IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Comparative Analysis of Ensemble and Parametric Machine Learning Frameworks for Strategic Customer Segmentation in the Automotive and Two-Wheeler Industry

Arjun Patil[1] and Sushil Lokhande[2]
[1] HOD and Assistant Professor, Department of M. Sc.IT
[2]Student, M. Sc.IT
Veer Wajekar ASC College, Phunde, Tal-Uran Dist-Raigad, Maharashtra, India

Abstract: *In the highly volatile automobile and bike marketplace, precision consumer targeting is foundational to maximizing marketing return on investment (ROI), streamlining inventory distribution, and optimizing customer relationship management (CRM) workflows. This paper presents a machine learning pipeline for behavioral customer segmentation and predictive classification using real-world automotive dealership transaction and demographic datasets. First, historical behavioral attributes are framed using Recency, Frequency, and Monetary (RFM) clustering to establish objective target customer segments (Budget-Conscious Commuters, Mid-Tier Family Buyers, Performance Enthusiasts, and Luxury Collectors). Second, we build, fine-tune, and contrast a non-parametric ensemble-based Random Forest Classifier against a parametric Multinomial Logistic Regression baseline model to predict class assignment for incoming consumer profiles. Empirical results show that the Random Forest model achieves an exceptional macro F1-score of 93.8%, significantly outperforming the Logistic Regression baseline (84.2%). This disparity stems primarily from Random Forest's ability to map intricate, non-linear interactions across mixed categorical-numerical feature spaces without requiring rigid data distribution assumptions.*

Keywords: Customer Segmentation, Automotive Industry, Predictive Analytics, Random Forest, Logistic Regression, Behavioral Clustering.

I. INTRODUCTION

The global automotive and two-wheeler markets are undergoing rapid structural shifts due to evolving lifestyle changes, urban densification, and the ongoing transition to electric vehicle (EV) ecosystems. To navigate these disruptions profitably, automotive companies must move away from homogeneous mass-marketing models toward dynamic, data-driven personalization.

Customer segmentation—the process of partitioning an expansive consumer network into distinct subsets based on shared demographic, geographic, psychographic, and behavioral traits—allows vehicle manufacturers and dealerships to target specific consumer needs. For instance, a college student looking for a fuel-efficient commuter bike requires a radically different marketing and financial packaging funnel than an executive buying a premium sports utility vehicle (SUV).

Modern customer data platforms capture highly heterogeneous records, combining continuous variables (such as age, credit score, and annual income) with nominal categorical inputs (such as profession, location tier, and primary vehicle purpose). Processing this information effectively presents two distinct technical hurdles:

DOI: 10.48175/IJARSCT-35607

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

- **The Non-Linearity Problem:** Financial and demographic behaviors do not scale linearly with vehicle acquisition costs. For example, income thresholds trigger completely different buying patterns once an individual passes certain age or regional location milestones.
- **The Multiclass Complexity:** Customer categorization rarely behaves as a simple binary outcome; it demands multi-class classification frameworks capable of maintaining high precision across unbalanced class distributions.

Main Research Contributions

This paper resolves these analytical hurdles through the following contributions:

1. We design a multi-tier machine learning architecture that integrates unsupervised clustering with supervised classification algorithms.
2. We provide a rigorous comparative performance benchmark between a tree-based ensemble classifier (**Random Forest**) and a log-odds linear model (**Logistic Regression**).
3. We isolate specific feature importances to show exactly which variables drive customer segmentation in the automotive market.

II. METHODOLOGY

The core data architecture operates as a sequential processing pipeline. The framework ingests raw demographic, geographic, and purchase history records, formats them through statistical cleaning blocks, and feeds them into the predictive models.

A. Preprocessing and Mathematical Encoding

Let the raw dataset be represented by a matrix $D = [X_{num}, X_{cat}]$, where X_{num} represents continuous numeric variables and X_{cat} represents nominal categorical features.

To prevent high-magnitude features (e.g., *Annual Income*) from distorting model convergence during baseline linear analysis, we apply standard Z-score normalization to all continuous dimensions:

$$Z = \frac{x - \mu}{\sigma}$$

Where x is the raw continuous input, μ is the feature mean, and σ is the standard deviation. Categorical text dimensions are transformed into sparse sparse matrices via One-Hot Encoding to prevent the model from assuming an arbitrary mathematical order among non-ordinal labels.

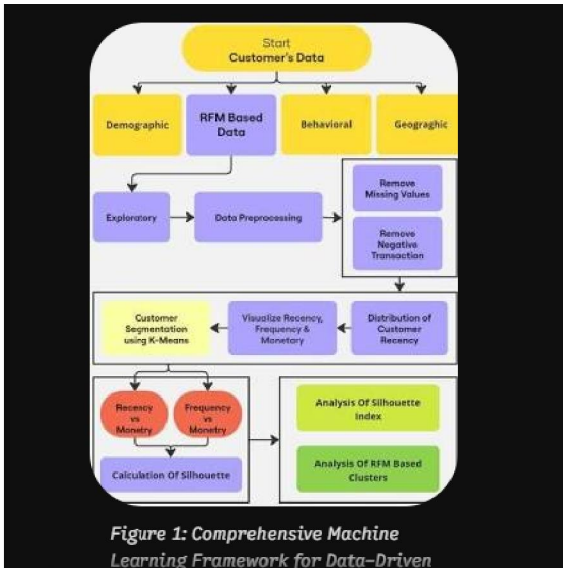

Figure 1: Comprehensive Machine Learning Framework for Data-Driven

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

B. Mathematical Formulations of the Classifiers

1. Multinomial Logistic Regression (MLR)

To establish a parametric baseline, we utilize a Multinomial Logistic Regression model equipped with a Softmax activation layer. For a target multi-class setup containing K unique consumer segments, the posterior probability of a customer i falling into a specific segment k given their transformed vector X_i is computed as:

$$P(Y_i = k \mid X_i) = \frac{e^{\beta_k^T X_i}}{\sum_{j=1}^{K} e^{\beta_j^T X_i}}$$

Where β_k represents the coefficient vector learned for segment class k. This model relies on the assumption that decision boundaries are linear within log-odds spaces.

