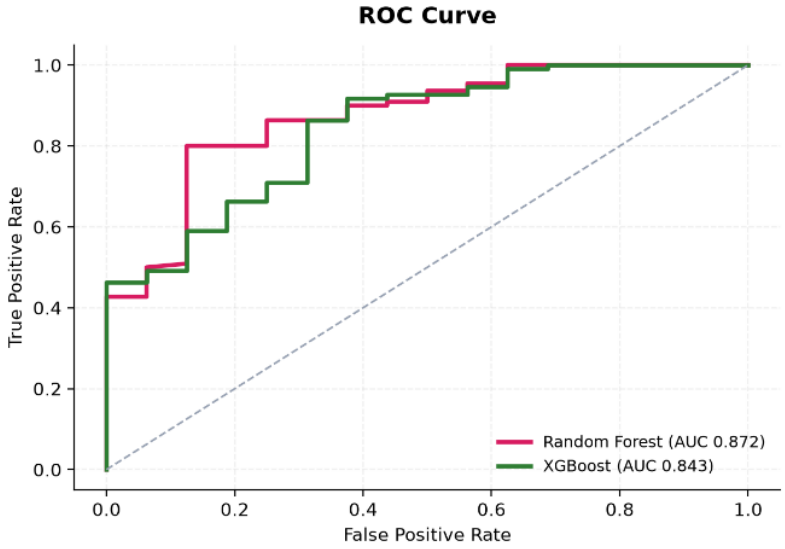



Figure 4 ROC curves of Random Forest and XGBoost on the 2026 questionnaire data

On the 2026 questionnaire dataset, Random Forest and XGBoost achieved identical Accuracy (0.8730), Precision (0.9196), Recall (0.9364), and F1-Score (0.9279). However, Random Forest obtained a higher AUC (0.8724) than XGBoost (0.8432). These results indicate that both models produced the same reported classification metrics at the selected decision threshold, while their discrimination performance across probability thresholds differed. Because Target 1 represented 87.1% of the full dataset, the results should be interpreted in light of the class distribution rather than Accuracy alone.

To further examine the classification results, the actual and predicted class distributions on the held-out test sets were compared for both algorithms and datasets. The number of actual observations in each target class and the number of observations predicted as Target 0 and Target 1 by each model in Table 7.

Table 7 Actual and Predicted Class Distributions on the Test Sets

| Dataset | Model | Actual Target 0 | Actual Target 1 | Predicted Target 0 | Predicted Target 1 |
|---|---|---|---|---|---|
| Historical 2022 | Random Forest | 48 | 78 | 49 | 77 |
| | XGBoost | 48 | 78 | 50 | 76 |
| Questionnaire 2026 | Random Forest | 16 | 110 | 14 | 112 |
| | XGBoost | 16 | 110 | 14 | 112 |

As shown in Table 7, the test-set class distribution differed between the two datasets. The 2022 test set contained 48 Target 0 and 78 Target 1 observations, whereas the 2026 test set contained 16 Target 0 and 110 Target 1 observations. Both models produced the same predicted class counts on the 2026 test set, with 14 observations classified as Target 0 and 112 as Target 1. These distributions provide additional context for interpreting the evaluation metrics, particularly because Target 1 dominated the 2026 test set.

### *3.4 Model Stability Based on Five-Fold Cross-Validation*

Five-fold cross-validation was conducted on the training data to assess the consistency of model performance across validation folds. Accuracy, Precision, Recall, F1-Score, and AUC were calculated for each fold. The mean and standard deviation of these evaluation metrics for Random Forest and XGBoost on both datasets in Table 8.

