Acta Scientific NUTRITIONAL HEALTH (ISSN:2582-1423)

Volume 10 Issue 5 May 2026

Research Article

A Comparative Study of Statistical and Machine Learning Techniques for Predicting Customers Shopping Behavior

Alaa A Elnazer[1]*, Fawzia Abdu Alsalam Al Tboli[2], Gehad Elgebaly[3] and Mahjoub A Elamin[4]

[1]Department of Marketing, College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

[2]Department of Statistics, Faculty of Science, Benghazi, University of Benghazi, Libya

[3]Department of Economics, Faculty of Business Administration, Delta University for Science and Technology, Gamasa, Egypt

[4]Department of Mathematics, University College of Umluj, University of Tabuk, Saudi Arabia

***Corresponding Author:** Alaa A Elnazer, Department of Marketing, College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia.

DOI: 10.31080/ASNH.2026.10.1623

Received: January 27, 2026

Published: May 08, 2026

Abstract

This study develops a comprehensive predictive framework by systematically comparing five classification models—Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), Artificial Neural Networks (ANN), and Extreme Gradient Boosting (XGBoost)—using the Online Shoppers' Purchasing Intention dataset. A diverse set of performance metrics, including Accuracy, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2, Correlation Coefficient (CC), Coefficient of Variation (COV), and Error Coefficient (EC), were employed to evaluate and benchmark the models. Descriptive statistics and correlation analysis provided a foundational understanding of the behavioral attributes shaping purchasing outcomes, while inferential analyses, including ANOVA and the Wilcoxon Signed-Rank Test, confirmed statistically significant differences among models and validated the robustness of the comparative framework. The findings suggest that the Random Forest model was the best in most of the evaluation measures as it had the lowest RMSE, the highest correlation with the actual outcomes, and the most stable. Even though Artificial Neural Networks displayed similar levels of accuracy, the Random Forest was more consistent and reduced the number of predictive errors, which highlights why this algorithm can be used to find out the customer behavior in very complex and nonlinear scenarios. The results show that ensemble techniques are significant in prediction of e-commerce and that hybrid methods have the potential to increase the accuracy and generalization. The research has both methodological and practical significance because it provides a strict standard of the classification algorithms and offers practical information to online retailer which needs to optimize its decision making process, customer satisfaction and long term customer loyalty.

Keywords: Online Shopping; Root Mean Squared Error (RMSE); Correlation Coefficient (CC)

Introduction

In the modern world of internet technologies, online shopping has become an essential part of life. With the ever rising e-commerce trends, companies are currently examining new ways of improving customer satisfaction and loyalty. Product reviews are considered to be one of the main factors which have a significant impact on online purchasing decisions. Positive testimonies do not only create trust among the potential consumers, but also play a notable role in enhancing sales as well as profitability. Consequently, it is paramount to promote more reviews based on the customer purchasing behavior to the success of a company.

Purchasing behavior of the customers provides valuable information that can be employed in order to improve the quantity and quality of the reviews. Through such behaviors, the companies would be in a position to establish trends and patterns that enhance formulation of desirable review strategies. As an illustration, repeat buyers are usually willing to do positive reviews. Similarly, the audience that engages with social media pages of a brand is always ready to give feedback. Having these tendencies enables businesses to make personalized requests of the reviews, thereby having a higher likelihood of getting a positive review. Also, learning the behavior of shoppers can indicate the areas that require modulation and business can work on their products to offer customers a better experience.

Literature Review

This study has been of great interest in the last few years, mainly due to the fast developments of e-commerce sites and the desire to predict online shopper satisfaction and purchasing intention. Extensive literature has shown that machine learning techniques are useful in the ability to identify latent behavior trends and predict the result of purchase.

In their study [1] compared the use of Customer satisfaction prediction based on Random Forest, Gradient Boosting, and Support Vector Machines with a deep learning model and identified the best accuracy results as 92% with a well-explained model of the former. In the same vein [2] indicated that XGBoost was more efficient than Random Forest in terms of precision, recall, and F1-score concerning the prediction of purchasing behavior. These findings are in direct correlation to the results of the current study, in which both Random Forest and XGBoost yielded good results, with the former demonstrating a more consistent high score in comparison with the latter.

A number of studies also highlighted the competence of tree-based ensemble in consumer analytics. As an illustration, [3] used XGBoost to behavioral and demographic variables and found significant determinants of purchase intention in addition to the interpretability strength of gradient boosting. Similarly, an XGBoost study by the authors revealed an XGBoost accuracy of 90.65 with feature selection and resampling, indicating the need to tackle the problem of class imbalance in e-commerce data [4]. This is consistent with the existing study which also found imbalance in the Revenue variable (~15% positive cases) and its effects on predictive performance.

The use of Artificial Neural Networks (ANN) in the satisfaction modeling is also popular. Although they can learn complex nonlinear patterns, variability and parameter adjustment have been reported to be difficult [5] compared to ensemble methods. This observation is not isolated to the current study findings with ANN having competitive accuracy and higher variability and lower stability than that of the Rand Forest and XGBoost.

Moreover, new inputs have examined hybrid and sophisticated frameworks. As an example, (Balasundaram., *et al*. 2024) combined K-means clustering and XGBoost in predicting loyalty, and a study in Vietnam (2024) used BERT-based sentiment features and XGBoost and achieved better results than control models. On the same note [6] suggested an RF–LightGBM ensemble that used SMOTE-ENN sampling to manage the imbalance, which was much better than single models. Improvements through optimization have also been reported, including an XGBoost (2024) optimized using a Sparrow Search Algorithm (SSA), which had improvements in both AUC and F1-score. Although the current study did not run these hybrid or optimization methods, the findings of it support the effectiveness of ensemble baselines (RF and XGBoost) and, thus, these hybridization methods can be viewed as potential ways to conduct research in the future.

In all the studied articles, there is one common observation: the ensemble-based methods (Random Forest, XGBoost, and LightGBM) are superior to simpler classifiers (Logistic Regression and KNN), and may perform equally with deep learning models in structured behavioural data. This consensus is also reflected by the findings of the present research as Random Forest turned out to be the most efficient model in terms of accuracy, reduction of error and stability. In the meantime, Logistic Regression and KNN were not performing well, which confirmed the fact that they have a limited ability to capture the nonlinear consumer behavior.

Overall, the literature offers solid support that the ensemble learning is more dominant in the prediction of e-commerce behavior. This is corroborated by the current investigation that reveals the use of Random Forest as the most viable, consistent and robust classifier, besides indicating that superior hybrid model, balancing, and optimization approaches, which have been the subject in many previous studies, present incredible opportunities to enhance predictive accuracy in future.

Methodology

Machine learning model

There were no preprocessing operations outside the training set to allow data leakage. Numerical features were scaled (either by a Z-score standardization or by min maximum normalization, depending on the algorithm) by training the scaler on the training data and then applying the parameters learned on the test set. In the same manner, categorical features were encoded (through one-hot encoding) based on categories constructed only based on the training data. These steps were taken with scikit-learn pipelines, which meant that they could apply the transformations to the data in a consistent and reproducible way both in cross-validation folds and in the final test set.

Logistic regression: A statistical framework for binary outcome modeling

Logistic regression is a basic statistical technique utilized for the modeling of binary outcome measures where the response takes on one out of two values (e.g., positive/negative, yes/no, event/non-event) [7]. Unlike linear regression, which is appropriate for continuous dependent measures, logistic regression is tailored to provide the estimate of the probability of occurrence of a given outcome in order to facilitate classification into one out of two categories.

The logic behind logistic regression lies in the fact that it is grounded in the theory of probability and the estimation of odds. The log-odds of the probability of the dependent variable is given by a linear combination of the set of input features. The function of logistic regression is given in mathematical form through the sigmoid function (the logistic function), which maps any real number into the range from 0 up to 1 [8].

$$\frac{1}{(e^{\omega 0} + \sum_{i=1}^{n} w_i x^i) + 1} = P(Y = 1|X) = (X|1 = P(Y$$

$$\frac{(e^{\omega 0} + \sum_{i=1}^{n} w_i x^i)}{(e^{\omega 0} + \sum_{i=1}^{n} w_i x^i) + 1} = P(Y = 1|X) = (x|0 = P(Y = 0|x)$$

Where:

X = represents the vector of input features $[x^1, x^2, \ldots \ldots . . x^n]$,

ω_0 is the intercept term (bias),

$w_i x^i$ are the coefficients (weights) assigned to each feature x^i

$p(y|X)$ denotes the conditional probability that the dependent variable Y takes the value 1 given the input vector X,

The parameters = W ($\omega_1 . \omega_2 \ldots \ldots , \omega_n$) are estimated through a method known as Maximum Likelihood Estimation (MLE),

MLE seeks to identify the parameter values that maximize the likelihood of the observed outcomes in the training data,

The likelihood function is defined as:

$$L(\beta) = \prod_{i=1}^{n} p(Y_i|X_i)^{y_i} \, (1 - p(Y_i|X_i)^{1-y_i}$$

T0 simplify computations, the log – likelihood function is often used:

$$log\, L(\beta) = \sum_{i=1}^{n} [Y_i . log(P(Y_i = 1|x_i)) + (1 - Y_i). log(1 - p(Y_i = 1|X_i))]$$

Estimation of the parameters for logistic regression is performed by maximizing the log-likelihood function by means typically executed using the application of iterative numerical optimization procedures such as the Newton-Raphson algorithm or Iteratively Reweighted Least Squares (IRLS). Such procedures estimate the coefficient values iteratively by utilizing the gradient (the first derivative) and the Hessian matrix (the second derivative) for the likelihood function. Optimization continues until the optimization algorithm has converged according to a predetermined criterion—typically when the variation in the values between parameters in two consecutive iterations is smaller in comparison to a predetermined threshold or until the maximum number of iterations has been reached [9].