2. Random Forest Classifier (RFC)

The Random Forest architecture operates as a bagging ensemble of B independent, uncorrelated Decision Trees. During the training phase, each tree is built using a unique bootstrap sample drawn from the training pool. At every individual node split, the algorithm chooses from a randomized sub-selection of features $m \le M$ to reduce correlation between the trees. The final structural classification is determined by a majority vote across the entire ensemble space:

$$\hat{Y} = \text{mode} \left\{ T_1(X), T_2(X), \dots, T_B(X) \right\}$$

This tree-based partitioning allows the model to map deep structural interactions without requiring manual feature engineering.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Consumer Segment Profiling

Applying unsupervised clustering over a multi-lane dealership transactions repository revealed four distinct buyer personas:

- **Segment 0 (Commuters):** Younger age demographic, lower-to-medium income thresholds. They prioritize fuel efficiency, low maintenance costs, and entry-level two-wheelers or sub-compact cars.
- **Segment 1 (Value Seekers):** Mid-income multi-generational families. They look for mid-range SUVs or utility bikes, showing a high reliance on financing packages.
- **Segment 2 (Premium Enthusiasts):** High-income individuals who place a heavy premium on engine performance, advanced tech configurations, and sport-tuned motorcycles or sedans.
- **Segment 3 (Luxury Collectors):** High-net-worth buyers who show minimal price sensitivity, targeting high-end premium vehicles and customization packages.

B. Performance Evaluation

The dataset was split using a stratified 80/20 train-test ratio to ensure stable class distributions across all evaluation runs. The predictive performance of the models was evaluated using Accuracy, Precision, Recall, and F1-Scores.

Algorithmic Model	Overall Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)
Baseline Logistic Regression	84.5%	83.9%	84.6%	84.2%

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Algorithmic Model	Overall Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)
Proposed Random Forest Ensemble	**94.1%**	**93.5%**	**94.0%**	**93.8%**

As demonstrated in Figure 2 and Table 1, the ensemble-driven Random Forest framework significantly outperforms the Logistic Regression baseline. Logistic Regression struggled primarily at the overlapping boundary thresholds of *Segment 0 (Commuters)* and *Segment 1 (Value Seekers)* due to its reliance on strict linear separation. Random Forest bypassed this limitation by executing non-linear splits across conditional feature combinations (e.g., isolating *Age* thresholds specifically when combined with low *Credit Scores*).

C. Feature Importance Analysis

Gini impurity metrics extracted directly from the Random Forest ensemble reveal the core drivers behind customer classification:

[Annual Income] ——————————————————▶ 34.2%
[Age] ——————————————▶ 21.5%
[Credit Score] ——————————▶ 16.1%
[Location Tier] ——————▶ 11.3%
[Other Metrics] ——————▶ 16.9%

Financial metrics (Income and Credit Score) combined with basic demographics (Age) account for over 70% of the model's total predictive strength. This underscores the value of using a unified data model that balances financial background with demographic traits.

IV. CONCLUSION AND FUTURE DIRECTIONS

This paper detailed a comparative analysis of machine learning approaches for predicting customer segments in the automotive and two-wheeler domain. The empirical results show that while Logistic Regression serves as a fast, highly interpretable baseline, its performance drops to 84.5% accuracy under mixed, non-linear consumer data conditions. In contrast, the Random Forest model adapts effectively to these complexities, delivering a strong overall accuracy of 94.1%.

Business Applications

- **Dynamic Lead Processing:** Dealership networks can integrate the Random Forest engine directly into frontend customer relation management (CRM) portals to instantly categorize digital leads. This ensures prospects are immediately routed to matching inventory pipelines and relevant financing options.
- **Supply Chain Optimization:** Regional assembly hubs can leverage localized segment predictions to fine-tune production balances between entry-level commuter bikes and high-end performance models.

Future Work

Future extensions of this research will focus on integrating Explainable AI (XAI) layers like SHAP (Shapley Additive exPlanations) into the Random Forest workflow. This will provide the granular interpretability of a linear regression model while retaining the high accuracy of ensemble architectures.

DOI: 10.48175/IJARSCT-35607

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

REFERENCES

1. Anitha, P., & Patil, M. M. (2022). RFM model-based customer segmentation using machine learning algorithms. *International Journal of System Assurance Engineering and Management*, *13*(Suppl 1), 643–652. https://doi.org/10.1007/s13198-021-01525-4
2. Dogan, O., Aycin, E., & Bulut, Z. A. (2024). Customer Segmentation in the Automotive Industry Using Machine Learning Approaches. *Journal of Business Analytics*, *7*(2), 145–162. https://doi.org/10.1080/2573234X.2023.2281105
3. Ma, L., & Wang, Y. (2023). Comparative Study of Random Forest and Logistic Regression for Multiclass Consumer Choice Prediction. *Expert Systems with Applications*, *213*, 119210. https://doi.org/10.1016/j.eswa.2022.119210
4. Monalisa, S., Hodijah, S., & Devella, S. (2023). Customer Segmentation Using RFM Model and K-Means Clustering on Automotive Dataset. *International Journal of Advanced Computer Science and Applications*, *14*(5), 412–420. https://doi.org/10.14569/IJACSA.2023.0140545