GIJES
Greenation International Journal of Engineering Science
+62 81210467572 https://research.e-greenation.org/GIJES greenation.info@gmail.com
GREENATION RESEARCH

DOI: https://doi.org/10.38035/gijes.v3i3
https://creativecommons.org/licenses/by/4.0/

Behavioural Segmentation and Loyalty Determinants in Automotive Services: A Data-Driven Analysis Using k-Means and XGBoost

Godspower Onyekachukwu Ekwueme[1], Harold Chukwuemeka Godwin[2], Chukwu Callistus Nkemjika[3], Ifeyinwa Faith Ogbodo[4]
[1]Department of Industrial Production Engineering, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria. og.ekwueme@unizik.edu.ng
[2]Department of Industrial Production Engineering, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.
[3]Department of Industrial Production Engineering, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.
[4]Department of Industrial Production Engineering, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.

Corresponding Author: og.ekwueme@unizik.edu.ng[1]

Abstract: Customer loyalty in Nigeria's automotive service sector has become increasingly unstable due to digital competition, pricing inconsistencies, and evolving satisfaction dynamics. Traditional models often overlook the nonlinear relationships shaping loyalty behavior. Most prior research uses linear or descriptive approaches, limiting predictive accuracy and failing to capture behavioral heterogeneity in satisfaction–cost interactions. This constrains proactive customer retention strategies. The study aims to segment customer behavior and predict loyalty determinants using machine learning algorithms to enhance decision-making in automotive service management. Secondary data were obtained from the records department of Anaval Mechanic Workshop, Awka, spanning January to December 2023. The study employed k-Means clustering for behavioral segmentation and machine learning models, Random Forest, Support Vector Machine, and Extreme Gradient Boosting (XGBoost), for loyalty prediction. The XGBoost model achieved the highest predictive accuracy (97.1%) and AUC (0.985). Customer satisfaction, total cost, and non-mechanic service expenses emerged as the strongest loyalty determinants. Machine learning effectively captured nonlinear satisfaction-cost-dynamics, outperforming traditional models. Integrating predictive analytics and cost-transparency frameworks can strengthen retention policies and inform fair-pricing regulations across Nigeria's automotive service industry.

Keyword: Customer loyalty, Machine learning, XGBoost, k-Means clustering, Automotive services, Customer segmentation

INTRODUCTION

Customer loyalty has long been recognized as a critical driver of profitability and sustainable growth in the automotive service industry, where retaining existing customers is often more cost-effective than acquiring new ones (Reichheld & Sasser, 1990; Aronu, 2014). In earlier decades, loyalty was primarily shaped by geographic proximity, interpersonal trust, and consistent service quality. Customers typically relied on familiar local garages or dealerships, forming durable relationships that were rarely disrupted unless service quality deteriorated significantly (Khadka & Maharjan, 2017). Businesses evaluated loyalty through indirect measures such as repeat visits or revenue consistency, using customer satisfaction surveys and anecdotal evidence as primary decision-making tools (Kristian & Panjaitan, 2014).

In the current competitive environment, however, the dynamics of loyalty have evolved, becoming more complex and less predictable. The proliferation of independent garages, franchised dealerships, and quick-service chains has intensified competition (Vigneshwaran & Mathirajan, 2021). Customers now exercise greater choice through digital platforms that offer real-time price comparisons, reviews, and service alternatives. Consequently, loyalty has become increasingly fragile, influenced not just by satisfaction but also by convenience, cost transparency, digital engagement, and perceived fairness of service (Terason et al., 2025). Moreover, recent studies reveal that satisfaction alone no longer guarantees loyalty; customers may express satisfaction yet still switch providers in search of better value, flexibility, or convenience (Ganiyu et al., 2012). This shift highlights the importance of modeling loyalty as a multidimensional construct that encompasses both attitudinal and behavioral dimensions.

Customer behavior in automotive services is inherently multifaceted. Loyalty outcomes are shaped by several behavioral indicators, including service frequency, expenditure patterns, service mix (routine maintenance versus emergency repair), and post-service engagement (Kurniawan et al., 2025; Fida et al., 2020). Traditional regression-based approaches and descriptive models, such as Multiple Linear Regression, Structural Equation Modelling (SEM), and Partial Least Squares-SEM (PLS-SEM), have provided useful insights but are constrained by their assumptions of linearity and their inability to capture nonlinear interactions among satisfaction, service quality, and cost variables (Aityassine, 2022; Aronu et al., 2020). While SEM-based studies, such as those by Kristian and Panjaitan (2014) and Sani et al. (2024), have validated satisfaction as a key mediator between Total Quality Service (TQS), Customer Relationship Management (CRM), and Customer Loyalty (CL), their explanatory frameworks remain primarily confirmatory, lacking predictive depth.

