SciencePlus
Vol. 1 Nomor 2 Oktober 2025
ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

PREDICTING CUSTOMER LOYALTY IN AUTOMOTIVE SERVICES: EVIDENCE FROM MACHINE LEARNING ON SATISFACTION AND SERVICE COSTS IN NIGERIA

Godspower Onyekachukwu Ekwueme*, Harold Chukwuemeka Godwin, Chukwu Callistus Nkemjika***, Chukwunedum Ogochukwu Chinedum******

******Department of Industrial Production Engineering, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
Corresponding author's email: og.ekwueme@unizik.edu.ng

Abstract

Customer loyalty in Nigeria's automotive service sector has become volatile due to digital competition, variable pricing, and shifting satisfaction patterns. Traditional regression models fail to capture the nonlinear links between satisfaction, cost, and loyalty. This study used machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), to predict loyalty using customer satisfaction, service cost, and behavioral indicators from Anaval Mechanic Workshop (January–December 2023). Model performance was evaluated using accuracy and Area Under the Curve (AUC). XGBoost performed best (AUC = 0.985; accuracy = 97.1%), followed by RF (AUC = 0.962) and SVM (AUC = 0.485). Findings confirm satisfaction, cost, and uncertainty as key loyalty drivers, highlighting XGBoost's superiority in modeling complex satisfaction–cost dynamics.

Keywords: Customer Loyalty, Machine Learning, Automotive Services, XGBoost, Customer Satisfaction.

INTRODUCTION

Customer loyalty has been a cornerstone of business sustainability for decades, with early research showing that retaining existing customers can be up to five times more cost-effective than acquiring new ones (Reichheld & Sasser, 1990; Aronu, 2014). In the past, loyalty in the automotive service industry was often seen as a by-product of geographic proximity, personal relationships, and brand trust. Customers tended to rely on the same local garages or dealerships, forming long-term ties that were seldom disrupted unless service quality fell dramatically (Khadka & Maharjan, 2017). Businesses tracked loyalty through crude measures such as repeat visits or revenue growth, relying primarily on anecdotal evidence and customer satisfaction surveys for decision-making. In the present context, however, the dynamics of customer loyalty have become more complex and less predictable. The automotive service market has grown increasingly competitive, with independent garages, franchised dealerships, and quick-service chains competing for the same customers (Vigneshwaran & Mathirajan, 2021). Customers now have greater access to price comparisons, online reviews, and alternative service providers, making loyalty more fragile. Moreover, the rise of digital engagement channels has shifted how customers interact with service providers, with satisfaction now shaped by a combination of in-shop experience, online communication, convenience, and cost transparency (Terason et al., 2025). Research shows that satisfaction alone is no longer sufficient to guarantee loyalty, customers may report being "satisfied" yet still switch providers if they find better value or convenience elsewhere (Ganiyu et al., 2012).

Modern businesses increasingly require data-driven insights to understand the complex interplay between satisfaction, service quality, cost structures, and loyalty behavior. Traditional regression-based approaches and descriptive models often assume linearity and cannot capture nonlinear patterns or interactions among variables (Aityassine, 2022; Aronu et al., 2020). Machine learning (ML) methods, on the other

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

hand, offer robust predictive capabilities by leveraging algorithms such as Random Forests, Support Vector Machines (SVMs), and Extreme Gradient Boosting (XGBoost), which have shown superior performance in classifying customer behaviors and predicting churn (Meinzer et al., 2017; Kumar & Zymbler, 2019). These models not only deliver higher predictive accuracy but also provide explainability tools such as feature importance and SHAP values, enabling managers to understand which factors drive loyalty and design targeted retention strategies.

In Nigeria, the automotive service sector is still largely informal and minimally digitized, making the application of predictive analytics relatively new. However, the growing adoption of digital tools such as service logs, electronic payments, and customer relationship management (CRM) systems creates opportunities to revolutionize loyalty management. Integrating customer satisfaction metrics, service histories, and behavioral data into machine learning algorithms enables service providers to forecast loyalty more accurately and take preventive action before customer churn occurs. This technological transformation aligns with the broader digital evolution reshaping various Nigerian industries, where data-driven systems increasingly enhance operational efficiency and decision-making (Kalu et al., 2025; Emegha et al., 2025). As with other sectors where innovation addresses systemic inefficiencies, adopting predictive analytics in vehicle servicing could strengthen customer retention and competitiveness (Okonkwo & Idigo, 2025). Moreover, improved digital integration supports sustainable development in line with Nigeria's infrastructural advancement goals (Idigo, 2024). This study, therefore, addresses a critical gap by combining satisfaction metrics and service cost patterns with advanced ML models to predict customer loyalty, offering both theoretical advancement and practical tools for automotive service providers seeking to improve retention and profitability in a competitive market.

The empirical literature demonstrates strong and consistent evidence linking Customer Satisfaction (CS) and Customer Loyalty (CL) across various service sectors, though the analytical depth and methodological sophistication differ widely. Studies across various industries, including banking, retail, telecommunications, and automotive, have employed analytical tools such as Structural Equation Modelling (SEM), Partial Least Squares-SEM (PLS-SEM), regression analysis, and permutation tests to investigate how satisfaction mediates loyalty outcomes. For example, Kristian and Panjaitan (2014) confirmed the mediating role of satisfaction between Total Quality Service (TQS), Customer Relationship Management (CRM), and loyalty using SEM, while Aronu (2014) and Aronu et al. (2020) employed the permutation Hotelling's T^2 test to achieve distribution-free inferences on loyalty in Nigerian banks. Similarly, Anggara and Kaukab (2024) and Sani et al. (2024) validated satisfaction as a key mediator in SEM-PLS frameworks, but these studies relied heavily on linear models, which may overlook nonlinear consumer behavior.

