DIJEFA:
Dinasti International Journal of Economics, Finance & Accounting
E-ISSN: 2721-303X
P-ISSN: 2721-3021
https://dinastipub.org/DIJEFA dinasti.info@gmail.com +62 811 7404 455
DOI: https://doi.org/10.38035/dijefa.v6i3

Customer Retention Strategy through Churn Prediction in Four-Wheeled Vehicle After-Sales Services Using Big Data Analytics

Bella Puspa Dewani[1*], Athor Subroto [2]
[1]Universitas Indonesia, Salemba, Jakarta, Indoensia, bella.puspa@ui.ac.id
[2]Universitas Indonesia, Salemba, Jakarta, Indionesia, athor.subroto@ui.ac.id

[*]Corresponding Author: bella.puspa@ui.ac.id

Abstract: Customer churn prediction has become a critical aspect of business analytics, particularly in the automotive after-sales service industry. This study aims to develop an effective predictive model for identifying customers at risk of churn using big data analytics and machine learning techniques. The research focuses on four-wheeled vehicle after-sales services provided by Brand X, leveraging historical customer data over a seven-year period. Two machine learning algorithms Decision Tree and Random Forest were applied to classify churn behavior. Feature importance analysis was conducted to identify key variables influencing churn, including Warranty Status, Total Service Frequency, and Dissatisfaction Level. The models were evaluated using accuracy, sensitivity, specificity, confusion matrix, and feature importance metrics.. The findings suggest that integrating big data analytics with ensemble machine learning methods enhances churn prediction accuracy, enabling targeted customer retention strategies. This research contributes both academically and practically by providing a robust predictive framework for churn management in the automotive after-sales sector.

Keyword: After-sales service; Customer churn; Customer retention; Big data analytics;; Machine learning.

INTRODUCTION

In the increasingly competitive landscape of the automotive industry, customer retention has emerged as a critical strategic priority for businesses aiming to sustain profitability and long-term growth (Sharda et al., 2018). The cost associated with acquiring new customers often surpasses the investment required to retain existing ones, making churn prediction and subsequent intervention strategies not only economically beneficial but also operationally essential (Ascarza, 2018). Customer churn, defined as the phenomenon where consumers cease their engagement with a company's products or services over time, significantly affects revenue streams and brand loyalty, particularly in after-sales service sectors (Subroto, 2020).

After-sales service plays a pivotal role in shaping customer satisfaction and brand perception, especially in industries such as automotive, where continued engagement is necessary post-purchase (Durugbo, 2020). High-quality after-sales support—including regular

maintenance, warranty claims handling, repair services, and responsive customer care—has been identified as a key driver of customer loyalty and repeat business (Nyadzayo & Khajehzadeh, 2016). However, many companies still struggle to effectively monitor and predict customer disengagement, leading to missed opportunities for proactive retention efforts.

The rapid advancement of big data analytics and machine learning technologies has revolutionized how organizations approach customer relationship management (Camm et al., 2019). These tools enable firms to process vast volumes of structured and unstructured data, uncover hidden behavioral patterns, and build predictive models that anticipate future customer actions with high accuracy. In particular, supervised machine learning algorithms such as Decision Tree and Random Forest have demonstrated superior performance in churn prediction tasks across various domains, including telecommunications and e-commerce (Abdullah-All-Tanvir et al., 2023; Prabadevi et al., 2023).

Despite these advancements, the application of such techniques within the Indonesian automotive sector remains underexplored. Many local businesses, especially small and medium-sized enterprises, lack the analytical capabilities to leverage big data for customer insights, resulting in reactive rather than proactive customer management strategies (Subroto, 2020). This research addresses this gap by focusing on four-wheeled vehicle after-sales services provided by Brand X, a major player in the Indonesian automotive market. By analyzing historical customer data spanning seven years, this study develops and evaluates several machine learning models to identify customers at risk of churn and proposes actionable retention strategies based on model outputs.

Moreover, this study contributes to both academic literature and practical applications in the field of customer analytics. Academically, it expands the body of knowledge regarding the use of ensemble machine learning methods in churn prediction within the automotive domain. Practically, it provides Brand X with a robust predictive framework that enables real-time decision-making, personalized marketing campaigns, and optimized service delivery. The findings of this research can serve as a blueprint for other automotive companies seeking to implement data-driven customer retention strategies in emerging markets.