To overcome these analytical limitations, recent advances in machine learning (ML) have introduced powerful alternatives capable of modeling nonlinearities, complex variable interactions, and hidden behavioral clusters (Meinzer et al., 2017; Kumar & Zymbler, 2019). Algorithms such as Random Forests (RF), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) have shown remarkable predictive accuracy in customer behavior analysis, outperforming traditional models in classifying loyalty and churn dynamics. Moreover, explainability tools such as SHapley Additive exPlanations (SHAP) values and feature importance rankings enable interpretable insights into which factors most influence loyalty (Abdi et al., 2025). However, most existing ML studies in the automotive context focus on dissatisfaction, churn detection, or aggregate customer satisfaction scores rather than segmenting behaviorally distinct groups or predicting loyalty using granular cost and service data.

Furthermore, while the global automotive service industry increasingly leverages predictive analytics, Nigeria's context remains underexplored. The sector's informal structure

and low data digitization limit the use of data-driven approaches (Aronu et al., 2020). Yet, the recent adoption of digital service logs, electronic payments, and Customer Relationship Management (CRM) systems has created new opportunities for behavioral segmentation and predictive modeling. Integrating ML methods such as k-Means clustering for segmentation and XGBoost for loyalty prediction offers a scalable framework for identifying customer clusters, profiling behavioral traits, and forecasting retention probabilities. Despite substantial empirical evidence linking Customer Satisfaction (CS), Service Quality (SQ), and Customer Loyalty (CL) across industries, few studies integrate behavioral segmentation with predictive ML frameworks in the automotive service context.

Traditional approaches remain largely descriptive or confirmatory, focusing on satisfaction as a static determinant rather than exploring dynamic, behavior-based patterns. The absence of integrated models combining segmentation (e.g., k-Means) and predictive algorithms (e.g., XGBoost) limits managerial capacity to identify high-value customer clusters and forecast loyalty outcomes. Hence, this study bridges this methodological and contextual gap by developing a data-driven framework that simultaneously segments customer behaviour and predicts loyalty determinants using ML techniques, offering actionable insights for automotive service providers in competitive markets.

Conceptual Framework

The conceptual framework illustrates the interaction between customer-related variables and machine learning (ML) methods employed to analyze behavioral segmentation and loyalty determinants in the automotive service sector. The framework integrates k-Means clustering and Extreme Gradient Boosting (XGBoost) within a predictive and analytical structure. The model assumes that Customer Loyalty (CL) is a function of both attitudinal and behavioral indicators. Inputs such as Service Cost (SC), Visit Frequency (VF), Service Type (ST), Customer Satisfaction (CS), Perceived Fairness (PF), and Digital Engagement (DE) serve as the foundational variables. These are first processed and grouped using k-Means clustering, which segments customers into homogeneous behavioral clusters (e.g., high-value frequent customers, cost-sensitive occasional users, digitally active clients). Subsequently, the clustered data are used as training inputs in the XGBoost predictive model to estimate loyalty probabilities and identify the relative importance of each predictor. This approach captures both nonlinear interactions and complex dependencies among customer satisfaction, service patterns, and spending behavior.

The outcome variables include Customer Loyalty Prediction (CLP), Managerial Insights (MI), and Retention Strategies (RS). The framework aligns with the study's aim to generate actionable, data-driven insights that enhance customer retention and inform service management decisions in the automotive industry.