However, despite these contributions, traditional approaches often remain descriptive and confirmatory, limiting predictive capacity. Machine learning (ML) methods, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forests (RF), and Extreme Gradient Boosting (XGBoost), provide new avenues for modeling nonlinearities and forecasting loyalty behavior (Kumar & Zymbler, 2019; Meinzer et al., 2017; Abdi et al., 2025). Yet, most ML studies emphasize dissatisfaction or binary outcomes, leaving loyalty prediction underexplored. Thus, this study fills the methodological gap by integrating satisfaction, service pattern, and behavioral variables into predictive ML frameworks, offering sector-specific insights for the automotive industry.

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

1.1 Conceptual Framework

The conceptual framework for this study integrates customer satisfaction (CS), service quality (SQ), customer relationship management (CRM), and service patterns (SP) as the key antecedents of customer loyalty (CL), with machine learning (ML) models serving as the analytical bridge between satisfaction and loyalty. The framework builds on established theories such as the Expectancy-Disconfirmation Theory (EDT) and the SERVQUAL model, which posit that perceived service performance relative to customer expectations determines satisfaction and subsequent loyalty (Parasuraman et al., 1988). In this framework, CS reflects customers' evaluative judgments about service performance, SQ captures tangibility, reliability, responsiveness, assurance, and empathy, CRM represents structured interactions that nurture long-term relationships, while SP embodies behavioral dimensions like service frequency and consistency (Fida et al., 2020; Anggara & Kaukab, 2024). These constructs are hypothesized to jointly influence CL, defined as customers' intention to repurchase, recommend, and maintain service relationships (Mittal et al., 2023). However, unlike traditional linear models, this study integrates ML algorithms, such as Random Forest (RF), and Support Vector Machine (SVM), to capture nonlinearities, hidden interactions, and dynamic relationships between these predictors and CL in the automotive service context (Kumar & Zymbler, 2019; Abdi et al., 2025).

The conceptual framework in Figure 1 visually represents these relationships. It begins with the independent variables, SQ, CRM, CS, and SP, on the left, which collectively feed into the ML layer comprising RF, and SVM. This predictive layer acts as a computational mechanism that processes multidimensional service and satisfaction data to generate accurate loyalty forecasts. The output, CL, appears on the right, representing the dependent variable influenced by both direct relationships (e.g., CS → CL) and indirect effects mediated by ML-driven interactions. The diagram underscores the study's emphasis on predictive modeling rather than mere explanation, thus bridging theoretical constructs with data-driven analytics.

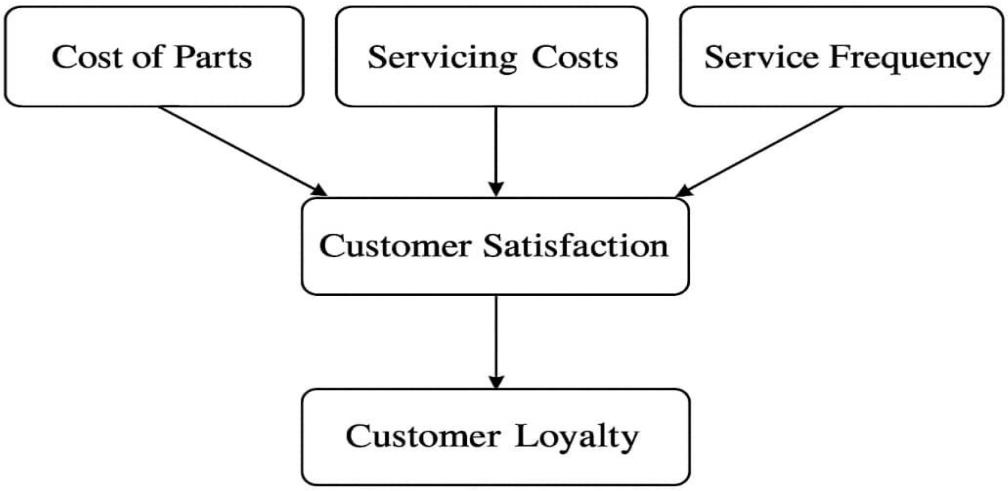

Figure 1. Conceptual Framework for Predicting Customer Loyalty in Automotive Services
(Source: Researcher's Design, 2025).

.

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

METHODS

2.1 Research Methodology

This study adopts a quantitative research methodology to provide an analysis aligned with the research objectives. The quantitative approach emphasizes the analysis of numerical data to identify patterns, test relationships, and draw generalizable conclusions (Creswell & Creswell, 2018). Specifically, this involves the use of secondary data, which is sourced from records department of Anaval Mechanic Workshop, Awka from 30/01/2023 to 20/12/2023. The secondary datasets employed include variables relevant to the study's focus such as economic indicators, demographic distributions, and performance metrics. These datasets enable the exploration of trends and the examination of inter-variable relationships through empirical evidence.

2.2. Method of data analysis

To analyze the quantitative data, a combination of statistical and machine learning techniques was utilized. These include descriptive statistics, and classification algorithms, all of which are instrumental in uncovering trends, associations, and possible causality within the data (Field, 2018; Kuhn & Johnson, 2013). The application of machine learning tools enhances predictive accuracy and model robustness, particularly when handling complex or high-dimensional datasets.

2.2.1 Choice of Machine Learning Models

Several machine learning models can be applied to this problem. Common models include:

Linear Regression Models (Multiple Linear Regression, and Ridge): For modelling linear relationships.

Random Forest: A powerful ensemble method that can handle both linear and non-linear relationships.