One of the most important merits of logistic regression is the ability of the model to incorporate both continuous and categorical independent measures and hence is applicable in the widest range of applications from credit risk prediction to clinical diagnosis, customer profiling, and public policy research. A logistic model might actually apply predictor measures like income level, employment duration, age, and education level—all continuous or ordinal measures—in order to estimate the probability of a binary outcome like the event of default on a loan or the extension of credit [10].

Key within this approach lays the utilization of the logit transform, the natural log of the odds occurring for the outcome event. The logit function is the link function in the generalized linear model framework and is described by:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

Logistic regression is a useful statistical technique commonly used in many fields like healthcare, marketing, or finance to generate insights from binary outcomes (e.g., sick vs. not sick). However, when applying logistic regression to complex survey data, which includes complex sampling designs, specific methodological issues are often overlooked. The use of the logit transformation ensures a linear relationship between the predictors and the log-odds of the response variable, thereby enhancing model interpretability and facilitating the estimation of parameters through linear methods [11].

From a modeling standpoint, the comparison between linear and logistic regression can be visualized through the following representations:

1. $linear\ model:\ \beta . X = y$
2. $logistic\ model: \sigma(x) = \frac{1}{e^{-x}+1}$

Here, σ(x)\sigma(x)σ(x) denotes the sigmoid activation that introduces the non-linearity required to bound predictions.

Logistic regression is not only theoretically well-motivated statistically but is also practically strong in the realm of predictive modeling. It utilizes probabilistic output, strong estimation methods, and transform functions in order to effectively model intervariable relations. With both mathematical strength and interpretability and flexibility accompanying it, it serves as a central approach in research as well as in practical machine learning applications [12].

K-Nearest Neighbor (KNN)

In this study, the K-Nearest Neighbor is a powerful nonparametric classifier which assigns an unclassified pattern to the class represented by a majority of its k nearest neighbors. The k-nearest neighbor›s algorithm (k-NN) is a traditional nonparametric method used for classification and regression. k-NN is a type of instance-based learning (a.k.a. lazy learning), which means that the training process only stores the samples and all the computation cost is induced in the test process to find the k nearest neighbors [13].

Given a training dataset $D = \{(x_n, y_n)\}_n^N = 1$and a test sample x0, the goal is to predict the category of x0. In the training process the dataset x0 is loaded and stored. After that, the test process searches k nearest neighbors from the training dataset, and k is the hyper parameter selected at the beginning. Generally, we can apply the voting method in the classification task, that is, choosing the most frequent label in the k nearest neighbors to be the label of x0 as shown in Figure (1). Similarly, in the regression task the mean value of the labels of k nearest neighbors is set to be the label of x0 [14].

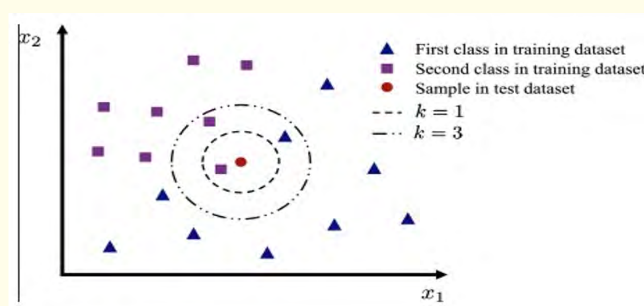


Figure 1

Obviously, the key factor of k-NN is to choose the k nearest neighbors. Generally, the distance metric is denoted as the Euclidean distance.

$$d(x_0, x_i) = \|(x_0, x_i)\|_2$$

Therefore, the k nearest neighbors can be selected based on the k minimal Euclidean distances from the training dataset D. We denote the set of the k nearest neighbors by D*K∈D, that is, the number of D*k is k. Finally, we utilize the voting method or mean value to execute classification tasks or regression tasks.

To improve the performance of k-NN, large margin nearest neighbor or neighborhood components analysis can be adopted to enhance the accuracy. The weighted nearest-neighbor classifier assigns weights to all neighbors to promote the performance and condensed nearest neighbor reduces the dataset to decrease the amount of calculation [15].

Feature scaling and normalization

Prior to applying the KNN model, feature scaling was a crucial preprocessing step. This ensured that all features contributed equally to the distance calculations. In particular, Z-score standardization and min-max normalization were applied to standardize the numerical variables. This step is essential because KNN relies heavily on the computation of distances between observations, and unnormalized features could disproportionately affect the model's predictions.

Other commonly used distance metrics include Manhattan distance and Minkowski distance, but Euclidean was selected due to its suitability for continuous, normalized features in our dataset.

Optimal K value selection

The parameter k, representing the number of neighbors considered for classification, plays a significant role in determining model flexibility and generalization. A smaller value of k increases model sensitivity and risk of overfitting, while a larger value enhances model stability. In this study, the optimal value of k was selected using cross-validation, ensuring the model achieved a balance between bias and variance.

Model application

Unlike parametric models, KNN is a lazy learner, meaning that it does not build a predictive model during training. Instead, it stores the entire training dataset and defers computation until a prediction is requested. During testing, the algorithm computes the distances between the test point and all training samples, identifies the k nearest neighbors, and assigns the most frequent class label among them as the final prediction [16]. By applying KNN to the processed dataset, customer satisfaction was classified into two binary categories (e.g., Revenue = 0 for unsatisfied, Revenue = 1 for satisfied), allowing for a clear evaluation of performance alongside other classifiers such as Logistic Regression, Random Forest, and Artificial Neural Networks (Zhang., *et al*. 2024).

Artificial neural network (ANN)

The model of an Artificial Neural Network (ANN) that was used in the paper and implemented in the current research with the help of the MLPClassifier of the scikit-learn library was used to estimate complex and nonlinear interactions between the variables of customer behavior and the target output of the purchase decision (Revenue). ANN is a supervised learning method based on the human brain structure, and it is an arrangement of layers of interconnected nodes (neurons) that interact to learn the functions on the basis of the data fed to it [17].

Network configuration

The architecture used comprised a single hidden layer with 100 neurons. Each neuron applied the Rectified Linear Unit (ReLU) activation function, defined as:

$$max\,(0, x) = f(x)$$

The reason why this activation was chosen is that it is computationally efficient and avoids the vanishing gradient issue. An output layer included one neuron that has a sigmoid activation function, which converts the output to a probability score (between 0 and 1) of whether or not the customer makes a purchase (i.e., Revenue = 1). The total output is mathematically shown as:

$$b_i^{(1)} + X_i^{(1)}\,\omega = \sigma\,(\omega_0 + \textstyle\sum_{i=1}^{100} \omega_i\,.\,ReLU\,(z_i))\ where\ z = \hat{y}$$

Data preprocessing and feature scaling

Before training the network, all input features were normalised by z-score, in order to make them have a common scale (zero mean and unit variance). This is required in neural network models, where some features may have a disproportional impact on the learning process and the optimization algorithm convergence is improved [18].

Model training

Training of the network was done with the stochastic gradient descent (SGD) optimization, which made use of the backpropagation algorithm. The cross-entropy loss was minimized to a maximum of 500 iterations or till convergence. The solver parameters and learning rate (0.001) were also empirically tested to ensure that a balance between the training speed and the accuracy of the model was achieved [19].

Prediction and use in evaluation

Upon training, the ANN model was used to generate probabilistic predictions via the predict_proba() function. These continuous scores were subsequently:

- Converted to binary classifications using a threshold of 0.5.
- Employed in statistical validation techniques, such as ANOVA and Wilcoxon Signed-Rank Test.
- Used to compute residuals (difference between actual and predicted probabilities) for diagnostic visualizations like the Q-Q Plot.

Justification and role in the study

The inclusion of the Ann model is appropriate from the ability to learn latent interactions and hierarchical representations within the dataset, which cannot be effectively captured by more traditional models such as logistic region or K-Nn. To determine its suitability in predicting customer procurement behavior, its performance was evaluated with other classifiers in terms of several matrix [20].

Random forest

Random forest is a combination of predictions of the tree such as each tree depends on the values of a random vector independently and with equal distribution for all trees in the forest. Random Forest (RF) is a cloth-based machine learning algorithm introduced by [21] and has performed strong performance in both classification and regression tasks. In terms of this research, RF was employed as one of the future models to assess online shopkeeper satisfaction based on behavioral features and transactions facilities.

The decision creates a crowd of trees during training through the technique known as the RF algorithm bootstrap collection, or the technique known as bagging. Each individual tree is trained on a random mastery of training data - in terms of both features and samples - swim with replacement from the original dataset. The final prediction is obtained by collecting outputs of all individual trees through majority voting (for classification) or average (for regression) (for regression).

In this study, the random forest (RF) model was applied through a systematic process to maximize the future accuracy and minimize on fittings. First, the appropriate number of decision trees were specified to ensure model diversity and generalization. Each tree was trained on a bootstrap sample generated with replacement, while unused cases are known as co-of-off-bags samples-maintained for technical verification. According to the RF approach, trees were allowed to grow to the entire depth without pruning complex data structures. The model performance was then estimated using OOB predictions, which is shown to provide a reliable and fair option for cross-satisfaction, reducing computational costs [22]. To ensure comprehensive performance evaluation, several assessment matrix was employed, including the root medium square error (RMSE), meaning full error (MAE), the coefficient of determination (RG), Visiting Score (EVS), and accuracy. Such measures are widely recommended in RF applications, for example, in short -term current flow forecast where RF performed a strong future stating power [23]. In addition, feature importance analysis -average reduction in accuracy (MDA) and decrease in impurities (MDI) such as using metrics, raised lecturer and held to highlight the most impressive behavior indicators of customers' satisfaction (Zhu., *et al*. 2021; [24]). Collectively, this method structure ensured both strength and explained the ability of RF model in predicting customer results.

Extreme gradient boosting (XGBoost)

Extreme Grade Bosting (XGBOOS) is a highly efficient and scalable implementation of the gradient boosting disease trees introduced by [25] and has been developed on the founding concepts of Grade Bosting Machines [26]. The algorithm is widely adopted due to its better computational speed, parallel and capacity for distributed teaching, and high future stating accuracy, in particularly structured data applications such as customer behavior modeling and prediction of satisfaction. In this study, Xgboost was employed as one of the main classification models to predict the satisfaction levels of online shopkeepers. The ability to capture complex non -linear feature interactions, handle missing values and incorporate advanced regularization techniques (both L1 and L2) effectively reduces the risk of fittings, causing it to suit especially for real-world e-commerce datasets.