METHOD

This study was conducted with the objective of developing a customer retention strategy through churn prediction in four-wheeled vehicle after-sales services by leveraging big data analytics and machine learning techniques. To achieve this goal, a systematic and structured research methodology was employed, encompassing data collection, preprocessing, model development, performance evaluation, and the formulation of actionable retention strategies. The overall approach was designed to ensure robustness, reproducibility, and practical applicability for automotive service providers, particularly Brand X operating in the Indonesian market (Subroto, 2020).

The process began with the acquisition of historical customer data related to vehicle maintenance and repair services over a seven-year period. This dataset was sourced from internal company records and included various attributes such as vehicle type, warranty status, service frequency, labor cost, and customer dissatisfaction level. Each customer interaction was recorded in detail, allowing for a comprehensive view of behavioral patterns over time. Given that customers who performed multiple services contributed multiple entries to the dataset, the final dataset comprised a total of 398,695 records, which served as the foundation for both descriptive and predictive analysis (Thangeda et al., 2024).

To prepare the data for modeling, several preprocessing steps were undertaken. First, inconsistencies and missing values were identified and addressed through appropriate imputation or removal techniques. Categorical variables were encoded using one-hot encoding,

while numerical features were normalized to ensure uniformity across the dataset (Camm et al., 2019). Subsequently, each customer was classified into either retained or churned categories based on the time interval between consecutive service visits. Specifically, customers who had not returned for service within 12 months of their last visit were categorized as churned (labeled "1"), while those who continued to return within the 12-month window were considered retained (labeled "0"). This classification formed the dependent variable for the supervised machine learning models developed in the subsequent stages (Ascarza, 2018; Murtiningrum et al., 2022).

Once the dataset was prepared, it was partitioned into training and testing subsets using varying test sizes ranging from 20% to 60%. This variation allowed for an assessment of how different proportions of test data influenced model performance. Two machine learning algorithms were selected for model development: Decision Tree and Random Forest. These algorithms were chosen due to their proven effectiveness in handling classification tasks, especially in scenarios involving large datasets and imbalanced classes—common characteristics of churn prediction problems (Abdullah-All-Tanvir et al., 2023; Omotehinwa et al., 2024).

Each algorithm was implemented with default hyperparameters initially, followed by iterative tuning using Bayesian optimization techniques to enhance model accuracy and generalization. The Decision Tree algorithm was used to generate interpretable rules by recursively splitting the data into increasingly homogeneous subsets. Random Forest built upon this by creating an ensemble of decision trees trained on bootstrapped samples of the data, combining predictions through majority voting to reduce variance and overfitting (Prabadevi et al., 2023).

Model performance was evaluated using a range of metrics including accuracy, sensitivity, specificity, confusion matrix, and feature importance analysis. These metrics provided insights into how well each model could distinguish between churned and retained customers. Feature importance analysis was conducted using the Random Forest model to identify the most influential predictors of churn, such as Warranty Status, Total Service Frequency, and Dissatisfaction Level. These findings were crucial in informing targeted retention strategies tailored to the specific needs of at-risk customer segments (Leong et al., 2024; Malik et al., 2025).

In addition to quantitative model evaluation, a flow diagram was developed to illustrate the entire research process from data collection to strategy formulation. This visual representation helped ensure clarity in understanding the sequence of activities and facilitated communication of the methodology to stakeholders involved in the implementation phase (Sharda et al., 2018).

By integrating big data analytics with advanced machine learning techniques, this research aimed not only to predict customer churn but also to provide actionable insights that could be directly translated into effective customer retention initiatives. The methodology described here forms the backbone of the empirical findings and recommendations presented in the subsequent sections of this paper.

RESULTS AND DISCUSSION

This chapter presents the results of data analysis and model development conducted to predict customer churn in four-wheeled vehicle after-sales services using big data analytics. The dataset used in this study was sourced from Brand X's internal records and included historical customer service data spanning seven years (2018–2024). After preprocessing and feature engineering, a total of 398,695 customer records were analyzed for both descriptive and predictive purposes.