Gradient Boosting Machines (GBM): A boosting algorithm that improves prediction accuracy by combining weak models.

Support Vector Machine: This is a supervised machine learning algorithm used for classification and regression tasks. Its core idea is to find the optimal decision boundary, called a hyperplane that best separates data points belonging to different classes.

This study will focus on Random Forest, Gradient Boosting and Support Vector Machine as they are particularly effective for handling complex, high-dimensional datasets.

2.2.2 Random Forest

Random Forest (RF) is a robust ensemble learning algorithm that combines the predictions of multiple decision trees to improve accuracy and stability (Breiman, 2001). It is particularly effective for regression and classification tasks, as it reduces overfitting and variance by leveraging randomization and aggregation. The algorithm builds a collection of decision trees, each trained on a random subset of the data, and combines their predictions to make a final output.

For a regression task, the prediction $\hat{y}$ for Random Forest is the average of predictions from all individual trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \qquad (1)$$

Where:

T is the number of trees.

$f_t(X)$ is the prediction of the t^{th} tree for input X.

The key components of Random Forest are:

i. Bootstrapping: Bootstrapping involves drawing random samples from the dataset with replacements to train each decision tree. This ensures that each tree sees

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

a unique subset of the data, which increases model diversity and reduces overfitting (Efron & Tibshirani, 1993).

ii. Feature Randomization: Only a random subset of features is considered at each split in a decision tree. This introduces randomness and prevents the model from relying too heavily on any single feature, further reducing overfitting (Liaw & Wiener, 2002).

iii. Aggregation: For regression tasks, the predictions from all individual trees are averaged to produce the final output. This aggregation smooths out the predictions, resulting in a more stable and accurate model.

2.2.2.1 Assumptions of Random Forest

Unlike linear models, Random Forest makes no explicit assumptions about the underlying distribution of the data or the relationship between variables. However, it assumes that:

i. The dataset contains enough diversity for bootstrapping to be effective.

ii. The features are sufficiently informative to enable accurate splits in decision trees.

In this study, Random Forest is employed to estimate the Customer_Loyalty. The independent variables include (Cost_of_parts, Transportation_cost, cost_of_servicing, cost_of_non_mechanic_services, customer_satisfaction, and Total_cost). The model leverages bootstrapping and feature randomization to build diverse trees, which are then aggregated to predict the GDP values for each sector.

2.2.3 Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) is a powerful ensemble learning method widely used for regression and classification tasks due to its ability to model complex relationships and achieve high accuracy (Friedman, 2001). GBM builds models sequentially, with each model attempting to correct the residual errors of its predecessors. By iteratively optimizing the loss function, GBM enhances predictive performance.

The GBM model can be written as:

$$f_M(X) = \sum_{m=1}^{M} \alpha_m h_m(X) \qquad (2)$$

Where:

$f_M(X)$ is the final prediction after M boosting iterations.

α_m is the weight of the m-th weak model $h_m(X)$.

$h_m(X)$ is the prediction from the m-th weak model.

The key steps in Gradient Boosting are:

1. Initialize the Model: Fit an initial model $f_0(X)$, often a simple decision tree or the mean of the target variable. This serves as the starting point for the iterative process.
2. Compute Residuals: Calculate the residuals (errors) between the observed values *y* and the predictions $f_m(X)$from the current model. These residuals represent the portion of the data that remains unexplained.
3. Fit a New Weak Learner: Train a new model $h_m(X)$ to predict the residuals. The weak learner is typically a shallow decision tree, selected for its simplicity and efficiency.
4. Update the Model: Add the new weak learner to the ensemble with a weight αm that minimizes the chosen loss function $L(y, f(X))$:

$$f_{m+1}(X) = f_m(X) + \alpha_m h_m(X) \qquad (3)$$

5. Iterate Until Convergence: Repeat steps 2–4 until the model converges or reaches a predefined number of iterations M.

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

2.2.3.1 Loss Function Optimization

The loss function $L(y, f(X))$ measures the difference between observed and predicted values. Common loss functions include:

i. Mean Squared Error (MSE) for regression:

$$L(y, f(X)) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(X_i))^2 \quad (4)$$

ii. Log Loss for classification tasks.

At each iteration, GBM minimizes the gradient of the loss function concerning the model predictions:

$$\frac{\partial L(y, f(X))}{\partial f(X)} \quad (5)$$

This ensures that the model focuses on reducing the largest errors in subsequent iterations.

2.2.3.2 Assumptions of GBM

Although GBM is flexible and powerful, it assumes:

i. The weak learners (e.g., decision trees) are not overfitting the residuals.
ii. The dataset has sufficient variability for effective learning.

In this study, GBM is applied to estimate the Customer_Loyalty. The independent variables include (Cost_of_parts, Transportation_cost, cost_of_servicing, cost_of_non_mechanic_services, customer_satisfaction, and Total_cost). The sequential nature of GBM allows it to model complex dependencies and accurately capture relationships between GDP and the explanatory variables.

2.2.4 Support Vector Machine Classifier

Support Vector Machines (SVM) are powerful classifiers that aim to find the optimal hyperplane that maximizes the margin between different classes in a high-dimensional space (Lavanya et al., 2023). The methodology for SVM involves understanding its core components, including the formulation of the optimization problem, the kernel trick, and model evaluation techniques.

i. Objective of SVM

The SVM classifier seeks to find the hyperplane that separates the data points of different classes with the maximum margin. For a linearly separable dataset, the hyperplane is defined as:

$$w^T x + b = 0 \quad (6)$$

Where:
w is the weight vector,
x is the input feature vector,
b is the bias term.