Table 8 Mean and standard deviation of five-fold cross-validation results

| **Dataset** | **Metric** | **Random Forest** | **XGBoost** |
|---|---|---|---|
| 2022 Historical | Accuracy | 0.7890 ± 0.0425 | 0.7631 ± 0.0329 |
| | Precision | 0.8491 ± 0.0532 | 0.8249 ± 0.0258 |
| | Recall | 0.8084 ± 0.0459 | 0.7892 ± 0.0720 |
| | F1-Score | 0.8269 ± 0.0341 | 0.8047 ± 0.0366 |
| | AUC | 0.8390 ± 0.0531 | 0.8102 ± 0.0572 |
| 2026 Questionnaire | Accuracy | 0.9004 ± 0.0172 | 0.8885 ± 0.0216 |
| | Precision | 0.9333 ± 0.0208 | 0.9324 ± 0.0192 |
| | Recall | 0.9542 ± 0.0080 | 0.9405 ± 0.0189 |
| | F1-Score | 0.9435 ± 0.0092 | 0.9362 ± 0.0123 |
| | AUC | 0.8629 ± 0.0685 | 0.8472 ± 0.0987 |

The cross-validation results show that Random Forest achieved a higher mean F1-Score than XGBoost on both datasets. Random Forest also exhibited a lower F1-Score standard deviation, indicating less variation across the validation folds. These findings support the relative consistency of Random Forest during cross-validation. However, cross-validation results should be distinguished from the final test-set metrics, which evaluate model performance on held-out observations.

### *3.5 SHAP-Based Model Interpretation*

SHapley Additive exPlanations (SHAP) was applied to the selected model to examine the relative contribution of the predictor features to the model's predictions. The mean absolute SHAP value (Mean |SHAP|) was calculated across observations in the test set. A higher value indicates a greater average contribution to the model output within the evaluated dataset. The results are presented in Table 9.

Table 9 Mean absolute SHAP values for the selected model

| Feature | 2022 Historical Data | 2026 Questionnaire Data |
|---|---|---|
| Hedonic Browsing (HB) | 0.1694 | 0.1843 |
| Social Environment (SE) | 0.0760 | 0.1254 |
| Utilitarian Browsing (UB) | 0.0601 | 0.0718 |
| Physical Environment (PE) | 0.0571 | 0.0426 |
| Time Perspective (TP) | 0.0471 | 0.0425 |

The SHAP analysis identified Hedonic Browsing (HB) as the feature with the greatest relative predictive contribution in both the 2022 historical dataset and the 2026 questionnaire dataset. This indicates that HB contributed more strongly to the model predictions than the other features included in the analysis. However, SHAP values explain the contribution of features to model predictions and do not establish causal relationships. Therefore, the results should be interpreted as predictive associations rather than evidence that HB causes impulsive buying behavior.

### *3.6 Comparison of Algorithm Performance*

Random Forest (RF) outperformed XGBoost across all reported test-set metrics on the 2022 historical dataset. On the 2026 questionnaire dataset, both models achieved identical Accuracy (0.8730), Precision (0.9196), Recall (0.9364), and F1-Score (0.9279), while RF obtained a higher AUC (0.8724) than XGBoost (0.8432).

Five-fold cross-validation also favored RF, which achieved higher mean F1-Scores and lower standard deviations than XGBoost on both datasets, indicating more consistent performance across validation folds. SHAP identified Hedonic Browsing (HB) as the feature with the greatest predictive contribution in both datasets. This represents a contribution to model predictions, not a causal effect on impulsive buying.

These findings should be interpreted in light of differences in data collection periods, sources, sampling procedures, and target-class distributions. Because the datasets were analyzed independently, the results do not establish temporal changes in consumer behavior or direct cross-period generalization. Future research should use broader and more representative samples, stronger cross-temporal validation, and independent datasets to assess model robustness and generalizability.

Although both datasets use the same constructs, indicators, and five-point Likert scale, they were collected in different periods using different non-probability sampling procedures: purposive sampling for the 2022 dataset and convenience sampling for the 2026 dataset. Therefore, the datasets cannot automatically be considered statistically equivalent or representative of the same population. Differences in class proportions and model performance may reflect differences in sample composition and data collection procedures, rather than changes in impulsive buying behavior alone. Accordingly, the findings should be interpreted as dataset-specific evaluations of model performance, not as evidence of temporal changes in consumer behavior or proof of cross-temporal generalization. Future research should use more comparable sampling procedures, examine respondent characteristics across datasets, and assess measurement invariance before drawing direct conclusions about behavioral changes over time.