The raw dataset consisted of detailed service history entries for each customer visit to authorized dealers. Each customer could appear multiple times in the dataset depending on the number of service visits they made over time. To prepare the dataset for modeling, several preprocessing steps were applied:
Missing value handling : Inconsistent or missing values were either imputed or removed.
Categorical encoding : Categorical variables such as VehicleType , WarrantyStatus , and Dissatisfaction were encoded using one-hot encoding.
Numerical normalization : Numerical features like Total_Service , Total_LaborCost , and Total_Parts were normalized to ensure uniformity across the dataset.
Churn classification : Customers who did not return for service within 12 months after their last visit were labeled as "churned" (1), while those who continued to return were labeled as "retained" (0).
Following preprocessing, the dataset was transformed into a binary classification format suitable for supervised machine learning algorithms.

Descriptive analysis was conducted to understand the general characteristics and distribution of the data. It revealed that approximately 25% of customers were classified as churned, indicating a significant portion of lost customers within the dataset. Figure 1 illustrates the overall composition of churned and retained customers. Additionally, Figure 2 shows how churn behavior varies across different independent variables such as WarrantyStatus , VehicleModel , and Dissatisfaction . Notably, customers with vehicles still under warranty showed higher retention rates compared to those out of warranty. Similarly, dissatisfaction levels were strongly correlated with churn probability, emphasizing the importance of customer satisfaction in after-sales service.

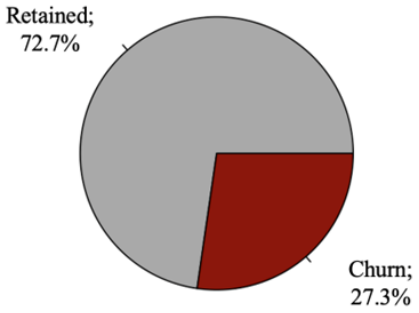

Figure 1
Source: Research data

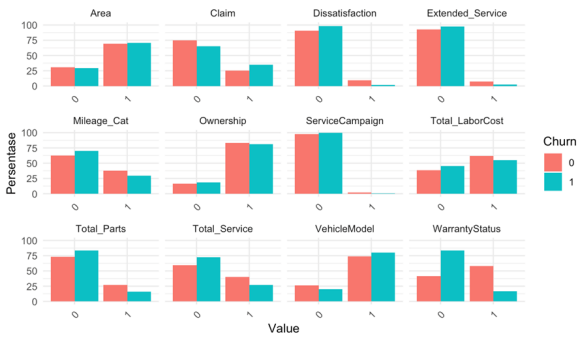

Figure 2
Source: Research data

Two machine learning algorithms were implemented to develop predictive models for customer churn: Decision Tree , Random Forest . Each algorithm was trained using varying

test sizes (20%–60%) to assess the impact of test data proportion on model performance. The Decision Tree algorithm was used to generate interpretable rules by recursively splitting the data into increasingly homogeneous subsets. Table 1 summarizes the accuracy of the Decision Tree model across different test sizes.

Table 1. Accuracy of Decision Tree Algorithm

Test Size	20%	30%	40%	50%	60%
Accuracy	81.94%	82.02%	81.99%	81.99%	81.96%

Source: Research data

A confusion matrix for the Decision Tree model is presented in Table 2 , showing the number of correctly and incorrectly predicted instances

Table 2 Confusion Matrix Algorithm Decision Tree

Confusion Matrix	Actual "0"	Actual "1"
Predict "0"	80,142	14,687
Predict "1"	6,82	17,96

Source: Research data

The Random Forest algorithm built upon Decision Trees by creating an ensemble of trees trained on bootstrapped samples. Table 3 presents the accuracy across different test sizes

Table 3. Accuracy of Random Forest Algorithm

Test Size	20%	30%	40%	50%	60%
Accuracy	81,91%	81,95%	81,94%	81,96%	81,95%

Source: Research data

The confusion matrix for the Random Forest model is shown in Table 4

Table 4 Confusion Matrix Algorithm Random Forest

Confusion Matrix	Actual "0"	Actual "1"
Predict "0"	133,84	24,873
Predict "1"	11,096	29,538

Source: Research data

CONCLUSION

This study demonstrates the effectiveness of machine learning models, particularly Decision Tree and Random Forest, in predicting customer churn within the four-wheeled vehicle after-sales service industry using big data analytics. By analyzing historical customer data from Brand X over a seven-year period, the research identified key churn drivers such as Warranty Status, Total Service Frequency, and Dissatisfaction Level. The predictive models achieved high accuracy and sensitivity, enabling proactive customer retention strategies. These findings provide valuable insights for automotive companies seeking to enhance customer loyalty through data-driven decision-making and personalized service offerings.