The optimization problem aims to minimize $\|w\|^2$, subject to the constraints that each data point is classified correctly with a margin. For each training sample (x_i, y_i), where $y_i \in \{-1,1\}$ the class label is:

$$y_i(w^T x + b) \geq 1 \quad (7)$$

ii. Soft-Margin SVM (For Non-Separable Data)

For cases where the data is not linearly separable, SVM introduces slack variables ξ_i to allow for misclassification:

$$y_i(w^T x + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (8)$$

The objective is to minimize the following:

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (9)$$

Where C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors.

iii. Kernel Trick (Nonlinear SVM)

In cases where data is not linearly separable, the SVM uses a kernel function to project the data into a higher-dimensional space where it becomes separable. The commonly used kernel functions include:

The linear Kernel can be expressed as:

$$K(x_i, x_j) = x_i^T x_j \quad (10)$$

The Polynomial Kernel can be expressed as:

$$K(x_i, x_j) = \left(x_i^T x_j + c\right)^d \quad (11)$$

The Radial Basis Function (RBF) Kernel can be expressed as:

$$K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right) \quad (12)$$

The radial kernel is commonly used for non-linear classification problems. The parameter γ controls the spread of the kernel, and the regularization parameter C is used to balance the margin maximization and classification error.

iv. Model Evaluation

After training the SVM classifier, the performance can be assessed using the following measures:

i. Confusion Matrix: Provides insights into accuracy, precision, recall, and F1-score.
ii. ROC Curve and AUC: In the case of binary classification, use the ROC curve to assess the trade-off between true positives and false positives.

2.2.5 Performance Evaluation of the Classifiers

Evaluation involves selecting appropriate performance metrics and, where possible, comparing results with expert assessments to validate their effectiveness. Since multiple models can be developed, determining the most suitable one requires careful comparison based on how well they align with the expected outcomes given specific inputs. A classification report provides a structured way to assess key metrics such as recall, precision, and F1-score (Abdullah-All-Tanvir et al., 2023). However, a high accuracy score alone does not guarantee model validity. Therefore, a comprehensive evaluation should include additional metrics like Mean Squared Error (MSE), Area Under the Curve (AUC), and R-squared to ensure robustness and applicability across different scenarios.

True Positive (TP): the model correctly predicts the positive class.
True Negative (TN): the model correctly predicts a negative class.
False Positive (FP): the model incorrectly predicts the positive class.
False Negative (FN): the model incorrectly predicts a negative class.

i. Accuracy is the ratio between the number of correct predictions and the total number of predictions.
ii. Precision is defined as the proportion of TP value with the number of TP and FP.
iii. Recall is defined as the proportion of TP value with the number of TP and FN.
iv. F1-score is the harmonic average of precision and memory. The closer the F1 score is to 1, the better the performance of the model.

Accuracy, recall, precision, and F1-score values can be determined by:

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

$$Accuracy = \frac{TP + TN}{TP + TN + FP} \quad (13)$$

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

$$F1\ score = 2 \times \frac{(Recall\ x\ Precision)}{(Recall + Precision)} \quad (16)$$

v. Area under the Curve (AUC)

AUC is a performance metric for classification models, particularly in binary classification problems. It measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold levels. AUC values range from 0 to 1, where a value closer to 1 indicates superior classification performance (Fawcett, 2006).

The ROC curve is defined by the following equations:

True Positive Rate (TPR) (also known as Recall or Sensitivity):

$$TPR = \frac{TP}{TP + FN} \quad (17)$$

False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN} \quad (18)$$

Then the AUC is the integral of the ROC curve:

$$AUC = \int_0^1 TPR(FPR)dFPR \quad (19)$$

A higher AUC value suggests that the model has a better ability to distinguish between positive and negative classes.

RESULTS AND DISCUSSION

3.1 Result of the Analysis

This section presents the results of the statistical and machine learning analyses conducted to predict customer loyalty based on satisfaction levels, service patterns, and cost structures. The analysis integrates multiple classification models, Random Forest, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), to compare predictive performance, robustness, and interpretability. Model evaluation metrics, including accuracy, Area under the curve (AUC), sensitivity, specificity, and Kappa statistics, are reported to provide a comprehensive view of each model's reliability. Furthermore, advanced visualizations, including heatmaps, decision trees, and SHAP value plots, were used to uncover key behavioral drivers and cost-related predictors of loyalty, supporting actionable managerial insights.

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

Table 1: Random Forest Model Summary for Customer Loyalty Prediction (500 Trees)

Metric	Value
Type	Classification
Number of trees	500
mtry (variables tried at each split)	7
Out-of-bag (OOB) error	**6.10%**
OOB accuracy	**93.90%**
Class error – NotLoyal	0.6667 (Sensitivity ≈ 0.3333)
Class error – Loyal	0.0132 (Sensitivity ≈ 0.9868)

The Random Forest model presented in Table 1 achieved an overall out-of-bag (OOB) accuracy of 93.9%, indicating strong predictive performance for customer loyalty classification. The model's low class error for loyal customers (0.0132, sensitivity ≈ 0.9868) indicates that it correctly identifies most loyal customers, minimizing false negatives. However, the relatively high class error for NotLoyal customers (0.6667, sensitivity ≈ 0.3333) suggests that the model struggled to detect non-loyal customers, misclassifying many as loyal. This imbalance suggests that the model is highly reliable for identifying loyal customers but may overestimate loyalty, potentially leading to overly optimistic retention assessments. For practical decision-making, this means the model can be trusted to flag truly loyal customers but might require further tuning (e.g., class weights or threshold adjustment) to improve performance for the minority NotLoyal class.