## 4. CONCLUSIONS

This study compared Random Forest and XGBoost for predicting impulsive buying behavior using the 2022 historical dataset and the 2026 questionnaire dataset. On the 2022 test set, Random Forest outperformed XGBoost across all reported evaluation metrics. On the 2026 test set, both models achieved identical Accuracy, Precision, Recall, and F1-Score values, while Random Forest obtained a higher AUC. Five-fold cross-validation also showed a higher mean

F1-Score and lower F1-Score standard deviation for Random Forest on both datasets, indicating more consistent performance across validation folds.

The SHAP analysis identified Hedonic Browsing as the feature with the greatest relative predictive contribution in both datasets. This finding describes the contribution of HB to the model predictions and should not be interpreted as evidence of a causal effect on impulsive buying behavior.

This study has limitations related to the use of datasets collected in different periods and through different sampling techniques. Consequently, differences in class distributions and model performance cannot be attributed solely to changes in consumer behavior over time, and the findings should not be generalized beyond the populations represented by the datasets. Future research should use more diverse samples, strengthen cross-temporal validation, and evaluate model stability across additional datasets and periods.

## REFERENCES

[1] S. Huang, A. D. K. Silalahi, I. J. Eunike, and D. Riantama, "Understanding impulse buying in E-commerce: The Big Five traits perspective and moderating effect of time pressure and emotions," *Telemat. Informatics Reports*, vol. 15, 2024, doi: 10.1016/j.teler.2024.100157.

[2] A. D. K. Silalahi, D. T. T. Phuong, A. P. Tedjakusuma, I. J. Eunike, and D. Riantama, "How does time pressure shape impulsive buying behavior ? Hedonic vs . utilitarian values emerges as a key driver on E -commerce platforms," *Digit. Bus.*, vol. 5, no. 2, p. 100138, 2025, doi: 10.1016/j.digbus.2025.100138.

[3] X. Chen, "Consumer Online Shopping Behavior Prediction Based on Machine Learning Algorithm," *Procedia Comput. Sci.*, vol. 262, pp. 1395–1401, 2025, doi: 10.1016/j.procs.2025.05.187.

[4] S. Budiman and T. Wijaya, "Mobile app impulsive buying : A situational factors dataset analysis," *Data Br.*, vol. 50, p. 109559, 2023, doi: 10.1016/j.dib.2023.109559.

[5] N. Nyoman, C. Villdina, M. J. Efata, E. Frederika, and M. S. Maulana, "APPLICATION OF THE RANDOM FOREST ALGORITHM IN PREDICTING IMPULSE BUYING BY CONSUMERS ON INDONESIAN MARKETPLACE APPLICATIONS," *PILAR Nusa Mandiri J. Comput. Inf. Syst.*, vol. 22, no. 1, pp. 78–86, 2026, doi: 10.33480/pilar.v22i1.7893.

[6] S. Nurfadhilah and M. Azhari, "Application of The Random Forest Algorithm in Classifying the Tendency of Impulsive Purchasing Behavior Among Gen Z Consumers in E-Commerce Based on Flash Sale Features," *Inf. Commun. Technol.*, vol. 17, no. 1, pp. 75–84, 2026.

[7] F. N. F. Insani and Denny, "Optimizing E-Commerce in Indonesia: Ensemble Learning for Predicting Potential Buyers," *Indones. J. Comput. Sci.*, vol. 13, no. 1, pp. 373–385, 2024.

[8] Y. Pathak, B. Kumar, S. A. Khan, and J. Makhijani, "Explainable Machine Learning for Modeling Online Impulse-Buying Tendency: A Generational Comparison of Gen Z and Millennials," *Int. J. Artif. Intell. Mach. Learn.*, vol. 6, no. 10s, pp. 1441–1452, 2026, doi: 10.51483/IJAIML.6.10s.2026.1441-1452.

[9] N. Cui, "Explaining and predicting consumer decision-making in E-commerce using an intelligent hybrid framework for impulsive and planned purchase behavior," *Sci. Rep.*, vol. 16, 2026, doi: 10.1038/s41598-026-47284-1.

[10] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-56706-x.