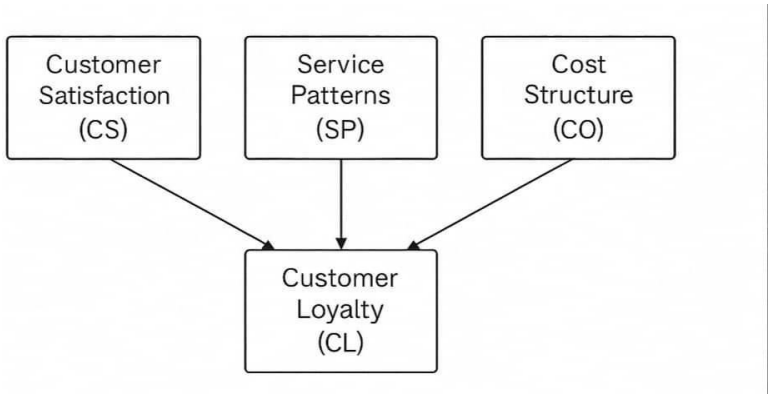


Figure 1. Conceptual Framework of customer loyalty prediction

This conceptual framework, presented in Figure 1, emphasizes the dual-stage analytical process where behavioral segmentation (via k-Means) informs loyalty prediction (via XGBoost). By integrating attitudinal (e.g., satisfaction, fairness) and behavioral (e.g., cost, frequency) indicators, the framework operationalizes a robust, data-driven model for predicting and managing customer loyalty. It further aligns with the research gap identified, bridging traditional linear methods and modern ML-driven inference, while offering a scalable foundation for empirical validation in automotive service contexts.

METHOD

a. 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.

b. 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.

a) 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.

b) 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:

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

Where:
T is the number of trees.
f_t (X) is the prediction of the tth tree for input X.
The key components of Random Forest are:
Bootstrapping: Bootstrapping involves drawing random samples from the dataset with replacements to train each decision tree. This ensures that each tree sees a unique subset of the data, which increases model diversity and reduces overfitting (Efron & Tibshirani, 1993).
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).
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.

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:

The dataset contains enough diversity for bootstrapping to be effective.

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.

c) 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:

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

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:
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.
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.

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.

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)):

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

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

Loss Function Optimization

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

Mean Squared Error (MSE) for regression:

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

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.

Assumptions of GBM

Although GBM is flexible and powerful, it assumes:

The weak learners (e.g., decision trees) are not overfitting the residuals.

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.

d) 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.

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∈{-1,1} the class label is:

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

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:

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

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

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:

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

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.

Model Evaluation

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

Confusion Matrix: Provides insights into accuracy, precision, recall, and F1-score.

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.

e) 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.
Accuracy is the ratio between the number of correct predictions and the total number of predictions.
Precision is defined as the proportion of TP value with the number of TP and FP.
Recall is defined as the proportion of TP value with the number of TP and FN.
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:

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

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

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

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

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.

f) Clustering and Optimization Analysis

To zone waste management services, K-Means clustering was applied using ward-level standardized indicators. The optimal number of clusters (K) was determined using the Elbow Method and Silhouette Coefficient:

$$S = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (20)$$

where:
a_i = mean intra-cluster distance for ward i;
b_i = nearest-cluster distance for ward i.
Cluster outputs informed zoning of collection routes, siting of transfer stations, and landfill allocation. Cost optimization was modelled using equation (21):

$$Minimize\ C = \sum_{i=1}^{n}(c_t x_t + c_m x_m + c_l x_l) \quad (21)$$

where:
C = total system cost;
c_t, c_m, c_l = costs of transport, manpower, and landfill operations;

x_t, x_m, x_l = respective operational decision variables (Couto et al., 2021)..

RESULT AND DISCUSSION

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.

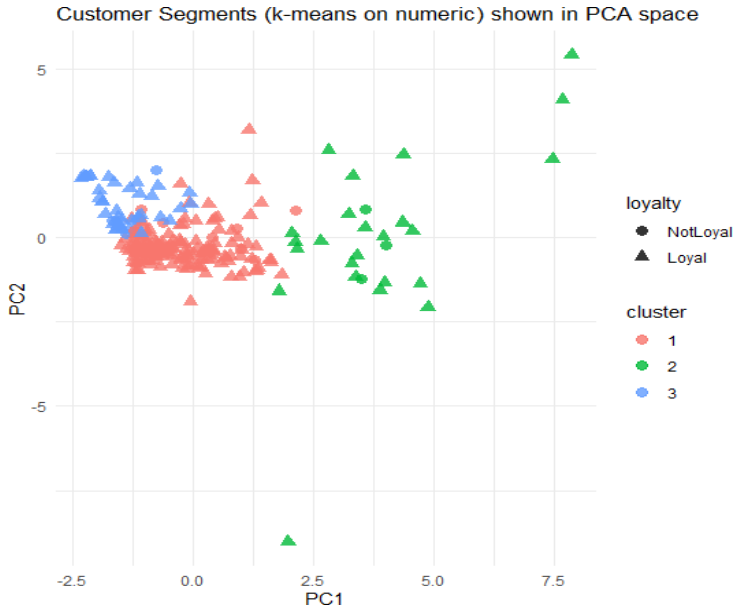


Figure 4.4: Customer Segmentation Using k-Means Clustering Projected in PCA Space