Let the training dataset be represented as = where= is the satisfaction label (binary classification). The XGBoost model predicts the target value using an ensemble of K additive regression trees as follows:

$$\mathcal{F} \ni k f_k(x_i), F(x_i) \sum_{k=1}^{K} = F_k(x_i) = \emptyset(x_i) = \hat{y}_i$$

Where: each f_k is a function in the space of regression trees , defined as

$$\{f(x) = \omega_{q(x)} \mid q: \mathbb{R}^m \to T, \omega \in \mathbb{R}^T\} = f$$

Here, q(X) maps an input feature vector to the index of a leaf node in the tree, is the number of leaves, .and represents the leaf scores.

To train the model, XGBoost minimizes a regularized objective function designed to balance the predictive accuracy and model complexity

$$\Omega(f_t) + \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} f_t(x_i)\right)$$

Where:

l is a differentiable convex loss function (e.g., logistic loss for binary classification),

$\hat{y}_i^{(t-1)}$ is the prediction from the previous iteration,

(f_t) is the new function (tree) added at iteration t,

$\Omega(f)$ is the regularization term given by:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$$

Regularization parameters of \ _ Gamma and λ \ lambdaλ control the complexity of individual trees, which helps reduce overfiting-a common issue in a user behavior and transaction records such as high-dimensional data.

In terms of this research, Xgboost model was tuned using cross-satyapan to adapt hyperpremes such as learning rates, depths of trees and estimates such as hyperpremes. The final model demonstrated a strong performance in making accurate prediction whether the customer was likely to be satisfied on the basis of their online shopping behavior.

To ensure methodological rigor and minimize overfitting, the dataset was partitioned into two mutually exclusive subsets: training set (80%) and a testing set (20%), with the split performed in a stratified manner to preserve the class distribution of the target variable (Revenue). All models were trained exclusively on the training set, while the unseen testing set was reserved for final performance evaluation. Hyperparameter optimization—for models such as Random Forest, ANN, and XGBoost—was carried out using stratified cross-validation applied only within the training set, thereby preventing information leakage from the test set. The performance metrics reported in this study are based on the independent testing set, ensuring that the results are unbiased and reproducible. Furthermore, all random processes (e.g., data splitting, weight initialization) were controlled by fixed random seeds to guarantee experimental reproducibility.

The target variable (Revenue) exhibited a clear class imbalance, with only about 15% of sessions resulting in a purchase. To address this issue and ensure fair evaluation, different strategies were applied depending on the model. For Logistic Regression and Random Forest, the class_weight='balanced' option was employed to adjust the loss function according to the inverse class frequencies. For K-Nearest Neighbors and Artificial Neural Networks, the imbalance was handled indirectly by training on the stratified dataset and validating performance on the held-out test set, while carefully reporting both error-based and accuracy-based metrics to capture potential bias toward the majority class. For XGBoost, a scale-pos-weight parameter was set based on the class distribution to compensate for the skewed labels. Importantly, all balancing strategies were applied strictly within the training set to avoid data leakage, while the testing set remained untouched to provide an unbiased estimate of model performance.

Evaluation metrics

Mean absolute error (MAE)

Root Mean Squared Error (RMSE) is another widely used evaluation metric in predictive modeling, offering a complementary perspective to Mean Absolute Error (MAE). RMSE is calculated as the square root of the average squared differences between predicted values and actual outcomes, thereby providing a measure of the standard deviation of prediction errors [27]. A lower RMSE value indicates that the model's predictions are closely aligned with the true values, with fewer large deviations. Unlike MAE, RMSE penalizes larger errors more heavily due to the squaring operation, making it particularly useful in contexts where large deviations from the actual outcomes are considered more critical. In this study, RMSE was employed to evaluate the predictive accuracy of models such as Logistic Regression, KNN, Random Forest, ANN, and XGBoost in classifying e-commerce customer satisfaction. By emphasizing larger errors, RMSE provides an additional diagnostic tool for identifying models that not only achieve overall accuracy but also minimize the occurrence of substantial miss predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Where:

n is the total number of observations,

y_i represents the actual value (e.g., whether the customer was satisfied), $\hat{y}_i$ represents the predicted probability or label by the model.

Mean absolute percentage error (MAPE)

Media is a widely used assessment metric to assess future accuracy, especially when an explanable measure of error in terms of percentage is desirable. It is calculated as an average of full percentage differences between real values and forecasted values, providing a spontaneous indication of model performance relative to the scale of the data [28]. In this study, Mape was employed to evaluate how developed machine learning models- Random Forest, Logistics Regression, Artificial Neural Networks (ANN), Extreme Graadant Boosting (Xgboost), and K-NEAREST neighbor (KNN)-KNN)-On the basis of customer satisfaction based on behavior shopping data. Unlike full metrics such as MAE or RMSE, which measures errors in the same unit as prediction, MAPE expresses errors as a percentage of real values, offering a generalized and easily interpretable performance indicator. This property suits MAPE for comparative model evaluation, especially in commercial contexts such as e-commerce, where relative deviations are important for effective decision making.

The coefficient of assessment (R g)) is one of the most employed statistical matrix to evaluate the good-to-fit of the future model. This determines the ratio of variance in dependent variables explained by independent variables, giving a sign of how well the model captures the underlying data pattern [29]. A R RAM value close to 1 indicates that the model accounts for a high ratio of variance, which reflects a strong future, while close to 0 values suggest weak explanatory ability. In this study, Random Forest, Logistics Region, N, XGBOST, and KNN were used to assess the explanatory strength of the machine learning models such as the e-commerce customer satisfaction. Unlike error-based matrix such as MAE, RMSE, or MAPE, which determines the magnitude of prediction errors, R, measuring that degree provides a supplementary perspective by measuring that degree that the model explains variability in customer behavior. This makes it particularly valuable not only to understand the forecast accuracy but also to understand the explanatory relevance of the developed model.

The formula for calculating MAPE is given as follows:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

where:

n denotes the number of observations,

y_i represents the actual value (true label),

$\hat{y}_i$ represents the predicted value from the model.

A low mean absolute percentage (MAPE) value indicates a better future performance performance, as it indicates that the estimated price is close to the actual results in terms of percentage. Nevertheless, a significant range of MAPE is its volatility when the actual values reach zero, which should be carefully addressed during data preprosying and interpretation [30]. In terms of prediction of customer satisfaction, Mape is particularly valuable as it allows researchers and physicians to compare the relative accuracy of several models in estimating user behavior patterns. By expressing the errors of prediction in percentage conditions, the mAPE enhances interpretation and facilitates decision-making, supporting personal marketing and customer experience strategies to support data-managed reforms.

Coefficient of determination (R-squared, R^2)

The coefficient of determination (R and) is a fundamental statistical metric that is widely used to evaluate the explanatory power of the future model. It determines the ratio of variance in dependent variables that can be explained by independent input variables, providing insight into the degree from which a model captures the underlying data pattern. In the context of this study, R OF was employed to assess that the machine learning models were employed to include logistics Regression, KNN, Random Forest, N, and XGBost -folk and to repeat the variability in customer satisfaction based on behavior and transaction characteristics. A high R, the value, indicates strong explanatory ability, not only reflects the ability of the model to achieve the future accuracy, but also pays attention to the variance contained in customer satisfaction results (Netter., *et al*. 1983).

Mathematically, R^2 is expressed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y^-)^2}$$

Where:

y_i is the actual observed value,

$\hat{y}_i$ is the predicted value y,

y^- is the mean of actual values ,

n is the number of observations n

The value of R2R^2R2 ranges from 0 to 1, where a value close to 1 indicates a strong correlation between approximate and real results, a good model fit. Conversely, an R2R^2R2 near 0 states that the model explains very little of variability in the target variable. This metric is particularly informative in assessing how well the model normalizes in different customer behavior patterns.

Root mean squared error (RMSE)

The root medium is one of the most widely planned matrix to evaluate the future accuracy of the route medium repayment model. It is calculated as square root of the square difference between the approximate values and real comments, which captures both the variance and magnitude of the prediction errors. A major feature of RMSE is that it punishes large deviations due to the square of residues, causing it to become informative especially in contexts where adequate errors are particularly undesirable. In this study, RMSE was implemented to evaluate the performance of the machine learning model - which includes logistic Regression, KNN, Random Forest, N, and XGBost - in which customer satisfaction was predicted. RMSE is particularly useful in this context, as the results of large prediction errors can be important for operational or strategic decisions. In addition, because the RMSE target retains the original unit of the variable, it provides an intuitive interpretation of the model's error magnitude, indicating low RMSE values high future accuracy and a model that reflects real customer behavior [31] more closely.

The RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Where:

y_i denotes the actual values,

$\hat{y}_i$ denotes the predicted values,

n , is the number of data points.

Coefficient of variation (COV)

The coefficient of variation (cov) is a generalized statistical remedy that is widely employed to assess the dispersion relative to the approximate values in relation to their mean. It is calculated as a ratio of standard deviations (μ) (μ), it forms a Unite Indicator that facilitates comparability in different parameters of data. This metric is particularly valuable in evaluating and comparing the forecast performance of the model in the dataset where the limit and magnitude of values vary. In the context of this study, cov was used to evaluate the stability and strength of the classification model - such as logistic Regression, KNN, Random Forest, Ann, and Xgboost- In predicting customers' satisfaction. A lower cov indicates that the predictions of a model are more consistent, with low variability relative to them, which is highly desirable when assessing the model reliability in real-world e-commerce applications [32].