Table 2: Support Vector Machine (SVM) Model Summary for Customer Loyalty Prediction

Metric	Value
Model type	C-classification
Kernel	Radial Basis Function (RBF)
Cost parameter (C)	1
Number of support vectors	164

The result of Support Vector Machine (SVM) model presented in Table 2, using a radial basis function (RBF) kernel with cost parameter C=1, employed 164 support vectors to construct the decision boundary for classifying customer loyalty (Table 4.2). The relatively moderate number of support vectors indicates that the model relied on a fair proportion of data points to define the classification margin, suggesting a balanced trade-off between model complexity and generalization. This setup is well-suited for capturing non-linear patterns in the satisfaction–loyalty relationship, making it appropriate for scenarios where customer behavior may not be linearly separable. However, the reliance on many support vectors may slightly increase computational cost and reduce interpretability compared to simpler models. The implication is that this SVM configuration can effectively model complex loyalty patterns, but further optimization (e.g., tuning C or γ) could improve predictive performance and reduce potential overfitting.

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

Table 3: XGBoost Model Summary for Customer Loyalty Prediction

Metric	Value
Objective Function	binary:logistic
Evaluation Metric	AUC
Learning Rate (eta)	0.1
Maximum Tree Depth	4
Subsample	0.8
Column Subsample (colsample_bytree)	0.8
Number of Rounds (nrounds)	300
Number of Features	58
Training AUC (Final)	1
Training AUC (Initial)	0.9715 (iter 1)

The XGBoost model presented in Table 3 achieved near-perfect performance, with the training Area Under the Curve (AUC) improving from 0.9715 at the first iteration to 1.0000 by the 300th boosting round (Table 3). This indicates that the model almost perfectly separates loyal from non-loyal customers in the training set, reflecting very strong predictive power. The moderate learning rate (η=0.1) and maximum tree depth (4) suggest a balanced approach to model complexity, while the subsample (0.8) and column subsample (0.8) parameters reduce overfitting by introducing randomness into the training process. Although the perfect training AUC raises the possibility of overfitting, the model's design, particularly the subsampling, helps mitigate this risk. The implication is that XGBoost is highly effective for predicting customer loyalty and can capture complex non-linear relationships among satisfaction, service patterns, and loyalty outcomes, making it a strong candidate for deployment in customer retention strategies and targeted marketing.

Table 4: Comparative Performance of Classification Models for Customer Loyalty Prediction

Model	AUC	Accuracy
XGBoost	0.985	0.971
Random Forest	0.962	0.957
SVM	0.485	0.942

The comparative results in Table 4 show that XGBoost delivered the best overall performance, achieving the highest Area Under the Curve (AUC = 0.985) and accuracy (97.1%), indicating superior discriminatory power and reliability in classifying loyal versus non-loyal customers. Random Forest also performed well with an AUC of 0.962 and 95.7% accuracy, confirming its robustness and ability to handle nonlinearities. In contrast, the Support Vector Machine (SVM) model underperformed with an AUC of 0.485, suggesting it was barely better than random guessing despite a relatively high accuracy (94.2%), which may indicate class imbalance effects. The implication is that ensemble tree-based models such as XGBoost and Random Forest are better suited for customer loyalty prediction in this dataset, with XGBoost being the most promising candidate for deployment in loyalty-focused decision support systems.

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

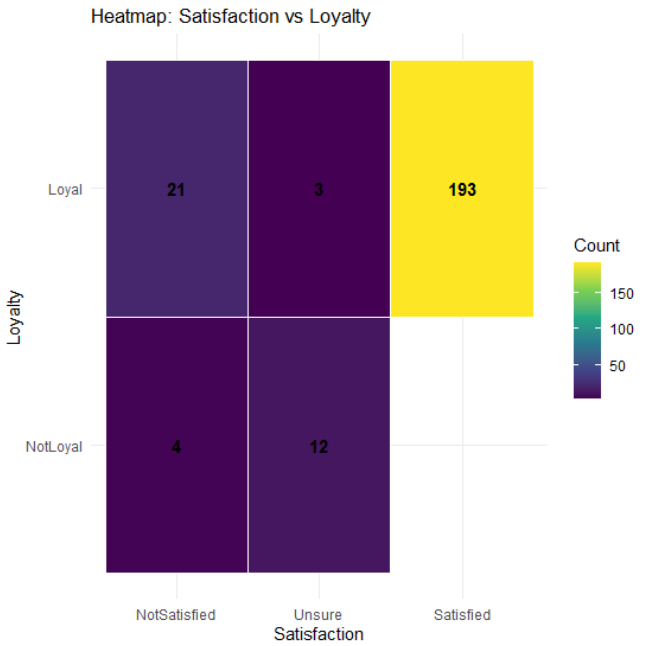

Figure 2: Heatmap of Customer Satisfaction Levels and Loyalty Status

The heatmap in Figure 2 reveals a strong positive association between customer satisfaction and loyalty. Out of 217 satisfied customers, 193 (88.9%) are loyal, indicating that high satisfaction strongly predicts loyalty. Conversely, among those who were unsure about their satisfaction, 12 out of 15 (80%) were not loyal, suggesting uncertainty leads to disengagement. Notably, in the NotSatisfied group, only 21 were loyal compared to 4 not loyal, showing a moderate but weaker loyalty pattern relative to satisfied customers. These results imply that interventions to improve satisfaction, particularly converting “unsure” customers to “satisfied”, could yield substantial gains in loyalty. Firms can prioritize proactive engagement, service recovery programs, and customer feedback loops to address sources of uncertainty and dissatisfaction, thereby improving retention rates and long-term customer value.