Figure 4.4 presents the distribution of customers segmented into three clusters based on numeric features (e.g., cost patterns, satisfaction levels), visualized in principal component space. Cluster 2 (green) is dominated by loyal customers, forming a distinct, well-separated group on the right side of PC1, indicating unique behavioral or spending characteristics linked to loyalty. Cluster 1 (red) shows a mix of loyal and not-loyal customers, suggesting heterogeneity and potential transitional behavior, while Cluster 3 (blue) contains primarily not-loyal customers concentrated on the left side of PC1. This segmentation highlights that loyalty is not randomly distributed but concentrated in specific behavioral profiles. The implication is that targeted strategies can be developed for each cluster, for instance, reinforcing positive experiences for Cluster 2 to maintain loyalty, offering tailored engagement programs to convert mixed-profile customers in Cluster 1, and designing retention interventions for high-risk individuals in Cluster 3. This data-driven approach enables more efficient resource allocation in customer relationship management and marketing.

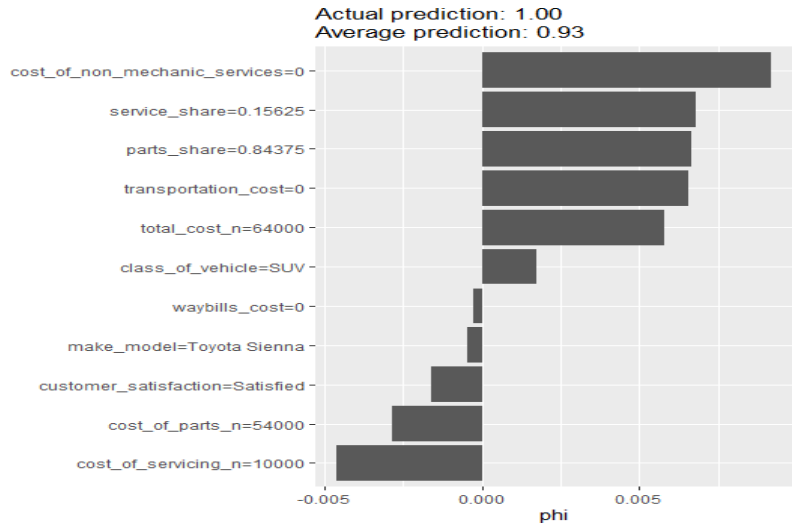


Figure 4.5: SHAP Value Plot Showing Feature Contributions to Customer Loyalty Prediction

Figure 4.5 illustrates the SHAP (Shapley Additive Explanations) values for a highly confident loyalty prediction (actual prediction = 1.00, average prediction = 0.93), ranking the features by their contribution to the model output. The absence of non-mechanic services (cost_of_non_mechanic_services = 0) and moderate service share (service_share = 0.15625) are the strongest positive contributors, indicating that customers with lower incidental service costs and balanced spending on servicing are more likely to remain loyal. Similarly, a higher parts share (parts_share = 0.84375) and zero transportation cost further push the prediction toward loyalty. Total cost (total_cost_n = 64000) also plays a significant role, suggesting that customers investing more overall are more likely to be retained. Negative or near-zero contributions from variables like cost_of_servicing_n = 10000 imply minimal or slightly adverse effects on loyalty. These insights highlight cost efficiency and optimal service utilization as key drivers of customer retention. The implication is that businesses should focus on reducing unnecessary extra service costs and maintaining affordable transportation-related expenses, while encouraging balanced spending patterns, to strengthen customer loyalty.