The formula for COV is:

$$COV = \frac{\sigma}{\mu}$$

Correlation coefficient (CC)

Correlation coefficient (CC) is a widely used statistical measure that determines the strength and direction of linear relations between the observed target values and model-future outputs. Between -1 and 1, the coefficient indicates to the extent that predictions follow the same trend as real results. A value of 1 reflects a strong positive correlation, indicating that the model closely aligns with the pattern seen, while 0 recommends the values weaker associations, and negative values mean reverse relationships [33]. In the context of customer satisfaction modeling, a higher CC value indicates that the future -staging model effectively catches the underlying behavior pattern and makes a replica of customer satisfaction trends with more reliability

CC is mathematically defined as:

$$\frac{cov\ (y^1, y)}{\sigma_{y^1}\ \sigma_y} = CC$$

Where:

Cov (y^1, y),, is the covariance between the actual and predicted values,

σ_y is the standard deviation of actual values,

σ_{y^1} is the standard deviation of predicted values.

Error coefficient (EC)

Error coefficient (EC) is a relative performance metric designed to evaluate the discrepancy between real and approximate values when accounting for the scale of data. It is calculated as a means of complete relative errors between predictions and comments, allowing deviations to normalize and allow fair comparison in dataset with different boundaries(Willmott & Matsuura, 2005). A low EC value refers to a close alignment between approximate and real values, highlighting high model reliability and accuracy. In the context of the prediction of customers' satisfaction, minimizing the EC ensures that model estimates effectively repeat the right customer feelings, which is necessary to achieve actionable insights and support strategic business decisions.

The formula for EC is:

$$\left|\frac{\hat{y} - y}{y}\right| \sum_{i=1}^{n} \frac{1}{n} = EC$$

Where:

y is the actual value,

$\hat{y}$ is the predicted value,

n , is the number of observations.

Overall accuracy and correlation matrix analysis

Overall, accuracy classification is one of the most widely adopted evaluation matrix, as it determines the ratio of the correct estimated results relative to the total numbers [34]. In terms of the prediction of customer satisfaction, overall accuracy provides a direct and explanatory measure of how a model effectively distinguish between satisfied and dissatisfied customers. A high accuracy value indicates strong forecast performance, indicating the ability of models to normalize in diverse customer behavior patterns and transaction references.Overall Accuracy=Number of Correct PredictionsTotal Number of Predictions\text{Overall Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}Overall Accuracy=Total Number of PredictionsNumber of Correct Predictions.

In the context of predicting revenue (purchase vs. no purchase), accuracy provides a direct and intuitive measure of how effectively a model can distinguish between buyers and non-buyers. A higher accuracy value reflects stronger classification performance. Nevertheless, accuracy alone may not fully capture model reliability in cases of class imbalance, so it is complemented in this study with additional error- and fit-based metrics.

In this study, two complementary evaluation approaches were adopted. First, the regression-style metrics-including Mae, MSE, RMSE, R and, and Correlation coefficients-was calculated using the approximate possibilities generated by the model. This approach provides a more fine understanding of how closely the approximate possibilities with real results are. Second, overall accuracy was calculated, which was done on the basis of binary classifications obtained by applying a certain range of 0.5 for potential outputs. The combination of these two approaches ensures functioning transparency: while the metric-style metrics occupy the quality of probability estimates, reflecting the ability to distinguish accuracy and differentiate between non-bred sessions.

However, relying perfectly on accuracy cannot provide a complete picture, especially when the dataset contains unbalanced classes or characteristics with complex intercourse. Therefore, a deep examination of relationships between features is necessary. To end this, a correlation matrix was formed to analyze linear associations between numeric variables within the dataset.

To examine the relationship between the input variables, a correlation matrix was formed to analyze linear associations between numerical characteristics in the dataset. Correlation matrix, illustrated in Figure 2, represents the Pearson correlation coefficient -1 to +1. Each coefficient indicates a degree in which two variables move simultaneously, which contains diagonal elements equal to 1 (full self-relation).

Correlation matrix analysis provides many important insight into user behavior. First, a very strong positive correlation * related to the product * and * related to the product is seen between \ _ Duration * (R = 0.86), suggests that customers who engage with a large number of products-related pages also spend more time proportionally. Second, * boom rates * and * exhaust rates * display an extraordinary high correlation (R = 0.91), showing that visitors who bounce from a page without meaningful interactions are also highly likely to completely get out of the website. In contrast, variables such as *traffic type *, *region *, and *browser *displays weak or negligible correlations with most other characteristics, which means that these features have limited impact in linear relations and cannot serve as strong predictions in later modeling functions. Such correlation analysis is essential in customer behavior research, as it helps not only identify fruitless or multi-collegir variables, but also provides action in user engagement patterns [35].

Understanding interrelations between the variables is essential for both functioning and practical reasons. From a methodical perspective, it supports effective convenience selection by identifying highly correlated predictions that can introduce excesses or multiple-linearity, allowing models to affect stability and lecturer [36]. From a practical perspective, it produces domain-specific insight into customer behavior, such as the relationship between time and shopping interest spent on the product pages, which enrich the relevant understanding of the online shopping pattern [37]. Overall, the operation of this analysis represents an important initial step before the model training, ensuring that the input variables contribute to the interpretation of both meaningful and accuracy, strength and future -stolen models [38].

Results

An initial descriptive analysis was performed to detect the fundamental characteristics of the dataset. Table 1 represents the summary statistics from 1 dependent variables (revenue) -the summary data for independent variables (X1 to x22x_ {22} x22), which represents purchasing results (1 = purchase, 0 = no purchase).

Reported measures include mean, standard deviation, minimum and maximum value, which provide observations of the central trend and spread of each variable. Such information is necessary to identify the overall distribution of data, to detect potential discrepancies such as obliqueness or outliers, and assessing the representation of the sample.

This descriptive examination model acts as an essential preparation step before development, as it ensures that the input variable is well understood and later the forecast analysis is based on reliable and balanced data.

Max	Q3	Q2	Std	Mean	Q1	Min	Feature
27.0000	4.0000	1.0000	3.3218	2.3152	0.0000	0.0	Administrative
3398.7500	93.2562	7.5000	176.7791	80.8186	0.0000	0.0	Administrative_Duration
24.0000	0.0000	0.0000	1.2702	0.5036	0.0000	0.0	Informational
2549.3750	0.0000	0.0000	140.7493	34.4724	0.0000	0.0	Informational_Duration
705.0000	38.0000	18.0000	44.4755	31.7315	7.0000	0.0	ProductRelated
63973.5222	1464.1572	598.9369	1913.6693	1194.7462	184.1375	0.0	ProductRelated_Duration
0.2000	0.0168	0.0031	0.0485	0.0222	0.0000	0.0	BounceRates
0.2000	0.0500	0.0252	0.0486	0.0431	0.0143	0.0	ExitRates
361.7637	0.0000	0.0000	18.5684	5.8893	0.0000	0.0	PageValues
1.0000	0.0000	0.0000	0.1989	0.0614	0.0000	0.0	SpecialDay
8.0000	3.0000	2.0000	0.9113	2.1240	2.0000	1.0	OperatingSystems
13.0000	2.0000	2.0000	1.7173	2.3571	2.0000	1.0	Browser
9.0000	4.0000	3.0000	2.4016	3.1474	1.0000	1.0	Region
20.0000	4.0000	2.0000	4.0252	4.0696	2.0000	1.0	TrafficType
1.0000	0.0000	0.0000	0.4225	0.2326	0.0000	0.0	Weekend
1.0000	0.0000	0.0000	0.3617	0.1547	0.0000	0.0	Revenue (Target)

Table 1: Descriptive Statistics for Features.

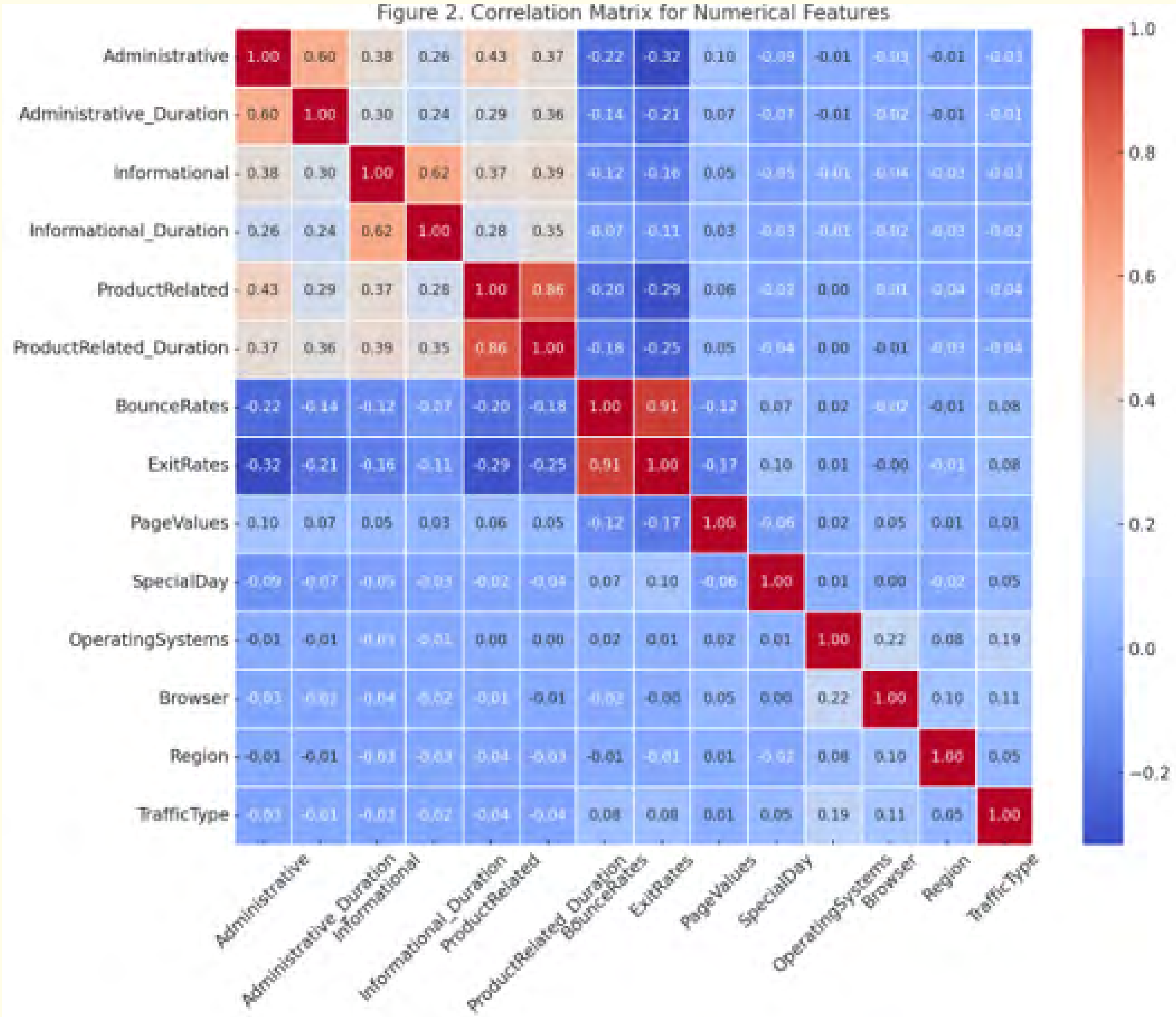


Figure 2: Correlation Matrix of Numerical Predictors for Online Shoppers' Purchase Satisfaction.