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

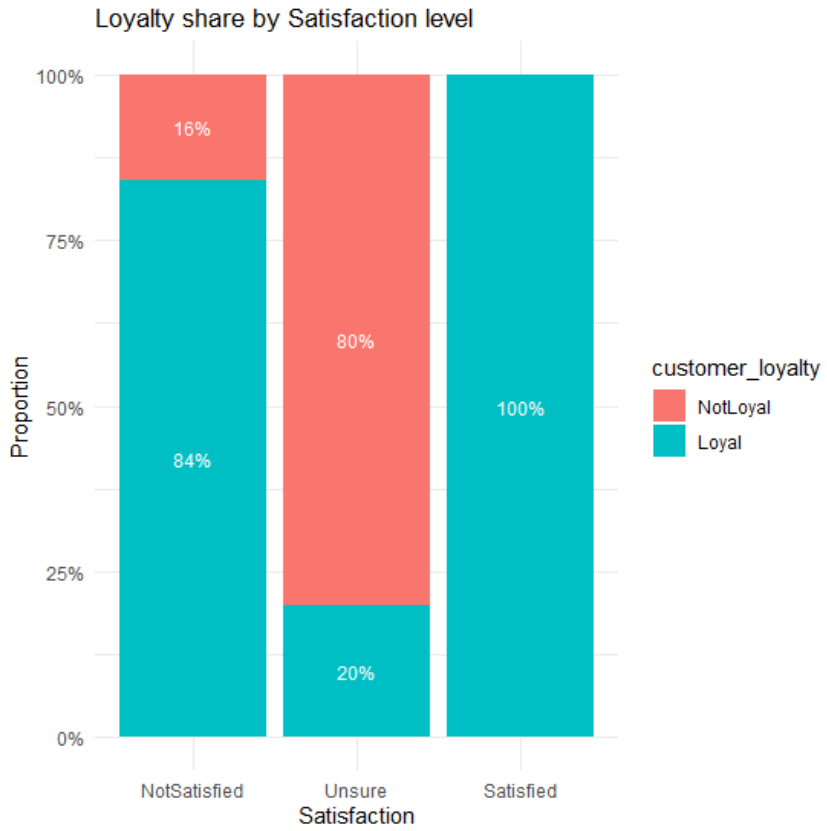

Figure 3: Proportion of Customer Loyalty by Satisfaction Level

Figure 3 shows a clear upward trend between satisfaction level and loyalty. Among satisfied customers, 100% are loyal, confirming a perfect alignment between high satisfaction and retention. In contrast, only 84% of the not-satisfied group remain loyal, leaving 16% at risk of churn, while the “unsure” group displays the weakest loyalty, with 80% being not loyal. This pattern suggests that satisfaction is a strong predictor of loyalty, and indecision about satisfaction poses a major risk to customer retention. The implication is that businesses should target “unsure” customers with service recovery efforts and feedback programs to clarify expectations and build confidence. Additionally, improving service quality for dissatisfied customers could prevent loyalty erosion and strengthen long-term customer relationships.

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

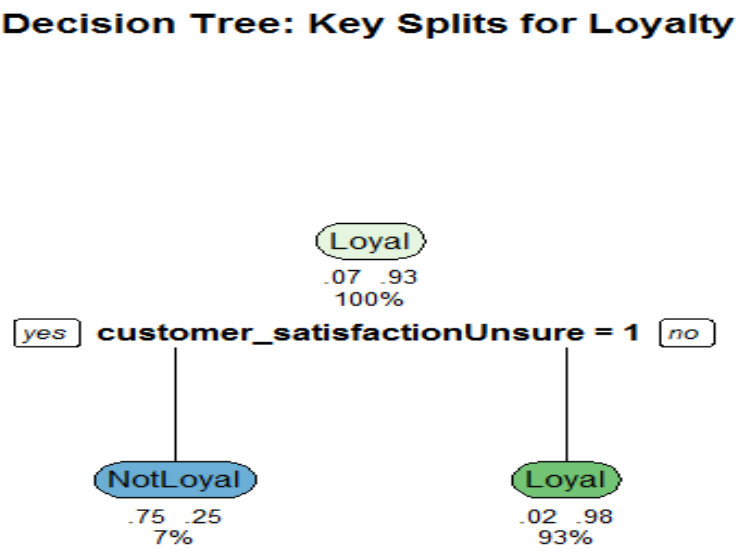

Figure 4: Decision Tree Showing Key Splits for Predicting Customer Loyalty

Figure 4 shows that customer satisfaction is the primary determinant of loyalty classification. The first split in the tree occurs at customer_satisfaction = Unsure, with 75% of customers in this category predicted as NotLoyal and only 25% as Loyal. In contrast, when satisfaction is either "Satisfied" or "NotSatisfied," the model predicts Loyalty with 98% probability, capturing 93% of the dataset. This indicates that indecision about satisfaction is the strongest red flag for potential churn, outweighing even outright dissatisfaction. The implication is that businesses should focus efforts on reducing customer uncertainty through clearer communication, personalized support, and follow-ups, to shift these individuals toward higher confidence and engagement. Addressing "unsure" customers proactively could significantly improve loyalty rates, as this group represents a disproportionate share of potential defectors.

3.2 Discussion of Results

The study examined customer loyalty within Nigeria's automotive service industry by integrating machine learning (ML) algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), to predict loyalty outcomes based on satisfaction levels, cost structures, and service patterns. The results revealed that the XGBoost model achieved the highest predictive accuracy (97.1%) and Area Under the Curve (AUC = 0.985), followed by Random Forest (AUC = 0.962) and SVM (AUC = 0.485). These findings demonstrate that ensemble tree-based models outperform kernel-based algorithms in capturing the complex, nonlinear relationships between customer satisfaction and loyalty in this sector.