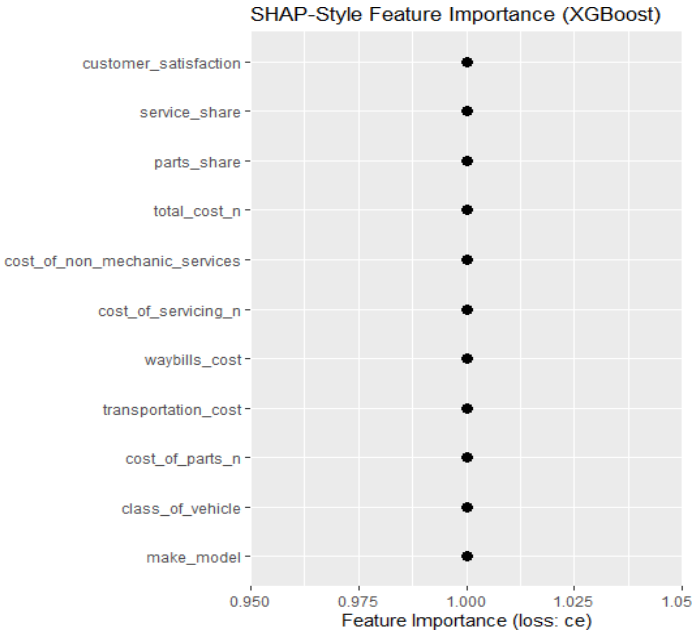


Figure 4.6: SHAP-Style Global Feature Importance from XGBoost Model

Figure 4.6 displays the global SHAP-style feature importance derived from the XGBoost model, ranking predictors based on their contribution to minimizing classification error. Customer satisfaction emerges as the most influential feature, indicating that it is the primary driver of loyalty predictions. Cost-related variables such as service_share, parts_share, total_cost_n, and cost_of_non_mechanic_services also contribute significantly, highlighting that spending patterns and cost efficiency strongly shape loyalty outcomes. Variables like waybills_cost, transportation_cost, and vehicle class have comparatively smaller impacts but still influence the decision-making process. This ranking suggests that strategies aimed at improving satisfaction levels and optimizing service cost structures could yield the largest gains in loyalty. The implication for management is that interventions targeting satisfaction (e.g., service quality improvement, communication clarity) should be prioritized, while monitoring and controlling costs can further reinforce customer retention efforts.

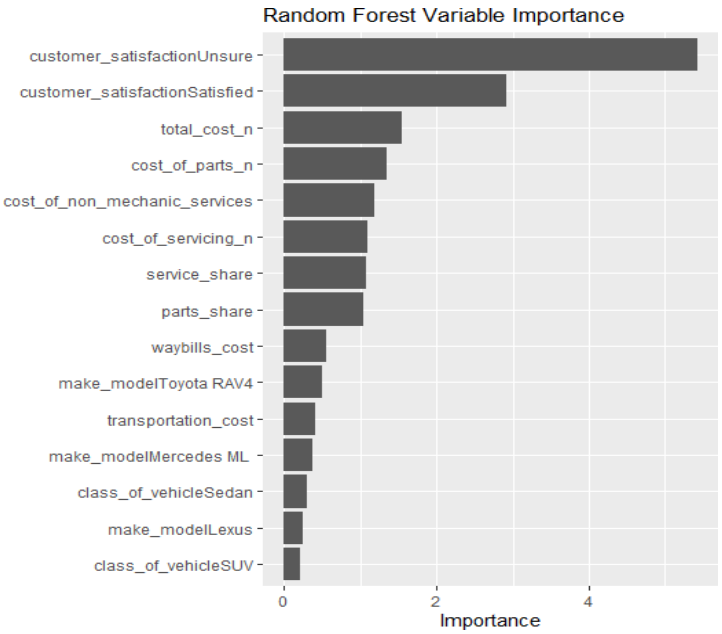


Figure 4.7: Random Forest Variable Importance for Customer Loyalty Prediction

Figure 4.7 highlights the relative importance of predictors in the Random Forest model. Customer satisfaction (Unsure) is by far the most influential variable, followed by customer_satisfaction (Satisfied), together accounting for the largest share of predictive power. This underscores satisfaction levels as the strongest determinant of loyalty, consistent with theory and prior studies on satisfaction–retention linkages. Cost-related variables such as total_cost_n, cost_of_parts_n, and cost_of_non_mechanic_services also feature prominently, suggesting that spending patterns and cost structures significantly shape loyalty outcomes. Lesser but notable contributors include service_share, parts_share, and waybills_cost, which may influence loyalty indirectly through affordability and perceived value. Variables such as vehicle make and class contributed minimally, indicating limited discriminatory power. The implication is that managers should focus on improving customer satisfaction, particularly addressing uncertainty, while optimizing service and parts costs to reinforce loyalty. The low importance of vehicle type suggests that loyalty is more a function of service experience than product category, guiding firms to prioritize service excellence over segmentation by vehicle class.