Descriptive statistical analysis of customer behavior features

Understanding the statistical characteristics of the dataset is an important first step in the modeling process that the future, as it provides insight into the variability and distribution of characteristics that affect purchasing behavior. Table 1 Summer the descriptive figures for independent variables (x22x_ {22} x22) and target variables, which indicates whether the purchase was completed (1) or not (0).

The analysis highlights several important patterns. Characteristics such as administrative, informative, and productive records relatively low mean value (2.31, 0.50, and 31.73, respectively). Combined with the fact that their first fourth (Q1) is equal to zero, it indicates that at least 25% of users did not reach these pages, suggesting diagonally distribution with several zero entries. Such sparsity may require preprosament strategies such as log transformation or bining to reduce the effects of extreme values.

In contrast, ProductRelated_Duration shows a much higher mean (1194.75) and a large standard deviation (1913.67), together with a wide interquartile range (Q1 = 184.14; Q3 = 1464.15). This indicates substantial variation in the time users spent on product-related pages, reflecting heterogeneous browsing behaviors and potential differences in purchase intent. While the high dispersion introduces noise, it may also serve as a valuable predictor of high-revenue sessions.

Behavioral indicators such as BounceRates and ExitRates exhibit low average values (0.02 and 0.04, respectively) with limited variability. Despite their small magnitude, these measures are behaviorally significant, as even minor increases in bounce or exit rates can strongly signal session abandonment or reduced purchase likelihood. Their narrow ranges suggest that careful normalization may be necessary to ensure stable model training.

Discrete and categorical features also play a key role. For example, Weekend is binary (0 or 1), while SpecialDay is ordinal, taking values between 0 and 1 (e.g., 0, 0.2, 0.4, ..., 1) to represent the proximity of the visit to a special occasion. Variables such as OperatingSystems and Browser are categorical identifiers with multiple classes. These variables require appropriate transformations (e.g., one-hot encoding) to enhance interpretability and to prevent distorted feature importance in classification models.

Finally, PageValues exhibits a mean of 5.88 and a large standard deviation of 18.56, reflecting a highly skewed distribution. While most sessions contributed no value, a minority of sessions accounted for disproportionately high values. This characteristic underscores the importance of this variable in predicting revenue-related outcomes, though standardization or log scaling may be necessary to reduce the influence of extreme outliers.

Overall, the descriptive statistics establish a clear picture of the dataset: while many features show sparse or skewed distributions, others capture substantial variability that can contribute meaningfully to predictive modeling of purchasing behavior.

The response variable Revenue is binary (0 = no purchase, 1 = purchase), with a mean of approximately 0.1547. This indicates a clear class imbalance problem, as only about 15.5% of sessions resulted in a purchase. Such imbalance necessitates the use of classification algorithms that can account for skewed class distributions, including Logistic Regression with class weighting, ensemble methods such as Random Forest or XGBoost, or resampling techniques like SMOTE. If left unaddressed, this imbalance may bias models toward the majority class (non-purchase) and reduce predictive performance.

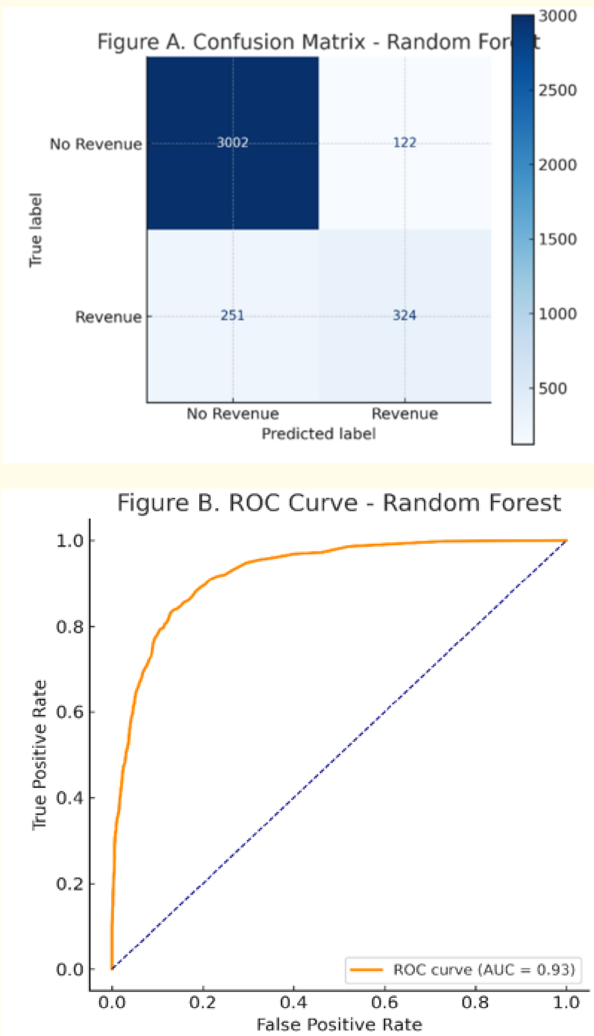
Another important observation from Table 1 is the substantial disparity in feature scales. Variables such as BounceRates and ExitRates are bounded between 0 and 1, while others, including ProductRelated_Duration and PageValues, span values in the thousands. This variation highlights the importance of normalization or standardization prior to training, particularly for scale-sensitive algorithms like K-Nearest Neighbors (KNN) or Neural Networks, to prevent high-magnitude features from dominating model learning.

Fourth data of IV further suggests that category identifiers such as region, smuggling, and operating systems are relatively balanced in their categories, which are no extreme oblique in their distribution. Since these variables are identifiers rather than continuous measures, they are more appropriately controlled through encoding techniques (eg, a-hot encoding) rather than relying only on descriptive figures.

In summary, descriptive analysis of Table 1 provides a clinical foundation for data preprosying by identifying major challenges, including slanting, outlair, square imbalance and asymmetrical feature scales. Addressing these issues is important for the creation of a strong and explanatory future of online purchasing behavior.

Figure 3 shows the evaluation diagnosis for the random forest model, including confusion matrix, ROC curve and lift curve. Confusion Matrix suggests that the model correctly classified 3002 true negative and 324 true positives, while 251 false negative and 122 false positivity is wrong. The ROC curve confirms the strong discriminatory potential of the model, acquiring AUC of 0.93, which indicates excellent separation between purchasing and non-cropped sessions. The lift curve displays the effectiveness of the model in identifying customers with a high probability of generating further revenue, underlining its utility to make target decisions in e-commerce applications.

To ensure a reliable evaluation, several machine learning algorithms from various families were applied, including logistic region, K-Nearest neighbor (KNN), Artificial Nervous Network (ANN), Random Forest and XGBOOST. These models were evaluated using a broad set of statistical and error-based matrix: meaning class error (MSE), means full error (MAE), R-Squard (RR AR), Root Medya Class Error (RMSE), Correlation coefficients (CC), coefficients of coefficients (COV), and error. Comparative results in Table 2 highlight business-bands in future stating performance, stability and generalization. In particular, random forests gained the lowest RMSE (0.3176) and consistently high accuracy, while XGboost equally produced competitive results, confirming the strengthening of modeling complex, non -relationship methods in non -relations.



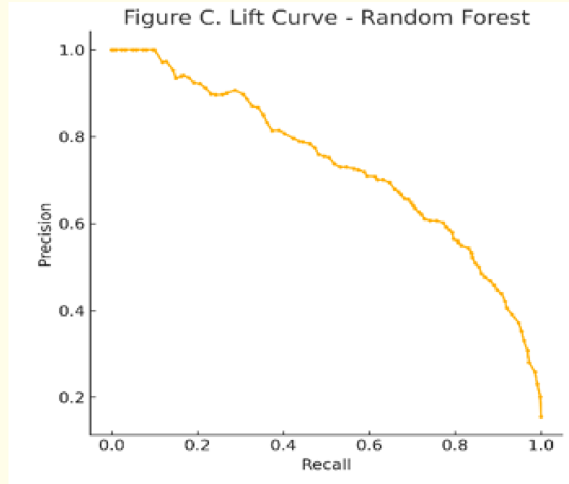


Figure 3: Diagnostic visualization of the Random Forest model: Confusion Matrix, ROC Curve, and Lift Curve for customer satisfaction prediction.