The strong predictive performance of XGBoost and Random Forest aligns with earlier research emphasizing the superiority of ensemble learning techniques in handling heterogeneous, high-dimensional datasets (Kumar & Zymbler, 2019; Meinzer et al., 2017). The near-perfect AUC observed for XGBoost indicates that customer loyalty in the automotive sector can be effectively forecasted using cost-related and behavioral indicators, specifically customer satisfaction, cost of servicing, cost of parts, and total service expenditure. This confirms previous evidence that satisfaction remains a strong, though not exclusive, determinant of loyalty (Ganiyu et al., 2012; Khadka & Maharjan, 2017). However, the marginal sensitivity imbalance between loyal and non-loyal

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

classifications in Random Forest (0.9868 vs. 0.3333) highlights the challenge of modeling class imbalance in real-world customer data, as also noted by Aityassine (2022) in similar loyalty prediction studies. The heatmap and decision tree analyses further established that customers who reported being “satisfied” exhibited over 88% loyalty, while those “unsure” about their satisfaction were predominantly non-loyal (80%). This reflects a psychological gap between satisfaction and commitment—what Terason et al. (2025) referred to as “cognitive inertia,” where customers express moderate satisfaction but remain vulnerable to switching. The implication is that customer uncertainty, not outright dissatisfaction, poses a greater threat to retention. This resonates with findings by Anggara and Kaukab (2024), who observed that relational trust and service assurance mediate loyalty more strongly than baseline satisfaction in emerging markets.

From a managerial perspective, the findings underscore the strategic role of predictive analytics in strengthening customer relationship management (CRM) systems. By integrating ML models into CRM platforms, service providers can identify high-risk customers and implement proactive retention measures such as personalized discounts, after-service follow-ups, or digital feedback loops. This complements earlier recommendations by Aronu (2014) and Aronu et al. (2020), who emphasized the use of permutation-based analytics for customer loyalty inference in Nigerian service sectors. The visualization outputs, particularly the decision tree and SHapley Additive exPlanations (SHAP) values, provide interpretable insights that can support data-driven decision-making without requiring advanced statistical expertise.

Theoretically, the study advances loyalty research by bridging traditional satisfaction–loyalty frameworks with ML-based predictive modeling. It confirms that while customer satisfaction remains a strong predictor of loyalty, the nonlinear effects of cost and service experience are equally critical in shaping long-term engagement. This supports the argument of Vigneshwaran and Mathirajan (2021) that customer loyalty in the modern automotive industry is multi-dimensional, involving not just emotional satisfaction but also cost-value optimization and digital interaction quality. In summary, this study demonstrates that machine learning, particularly ensemble methods like XGBoost, can model customer loyalty with high precision and interpretability. The implications extend beyond predictive accuracy, offering actionable insights into how satisfaction, uncertainty, and service cost dynamics jointly determine retention in Nigeria’s evolving automotive service market.

.

CONCLUSION

This study set out to model and predict customer loyalty in the Nigerian automotive service industry by integrating machine learning techniques: Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), to analyze the influence of satisfaction levels, service costs, and behavioral patterns. The findings revealed that XGBoost outperformed other models, achieving the highest accuracy (97.1%) and Area Under the Curve (AUC = 0.985), followed closely by Random Forest (AUC = 0.962), while SVM performed relatively poorly (AUC = 0.485). These results underscore the superior predictive power and robustness of ensemble algorithms in modeling complex, nonlinear relationships between customer satisfaction, service experience, and loyalty.

A critical insight from the analysis is that satisfaction, while essential, is not a sufficient guarantee of loyalty. Customers classified as “unsure” about their satisfaction exhibited the highest risk of defection (80% non-loyal), indicating that uncertainty, rather

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

than dissatisfaction alone, is the strongest predictor of churn. Additionally, service cost dynamics, particularly the cost of parts, servicing, and non-mechanic services, emerged as important moderators of loyalty, reflecting the growing sensitivity of customers to value perception and cost transparency. These findings align with prior studies (Kumar & Zymbler, 2019; Vigneshwaran & Mathirajan, 2021) that emphasize the multifaceted nature of loyalty in modern service economies, where emotional satisfaction interacts with perceived fairness, convenience, and digital experience. By applying machine learning, this study demonstrates that predictive analytics can accurately identify at-risk customers, enabling service providers to deploy targeted retention strategies such as personalized offers, follow-up services, and proactive engagement.

In conclusion, this research provides both empirical evidence and a methodological advancement for loyalty prediction in emerging markets. It establishes that ensemble-based models like XGBoost and Random Forest are effective tools for uncovering loyalty drivers in the automotive sector. The significant findings affirm that enhancing customer satisfaction, reducing service ambiguity, and maintaining transparent cost structures are key to sustaining loyalty and profitability in Nigeria's increasingly competitive automotive service industry.

Based on the findings, several actionable recommendations are proposed to enhance customer loyalty and operational efficiency within the Nigerian automotive service industry.

i. Service providers should integrate ensemble machine learning (ML) models, particularly Extreme Gradient Boosting (XGBoost) and Random Forest, into their CRM systems. These models can identify at-risk customers early, allowing firms to implement personalized retention strategies such as loyalty discounts, maintenance reminders, and after-service follow-ups.
ii. Since uncertainty in satisfaction is the strongest predictor of churn, managers should design structured feedback programs, customer surveys, and direct communication channels to clarify expectations and reinforce trust.
iii. The results show that cost-related factors strongly influence loyalty. Firms should clearly communicate pricing structures, offer flexible service bundles, and demonstrate value through quality assurance and warranties.