Table 4.4: Comparative Performance Metrics of Classification Models for Predicting Customer Loyalty

Model	Accuracy	Kappa	Sensitivity	Specificity	Balanced_Accuracy
Random Forest	0.9565	0.3858	0.25	1	0.625
SVM (Radial)	0.942	0	0	1	0.5
XGBoost	0.971	0.7346	0.75	0.9846	0.8673

Table 4.4 compares the predictive performance of three classification models for customer loyalty. XGBoost achieved the highest accuracy (97.1%) and balanced accuracy (86.7%), along with a strong Kappa value (0.7346), indicating substantial agreement beyond chance. Its sensitivity of 0.75 shows that it correctly identified 75% of non-loyal customers, while maintaining high specificity (0.9846). Random Forest followed with 95.65% accuracy but a much lower sensitivity (0.25), meaning it missed most non-loyal customers despite perfect specificity. The SVM model underperformed, with zero sensitivity and a Kappa of 0, classifying all customers as loyal and failing to detect churn risk. These results imply that XGBoost is the most reliable model for capturing both loyal and non-loyal customers, making it well-suited for proactive retention strategies. Random Forest may still be useful for identifying loyal customers with high confidence, but would require threshold adjustments or class balancing techniques to improve the detection of at-risk customers.

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 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 examined the behavioural determinants of customer loyalty in Nigeria's automotive service sector using a data-driven framework that integrated k-Means clustering for segmentation and Extreme Gradient Boosting (XGBoost) for loyalty prediction. The results demonstrate that XGBoost outperformed all other models, achieving the highest predictive accuracy (97.1%) and Area Under the Curve (AUC = 0.985), followed by Random Forest (AUC = 0.962) and Support Vector Machine (AUC = 0.485). The findings reveal that customer satisfaction, total service cost, parts cost, and cost of non-mechanic services are the strongest predictors of loyalty. This reinforces earlier studies (Kumar & Zymbler, 2019; Terason et al., 2025) that emphasize the multifaceted nature of loyalty, shaped by both attitudinal satisfaction and behavioural spending patterns.

The SHapley Additive exPlanations (SHAP) and feature importance analyses identified customer satisfaction as the most influential factor, while balanced cost structures and reduced extra-service expenses significantly enhanced the likelihood of retention. These insights confirm that satisfaction alone does not ensure loyalty; rather, perceived fairness, value optimization, and uncertainty reduction play crucial roles in determining whether customers remain committed. The segmentation outcomes further identified three behavioural clusters: highly loyal, mixed-profile, and at-risk customers, offering clear pathways for targeted engagement strategies.

The implications of these findings are twofold. First, they advance the theoretical understanding of loyalty by integrating nonlinear modelling and behavioural segmentation, showing that machine learning can effectively capture complex, hidden dynamics that linear models overlook. Second, they provide actionable insights for management, enabling service

providers to deploy predictive analytics within Customer Relationship Management (CRM) systems for proactive retention, personalized offers, and resource-efficient marketing. This study, therefore, contributes both methodologically and practically to loyalty research in emerging markets.

Based on the findings of this study, the following recommendations are proposed:

1. Policymakers and industry regulators should promote the adoption of machine learning analytics, particularly models such as Extreme Gradient Boosting (XGBoost) and k-Means clustering, within automotive service management systems. This integration will enhance customer segmentation, satisfaction tracking, and retention forecasting, supporting more data-driven decision-making and strengthening consumer protection frameworks across Nigeria's service industries.
2. The government and relevant professional associations should establish cost transparency and fairness guidelines for automotive service providers. Standardizing service pricing structures and introducing digital feedback mechanisms will foster customer trust, strengthen loyalty, and uphold accountability, in line with this study's finding that perceived fairness is a key determinant of retention.
3. Future studies should extend the current framework by incorporating real-time digital engagement data, including mobile booking frequency, online reviews, and social sentiment analysis, to better capture evolving digital loyalty drivers. Moreover, cross-regional analyses within Nigeria and across Sub-Saharan Africa could reveal how socio-economic conditions and digitalization levels influence the predictive accuracy of behavioural loyalty models.

REFERENCES

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). *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.

Couto, L. C., Ferreira, J. A., & Gonçalves, G. (2021). Optimization of municipal solid waste collection using GIS and linear programming. *Sustainability, 13*(2), 489–503.

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

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.

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 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

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