Accuracy	EC	COV	CC	R^2	MAE	RMSE	MSE	Model
0.8832	0.7513	3.5708	0.4717	0.1104	0.1168	0.3417	0.1168	Logistic Regression
0.8719	0.8243	3.4095	0.4184	0.0239	0.1281	0.3580	0.1281	KNN
0.8816	0.7617	2.5149	0.5270	0.0981	0.1184	0.3441	0.1184	Neural Network
0.8992	0.6487	2.7007	0.5835	0.2319	0.1008	0.3176	0.1008	Random Forest
0.8927	0.6904	2.5591	0.5675	0.1825	0.1073	0.3276	0.1073	XGBoost

Table 2: Machine Learning Evaluation Metrics.

Model evaluation metrics

The performance of the machine learning model was systematically evaluated using a set of statistical and error-based matrix. These matrix provide a complementary approach to future accuracy, stability, explanatory strength and reliability, allowing a broader comparison of an alternative classifier.

Mean squared error (MSE)

The MSE determines the average of the square difference between prediction and real results, more emphasis on large deviations. Low values are therefore indicating accuracy that gives better future. Among all the assessed models, random one classifier gained the lowest MSE (0.1008), outlining its better ability to reduce predicted errors. Conversely, the K-Nikat Saamasi

(0.1281) and Logistics Regression (0.1168) recorded high error levels, suggesting a low accurate estimate of purchasing results. These findings highlight the strong capacity of random forest in reducing large deviations that can significantly distort classification performance.

Mean absolute error (MAE)

MAE measures the average full difference between predictions and true values, considering all errors equally regardless of direction. Once again, the random forest model performed the best performance with a minimum MAE of 0.1008, which reflects low average deviation per prediction. In contrast, the artificial nerve network (0.1184) and the logistics region (0.1168) produced large errors, indicating low coherent prognosis accuracy. It confirms the stability of random forest in individual cases, as it maintains intimate alignment with real purchase results.

Coefficient of determination (R^2)

R^2 assesses the proportion of variance in the dependent variable explained by the independent features. A higher R^2 suggests stronger explanatory power and model fit. The Random Forest attained the highest R^2 (0.2319), showing it was able to capture more than 23.19% of the variance in purchasing behavior. This was followed by XGBoost (0.1825), which indicated moderate explanatory capability. By contrast, KNN lagged behind with a much lower R^2 (0.0239), reflecting limited ability to explain variability in customer purchase patterns. This result demonstrates that ensemble methods such as Random Forest provide greater flexibility in modeling complex, nonlinear relationships compared to linear techniques.

Root mean squared error (RMSE)

RMSE is the square root of MSE and, unlike MAE, it penalizes larger deviations more heavily. This makes RMSE particularly informative when substantial errors are more detrimental to predictive reliability. The Random Forest obtained the lowest RMSE (0.3176), confirming its robustness in minimizing significant deviations. XGBOOST and Logistic Regression followed with values of 0.3276 and 0.3417, respectively, while Logistic KNN trailed with weaker results. The evidence here consolidates the dominant performance of Random Forest in ensuring stability across diverse error scales.

Coefficient of variation (COV)

COV provides a normalized measure of variability by relating the standard deviation of predictions to their mean. It is particularly valuable in assessing the stability of model outputs across different samples. A lower COV indicates greater reliability. The Artificial Neural Networks model recorded the lowest COV (2.5149), demonstrating more stable predictions. Higher COV values were observed for Logistic Regression (3.5708) and KNN (3.4095), suggesting increased instability and reduced generalization power. This implies that ensemble approaches are better suited for producing consistent results in varying contexts.

Correlation coefficient (CC)

CC evaluates the strength and direction of the linear association between predicted probabilities and observed outcomes. A higher CC indicates stronger alignment between the two. The Random Forest achieved the highest correlation (0.5835), confirming its strong predictive alignment with actual purchasing behavior. XGBOOST showed a moderate CC of 0.5675, while Artificial Neural Networks (0.5270) and Logistic Regression (0.4717) demonstrated weaker associations, reflecting less dependable alignment between predictions and true labels.

Error coefficient (EC)

EC measures the magnitude of normalized prediction error, allowing for direct comparisons of reliability across models. Here, the Random Forest once more emerged as the most effective, producing the lowest EC (0.6487). This indicates that, relative to the scale of the data, its errors were minimal. By contrast, KNN (0.8243) and Artificial Neural Networks (0.7617) exhibited higher EC values, implying a greater likelihood of misclassification and cumulative deviation from true values.

Overall accuracy

Accuracy is one of the most intuitive and widely adopted metrics in classification tasks, representing the proportion of correctly classified observations. Both Random Forest and XGBoost achieved the highest accuracy (0.8992) and (0.8927) , highlighting their excellent ability to generalize predictions to unseen cases. Logistic Regression followed closely with 0.8832, while Artificial Neural Networks (0.8816) and KNN (0.8719) produced slightly weaker results. Although XGBoost matched Random Forest in accuracy, the latter's consistent superiority across almost all other evaluation metrics underscores its robustness.

Summary

Taken together, the evaluation metrics consistently demonstrate the dominance of the Random Forest model across multiple dimensions of performance. It achieved the lowest error measures (MSE, MAE, RMSE, COV, and EC) and the highest measures of fit and association (R^2 and CC). While Random Forest and XGBOOST tied in terms of accuracy, Random Forest outperformed XGBOOST and the other models on nearly all supplementary measures. These results establish Random Forest as the most effective and reliable model for predicting purchasing behavior within this dataset, providing strong evidence of its suitability for applications in e-commerce decision-making.

A comprehensive statistical evaluation was conducted to assess the distributional characteristics and predictive behavior of the five classification models: XGBoost, Random Forest, Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Logistic Regression. Table 3 presents the descriptive statistics (mean, standard deviation, standard error, median, and percentiles) for both the actual values and model predictions. The actual revenue variable shows a mean of 0.1547 and a standard deviation of 0.3618, which reflects the imbalanced binary nature of the dataset where only around 15% of sessions resulted in a purchase.

Across the models, ANN yielded the highest mean prediction value (0.3000) and the largest standard deviation (0.4606), indicating a tendency to over-predict the positive class (purchases) with substantial variability. In contrast, Logistic Regression produced the lowest mean prediction (0.1400) and the lowest standard deviation (0.3487), suggesting greater stability in its predictions but at the expense of under-representing the purchase class. These findings highlight substantial variation in how the models capture purchasing behavior.

The Analysis of Variance (ANOVA) results, summarized in Table 4, confirmed that the mean differences across models were statistically significant. Specifically, the test produced an F-statistic of 2.65037 and a p-value of 0.03265, which falls below the conventional 0.05 threshold. This indicates that at least one model exhibits a significantly different mean predictive behavior compared to the others, thereby justifying a deeper comparative analysis.

To complement these results, Root Mean Squared Error (RMSE) values were computed based on ten prediction samples, consistent with the Wilcoxon-based evaluation framework (see Figure 4). Within this reduced-sample comparison, Random Forest achieved the lowest RMSE (0.3593), confirming its superior accuracy in minimizing prediction errors. By contrast, Logistic Regression (0.4126) and KNN (0.4019) exhibited the highest RMSE values, underscoring larger deviations between predicted and actual values. These findings reinforce the robustness of Random Forest even under limited-sample evaluation.

The ANOVA heatmap in Figure 5 provides a visual interpretation of the statistical differences. Logistic Regression and KNN exhibited the largest deviations, with F-values of 127.3519 and 105.2443 and p-values of 0.0000, confirming highly significant differences. ANN recorded a moderate deviation ($F = 5.3150$, $p = 0.0212$), while XGBoost and Random Forest also showed meaningful but varied differences ($F = 7.9312$, $p = 0.0049$ and $F = 18.9516$, $p = 0.0000$, respectively).

However, when the Wilcoxon Signed-Rank test was applied to the same models, no statistically significant differences were detected between medians. This apparent contradiction can be explained by the different sensitivities of the two methods. ANOVA is designed to detect differences in group means and is highly sensitive to variance across samples, making it more likely to highlight differences when models produce outputs with distinct average levels. In contrast, the Wilcoxon test focuses on the ranking of paired differences and the alignment of medians, making it more robust against skewness, class imbalance, and outliers.

In this dataset, the imbalance in the target variable (Revenue = 1 in only ~15.5% of cases) and the high dispersion of some features (e.g., ProductRelated_Duration with SD > 1900) contribute to mean-level differences that ANOVA detects. However, because the median predictions across models remain relatively aligned, Wilcoxon does not register significant differences. From an academic standpoint, this does not represent a methodological flaw but rather illustrates how different statistical tests provide complementary insights: ANOVA reveals divergence in average predictive behavior, while Wilcoxon emphasizes consistency in central tendencies. Together, these findings suggest that although the models vary in their mean performance, their median predictive tendencies are statistically comparable.

XGBoost	Random Forest	ANN	KNN	Logistic Regression	Actual	Statistic
100	100	100	100	100	100	Number of values
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	Minimum
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	Percentile % 25
0.0000	0.0000	0.0000	0.0000	0.0000	0.5000	Median
1.0000	1.0000	1.0000	0.0000	0.0000	1.0000	Percentile % 75
1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	Maximum
1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	Range
0.2700	0.2700	0.3000	0.1800	0.1400	0.5000	Mean
0.4462	0.4462	0.4606	0.3861	0.3487	0.5025	Std. Deviation
0.0446	0.0446	0.0461	0.0386	0.0349	0.0503	Std. Error of Mean
27	27	30	18	14	50	Sum

Table 3: Statistical analysis.

P-value	F (DFn, DFd)	MS	SS	DF	Source of Variation
0.03265	2.65037	0.4670	1.868	4	Treatment (between columns)
—	—	0.1762	87.220	495	Residual (within columns)
—	—	—	89.088	499	Total

Table 4: ANOVA Test Results.