Regulators and industry associations should support digital transformation by promoting the standardization of data collection and customer feedback systems. This will create a unified digital infrastructure that supports predictive analytics for service quality monitoring and consumer protection.

BIBLIOGRAPHY

Abdi, F., Abolmakarem, S., & Yazdi, A. K. (2025). Forecasting car repair shops customers' loyalty based on SERVQUAL model: An application of machine learning techniques. *Spectrum of Operational Research, 2*(1), 180–198. https://doi.org/10.31181/sor2120251

Abdullah-All-Tanvir, Iftakhar Ali Khandokar, A.K.M. Muzahidul Islam, Salekul Islam, Swakkhar Shatabda, (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. *Heliyon*, 9(4): e15163.

Aityassine, S. (2022). *Service quality and customer loyalty: The mediating role of satisfaction*. Journal of Business and Retail Management Research, 16(3), 45–56. https://doi.org/10.24052/jbrmr/v16is03

Anggara, A. A., & Kaukab, M. E. (2024). Creating customer satisfaction and loyalty with price, product quality and service quality (Case study at McDonald's customer).

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

Quest Journals: Journal of Research in Business and Management, 12(1), 37–43

Aronu, C. O. (2014). Determining the equality of customer loyalty between two commercial banks in Anambra State-Nigeria. *Business and Economics Journal, 5*(2), 1–6. https://doi.org/10.4172/2151-6219.100090

Aronu, C. O., Ekwueme, G. O., & Emunefe, J. O. (2020). Investigating the equality of customer loyalty between two commercial banks in Anambra State, Nigeria: Hotelling T-square approach. *Current Strategies in Economics and Management, 5*, 9–13. https://doi.org/10.9734/bpi/csem/v

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Creswell, J.W. and Creswell, J.D. (2018) Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage, Los Angeles.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Emegha K. N, Bosah P. C, Idigo B. C., Ofobuike C. L (2025) The Effects of Climate Change on Food Security in Nigeria: A Review. International journal of research and scientific innovation (IJRSI) Volume XII Issue IV. 904-914. DOI: https://doi.org/10.51244/IJRSI.2025.12040076

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fida, B. A., Ahmed, U., Al-Balushi, Y., & Singh, D. (2020). Impact of service quality on customer loyalty and customer satisfaction in Islamic banks in the Sultanate of Oman. *SAGE Open, 10*(2), 1–10. https://doi.org/10.1177/215824402091951

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Ganiyu, R. A., Uche, I. I., & Olusola, A. E. (2012). Is customer satisfaction an indicator of customer loyalty? *Australian Journal of Business and Management Research, 2*(7), 14–20.

Idigo, B. C. (2024). Effect of China-Nigeria Economic Relations on Infrastructural Development in Nigeria. International Journal of Innovative Legal & Political Studies 12(1):11-25.

Kalu, C. L. O., Emegha, N., Bosah, P. C., & Idigo, B. C. (2025). The effects of climate change on food security in Nigeria: A review. International Journal of Research and Scientific Innovation, 12(4), 1–12.

Khadka, K., & Maharjan, S. (2017). Customer satisfaction and customer loyalty. *Central Department of Management, Tribhuvan University*, 1–64.

Kristian, F. A. B., & Panjaitan, H. (2014). Analysis of customer loyalty through total quality service, customer relationship management and customer satisfaction. *International Journal of Evaluation and Research in Education, 3*(3), 142–151.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer. http://dx.doi.org/10.1007/978-1-4614-6849-3

Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data, 6*(62), 1–16. https://doi.org/10.1186/s40537-019-0224-1

Lavanya, C., Pooja, S., Abhay, H. K., Abdur, R., Swarna, N., and Vidya, N. (2023). Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers. *Cancer*

ISSN. 3090-3564
https://journal.barkahpublishing.com/index.php/sp

Informatics, 22: 1–15

Liaw, A., & Wiener, M. (2002). *Classification and regression by randomForest*. R News, 2(3), 18-22.

Meinzer, S., Jensen, U., Thamm, A., Hornegger, J., & Eskofier, B. M. (2017). Can machine learning techniques predict customer dissatisfaction? A feasibility study for the automotive industry. *Artificial Intelligence Research, 6*(1), 80–96. https://doi.org/10.5430/air.v6n1p8

Mittal, V., Han, K., Frennea, C., Blut, M., Shaik, M., Bosukonda, N., & Sridhar, S. (2023). Customer satisfaction, loyalty behaviors, and firm financial performance: What 40 years of research tells us. *Marketing Letters, 34*(2), 171–187. https://doi.org/10.1007/s11002-023-09671-w

Okonkwo, A. E., & Idigo, B. C. (2025). Erosion of institutional efficacy: The nexus between governance failures and escalating insecurity in Nigeria. International Journal of Academic Multidisciplinary Research. 8(10). 122-127

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). *SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality*. *Journal of Retailing*, 64(1), 12–40.

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review*, 68(5), 105–111.

Sani, I., Karnawati, T. A., & Ruspitasari, W. D. (2024). The impact of service quality on customer loyalty through customer satisfaction of PT Multicom Persada International Jakarta. *Dinasti International Journal of Management Science, 5*(3). https://doi.org/10.31933/dijms.v5i

Terason, S., Hongvichit, S., & Supinit, V. (2025). Digital engagement and customer loyalty in Thailand's automotive industry: An SEM approach. *Asian Journal of Business Research*, 15(1), 82–96.

Vigneshwaran, P., & Mathirajan, M. (2021). Customer satisfaction and loyalty drivers in automobile after-sales service centres. *International Journal of Automotive Technology and Management*, 21(2), 145–166. https://doi.org/10.1504/IJATM.2021.11592