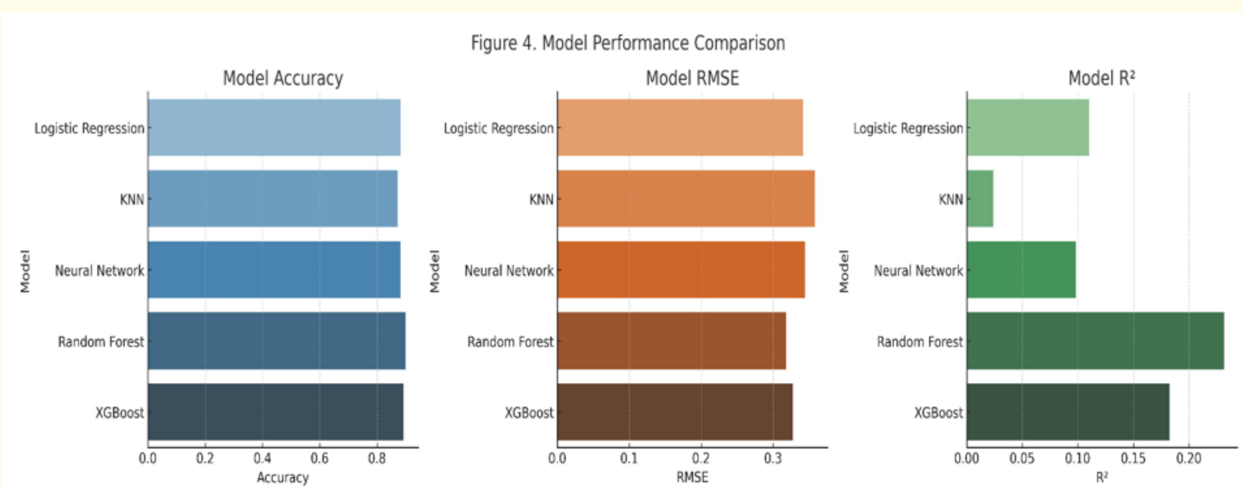


Figure 4: Visualization of Predictive Performance Across Models Using Accuracy, RMSE, and R^2.

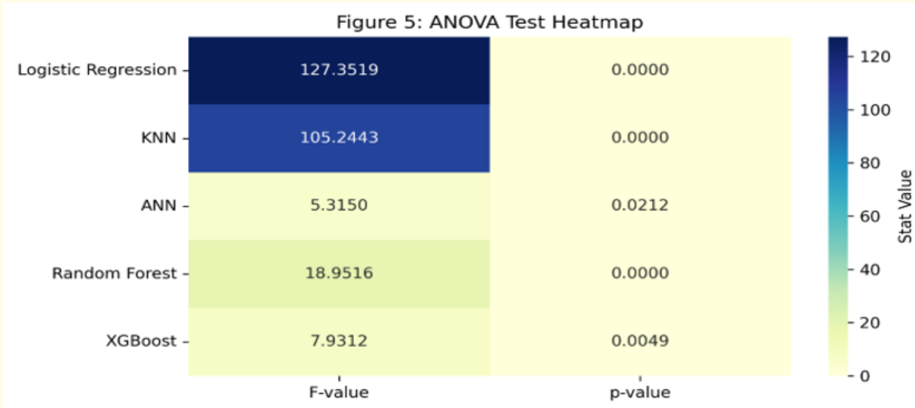


Figure 5: ANOVA Test Heatmap.

Discrepancy	?Significant	P-value summary	Exact or Estimate	P-value	W-stat	Number of values	Actual median	Theoretical median	Model
0.0	No	ns	Exact	0.250	0.0	10	0.0	0	Logistic Regression
0.0	No	ns	Exact	0.375	3.0	10	0.0	0	KNN
0.0	No	ns	Exact	0.625	2.5	10	0.0	0	ANN
0.0	No	ns	Exact	1.000	2.0	10	0.0	0	Random Forest
0.0	No	ns	Exact	0.625	2.5	10	0.0	0	XGBoost

Table 5: Wilcoxon signed rank test.

Statistical comparison and model evaluation

The joint consideration of ANOVA and Wilcoxon Signed-Rank results offers complementary insights into the comparative performance of the evaluated classifiers. While the ANOVA test (Table 4) identified statistically significant differences in mean predictive behaviors across models, the Wilcoxon analysis (Table 5) revealed no significant discrepancies in the medians of prediction distributions. This apparent divergence is not contradictory but rather reflective of the methodological differences between the two tests: ANOVA is sensitive to variations in means and assumes normally distributed residuals, whereas the Wilcoxon test is non-parametric and evaluates median alignment without distributional assumptions. Together, these findings suggest that although the models differ in their average predictive tendencies, their central distributions remain statistically consistent.

It is important to note that the RMSE values presented in Figure 6 differ slightly from those in Table 2 because they are derived from repeated subsample evaluations (n = 10). These values are reported solely for enabling ANOVA and Wilcoxon statistical comparisons, whereas Table 2 reflects the final test set performance. The RMSE distribution analysis (Figure 6) reinforces these conclusions by providing a granular perspective on error magnitudes across multiple samples. Random Forest consistently outperformed other classifiers with the lowest RMSE (0.3593), confirming its robustness in minimizing predictive errors. Logistic Regression and KNN, in contrast, demonstrated relatively higher RMSE values (0.4126 and 0.4019, respectively), indicating less reliable accuracy. ANN and XGBoost achieved moderate error levels, situating their performance between ensemble-based and simpler models. These results underscore the comparative strength of ensemble learning techniques in handling high-dimensional, nonlinear customer behavior data.

Finally, the residual diagnostics from the Quantile–Quantile (QQ) plot of the Random Forest model (Figure 7) revealed notable departures from normality, especially in the distribution tails. This outcome highlights the complexity and skewness inherent in online shopping behavior data and further justifies the reliance on non-parametric approaches—such as Wilcoxon—for robust model evaluation. Collectively, the integration of parametric and non-parametric statistical tools strengthens the interpretability of the analysis, ensuring that the comparative assessment of classifiers accounts for both mean-level differences and distributional consistencies.

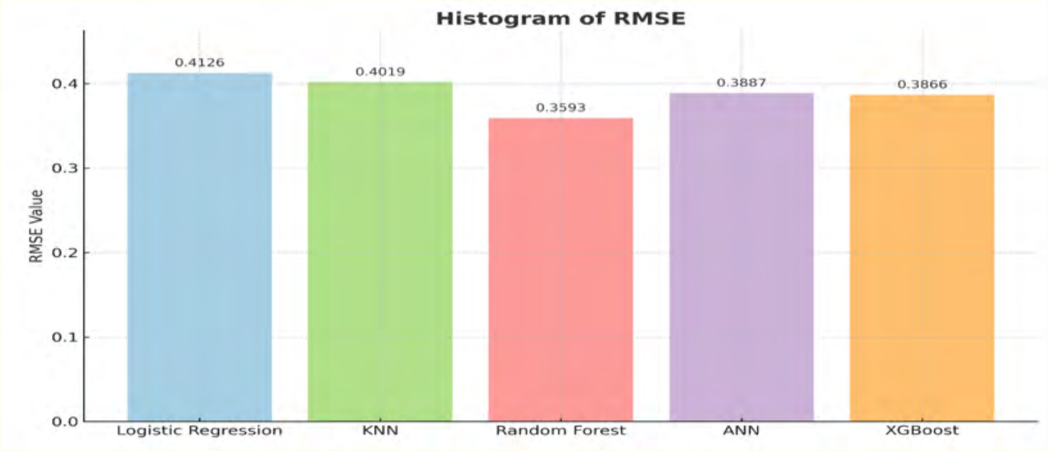


Figure 6: Histogram of RMSE.

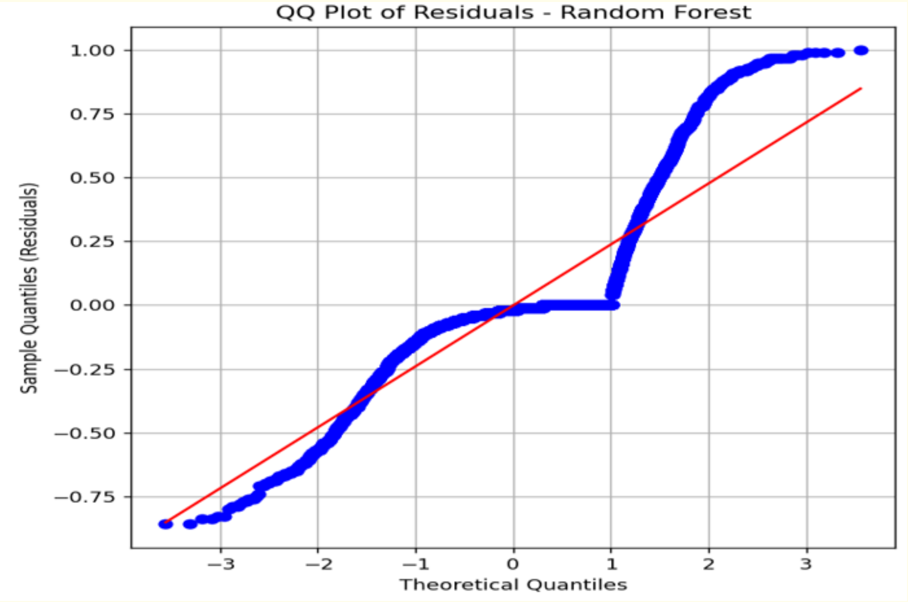


Figure 7: QQ Plot.

Conclusion

This study developed a comprehensive framework for predicting Revenue—used here as a proxy for purchase intention—within the retail and e-commerce sector by systematically evaluating five classification models: Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), Artificial Neural Networks (ANN), and Extreme Gradient Boosting (XGBoost). By employing a wide range of performance metrics, including RMSE, MAE, R^2, Accuracy, Correlation Coefficient, Coefficient of Variation, and Error Coefficient, the analysis revealed clear performance disparities across the models.

Between all classifiers, random forests demonstrated a consistent federal capacity, the lowest RMSE (0.3176), the highest correlation coefficient with real results (0.5835), and achieved the most favorable variability and error coefficient. While N in terms of random forest in terms of accuracy (89.92%), random forest proved to be the most stable and reliable overall, convenience balances future accuracy with interpretation through the importance analysis. This strength highlights its ability to catch the dynamics of non-behavior behavior while maintaining practical transparency to make decisions.

Beyond comparative accuracy, the results emphasized many major behavior predictions of the intention of purchasing. For example, strong correspondence between products and products between products and products indicates that the time spent in search of product materials is an important determinant of conversion. Similarly, high pages were firmly associated with revenue creation, while high bouncers and exits were aligned with session abandonment and reduced the possibility of purchase. These findings not only provide the future insight, but also provide actionable guidance for e-commerce operators that are willing to adapt the user engagement and reduce the drop-off.

While Ann and Xgboost offered a strong future, they required more intensive parameters tuning and computational resources. Conversely, logistic region and KNN provided simple and more explanatory models, but there was a lack of accuracy and strength obtained by the dress-based methods. Together, these comparisons confirmed that random forest provides the most effective balance of forecasting performance, computational efficiency and genuine -world rivalry in this context.

In summary, this research contributes methodologically by benchmarking diverse machine learning classifiers under rigorous statistical and error-based evaluation, and practically by offering insights into the behavioral attributes most associated with online purchasing. Future research can extend this framework by incorporating advanced resampling strategies to address class imbalance, exploring hybrid ensemble approaches, and applying the models across broader cross-industry datasets to enhance generalizability.

Limitations and Future Work

Despite the contributions of this study, several limitations must be acknowledged. First, the analysis relied exclusively on the Online Shoppers' Purchasing Intention Dataset, which captures behavioral data from a single e-commerce platform. This restricts the generalizability of the findings to other retail settings or industries with different consumer dynamics. Second, the dataset provides a cross-sectional snapshot of user sessions rather than longitudinal behavioral trajectories. As such, it may not fully capture evolving purchasing trends, seasonal effects, or shifts in consumer intent over time. Third, the Revenue variable exhibits a significant class imbalance (approximately 15% positive cases), which challenges the robustness of classification models. Although Random Forest demonstrated resilience in handling this imbalance compared to Logistic Regression or KNN, imbalance remains an inherent limitation that could bias model predictions toward the majority class.

Future research could address these issues through several avenues. The implementation of advanced data balancing strategies, such as SMOTE, ADASYN, or cost-sensitive learning, may further mitigate the skewed distribution of the target variable. The application of deep learning architectures, including recurrent neural networks (RNNs) for sequential behavior modeling or convolutional neural networks (CNNs) for feature extraction, could uncover richer and more complex behavioral dynamics. Additionally, hybrid ensemble frameworks that integrate Random Forest with gradient boosting or neural models may provide a stronger balance between predictive accuracy and stability. Expanding the analysis to multi-platform or cross-industry datasets would also enhance external validity and generate broader insights into consumer purchase behavior.

Finally, while a Q–Q plot of residuals was employed as a diagnostic tool in this study, it is recognized that this method is more appropriate for regression tasks, where residuals are continuous. In binary classification, alternative visualizations such as calibration plots, ROC curves, and Precision–Recall curves would offer more reliable insights into model calibration and performance under class imbalance. Accordingly, future work should integrate these specialized evaluation tools to strengthen the interpretability and robustness of classification outcomes.

Despite the strong performance of the models analyzed in this study, hybrid approach can play an important role in improving the forecast accuracy and balanced the strength of the model. A hybrid framework that integrates random forest with models such as gradient boosting and neural network can combine the strength of each technique. For example, random forest provides excellence and stability in handling square imbalance, while gradient boosting can refine the future accuracy, especially for tasks required by accurate decision limitations. Additionally, integrating nerve network with random forest can increase the ability of the model to catch complex patterns, such as sequential behavior in customer functions over time.

Including the hybrid model, the forecast may offer a better trade-off between accuracy and the model stability, addresses the challenges generated by the square imbalance and improves overall performance. This approach will help reduce the underlying bias towards the majority class in unbalanced dataset and can provide more reliable insight into the customer behavior pattern. In addition, these hybrid techniques can enable more comprehensive modeling of non-relations and dynamic consumer behavior, especially with the trend of fluctuations in the e-commerce environment.

Although the ANOVA and Wilcoxon signed-rank test was employed in this study to compare the performance of various classification models, it should be noted that the results of all models were obtained from the same dataset. This design implies that the model performance is not statistically independent, which violates one of the main beliefs of one-way Anova. While the additional use of the wilcoxon test helped partially address this limit by providing coupled comparison, future studies can be benefited by applying statistical tests that are more suitable for repeated-types, such as Freedman test or repeat-reiterate-olava, to ensure greater strong evaluation of model differences.

Bibliography

1. Abdullah-Al-Tanvir M., *et al*. “A gradient boosting classifier for purchase intention prediction of online shoppers”. (2023).

2. Armstrong JS and Collopy F. “Error measures for generalizing about forecasting methods: Empirical comparisons”. *International Journal of Forecasting* 8.1 (1992): 69-80.

3. Balasundaram E., *et al*. “A hybrid approach for customer segmentation and loyalty prediction in e-commerce”. *Prabandhan: Indian Journal of Management* 17.10 (2024): 56-69.

4. Bartroff J., *et al*. “Sequential experimentation in clinical trials: Design and analysis (Vol. 298)”. Springer (2012).

5. Benesty J., *et al*. “Pearson Correlation Coefficient. In Noise Reduction in Speech Processing”. Springer (2009).

6. Best H and Wolf C. “Logistic regression”. In The SAGE handbook of regression analysis and causal inference (2015): 153-171.

7. Bottou L. “Large-scale machine learning with stochastic gradient descent”. In Proceedings of COMPSTAT 2010 (2010). Physica-Verlag.

8. Breiman L. “Random forests”. *Machine Learning* 45 (2001): 5-32.

9. Cai K and Rodavia MR. “XGBoost analysis based on consumer behavior”. *Frontiers in Computing and Intelligent Systems* 6 (2023): 1-10.

10. Chai T and Draxler RR. “Root mean square error (RMSE) or mean absolute error (MAE)?” *Geoscientific Model Development* 7 (2014): 1247-1250.

11. Chen T and Guestrin C. “XGBoost: A scalable tree boosting system”. In Proceedings of the 22nd ACM SIGKDD (2016): 785-794. ACM.

12. Dey D., *et al*. “The proper application of logistic regression model in complex survey data: A systematic review”. *BMC Medical Research Methodology* 25 (2025): Article 15.

13. Dormann C F., *et al*. “Collinearity: A review of methods to deal with it”. *Ecography* 36.1 (2013): 27-46.

14. Ertan E and Akay K U. “Identifying a class of ridge-type estimators in binary logistic regression models”. *Statistics* 58.5 (2024): 1092-1116.

15. Everitt BS and Skrondal A. "The Cambridge dictionary of statistics". Cambridge University Press 4 (2010).
16. Friedman JH. "Greedy function approximation: A gradient boosting machine". *Annals of Statistics* 29.5 (2001): 1189-1232.
17. Friedman JH. "Stochastic gradient boosting". *Computational Statistics and Data Analysis* 38.4 (2002): 367-378.
18. Guyon I and Elisseeff A. "An introduction to variable and feature selection". *Journal of Machine Learning Research* 3 (2003): 1157-1182.
19. Hair JF., *et al*. "Multivariate data analysis (8th ed.)". *Cengage Learning* (2019).
20. Han J., *et al*. "Data mining: Concepts and techniques". *Morgan Kaufmann* (2012).
21. James G., *et al*. "An introduction to statistical learning". *Springer* (2013).
22. Kim S and Kim H. "A new metric of absolute percentage error for intermittent demand forecasts". *International Journal of Forecasting* 32.3 (2016): 669-679.
23. LeCun Y., *et al*. "Deep learning". *Nature* 521 (2015): 436-444.
24. Li Y., *et al*. "Customer online behavior analysis and purchase prediction in e-commerce". *Electronic Commerce Research and Applications* 40 (2020): 100935.
25. Midha M., *et al*. "Empathetic analytics: Understanding depression through AI". In APCIT 2024. IEEE (2024).
26. Neter J., *et al*. "Applied linear regression models". Richard D. Irwin (1983).
27. Pagan M., *et al*. "Investigating the impact of data scaling on the k-nearest neighbor algorithm". *Computer Science and Information Technologies* 4.2 (2023): 135-142.
28. Pedregosa F., *et al*. "Scikit-learn: Machine learning in Python". *Journal of Machine Learning Research* 12 (2011): 2825-2830.
29. Peng C Y J., *et al*. "An introduction to logistic regression analysis and reporting". *Journal of Educational Research* 96.1 (2020): 3-14.
30. Pham LT., *et al*. "Evaluation of random forests for short-term daily streamflow forecasting". *Hydrology and Earth System Sciences* 25 (2021): 2997-3015.
31. Qu Y., *et al*. "Product-based neural networks for user response prediction". In ICDM 2016. IEEE (2016).
32. Song P and Liu Y. "An XGBoost algorithm for predicting purchasing behaviour". *Tehnički vjesnik* 27.5 (2020): 1467-1471.
33. Sreesouthry S., *et al*. "Loan prediction using logistic regression". *Annals of the Romanian Society for Cell Biology* 25.4 (2021): 2790-2794.
34. Stoltzfus J C. "Logistic regression: A brief primer". *Academic Emergency Medicine* 18.10 (2011): 1099-1104.
35. Syaliman KU., *et al*. "Improving the accuracy of features weighted k-NN". *ICoSET* (2020): 326-330.
36. Willmott C J and Matsuura K. "Advantages of MAE over RMSE". *Climate Research* 30.1 (2005): 79-82.
37. Yang L., *et al*. "RF-LightGBM". arXiv (2021).
38. Zaghloul M., *et al*. "Predicting e-commerce customer satisfaction". *Journal of Retailing and Consumer* Services 79 (2024): 103865.
39. Zhang S., *et al*. "Traffic accidents severity using ordinal logistic regression". In ICAI 2024 (2024): 1007-1012.
40. Zhu S., *et al*. "Evaluation of random forests for streamflow forecasting". *Hydrology and Earth System Sciences* 25.6 (2021): 2997-3013.