REFERENSI

Abdullah-All-Tanvir, Ali Khandokar, I., Muzahidul Islam, A. K. M., Islam, S., & Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. Heliyon, 9(4). https://doi.org/10.1016/j.heliyon.2023.e15163

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research, 55(1), 80–98. https://doi.org/10.1509/jmr.16.0163

Camm, J. D., Cochran, J. J., Fry, M. J., Ohlmann, J. W., Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2019). Business Analytics (3rd Edition). Cengage. www.solver.com/aspe

Cao, P. (2021). Big Data in Customer Acquisition and Retention for eCommerce-Taking

Walmart as an Example.
Capponi, G., Corrocher, N., & Zirulia, L. (2021). Personalized pricing for customer retention: Theory and evidence from mobile communication. Telecommunications Policy, 45(1). https://doi.org/10.1016/j.telpol.2020.102069
Chandra Setiawan, I., Indarto, & Deendarlianto. (2021). Quantitative analysis of automobile sector in Indonesian automotive roadmap for achieving national oil and CO2 emission reduction targets by 2030. Energy Policy, 150. https://doi.org/10.1016/j.enpol.2021.112135
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794. https://doi.org/10.1145/2939672.2939785
Díaz, V. G.-P., & Márquez, A. C. (2014). After–sales Service of Engineering Industrial Assets: A Reference Framework for Warranty Management. Springer International Publishing.
Durugbo, C. M. (2020). After-sales services and aftermarket support: a systematic review, theory and future research directions. In International Journal of Production Research (Vol. 58, Issue 6, pp. 1857–1892). Taylor and Francis Ltd. https://doi.org/10.1080/00207543.2019.1693655
Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. In Journal of Animal Ecology (Vol. 77, Issue 4, pp. 802–813). https://doi.org/10.1111/j.1365-2656.2008.01390.x
Exactitude Consultancy. (2024, December). Automotive After-Sales Service Market. https://exactitudeconsultancy.com/id/reports/41416/automotive-after-sales-service-market
Guerra, E. (2019). Electric vehicles, air pollution, and the motorcycle city: A stated preference survey of consumers' willingness to adopt electric motorcycles in Solo, Indonesia. Transportation Research Part D: Transport and Environment, 68, 52–64. https://doi.org/10.1016/j.trd.2017.07.027
Homburg, C., Koschate, N., & Hoyer, W. (2006). The Role of Cognition and Affect in the Formation of Customer Satisfaction: A Dynamic Perspective. Journal of Marketing, 70.
IEA. (2023). Net Zero Roadmap A Global Pathway to Keep the 1.5 °C Goal in Reach. Www.Iea.Org/T&c/.
IQAir. (2023). Kualitas udara di Jakarta.
Kim, M. K., Oh, J., Park, J. H., & Joo, C. (2018). Perceived value and adoption intention for electric vehicles in Korea: Moderating effects of environmental traits and government supports. Energy, 159, 799–809. https://doi.org/10.1016/j.energy.2018.06.064
Kotler, Philip., Keller, K. Lane., Tan, C. Tiong., Ang, S. Hoon., & Leong, S. Meng. (2018). Marketing management : an Asian perspective. Pearson Education Limited.
Kumar, A. (2021). Analysing the drivers of customer happiness at authorized workshops and improving retention. Journal of Retailing and Consumer Services, 62. https://doi.org/10.1016/j.jretconser.2021.102619
Lamrhari, S., Ghazi, H. El, Oubrich, M., & Faker, A. El. (2022). A social CRM analytic framework for improving customer retention, acquisition, and conversion. Technological Forecasting and Social Change, 174. https://doi.org/10.1016/j.techfore.2021.121275
Leong, C. M., Cheah, J. H., Ting, H., Lim, R., Ariffin, A. B. B., & Lim, X. J. (2024). Enhancing Customer Retention: The Role of Customer Satisfaction and Delight in the Authorized Automotive After-Sales Service Sector. Journal of Applied Structural Equation Modeling, 8(1), 1–26. https://doi.org/10.47263/JASEM.8(1)05
Liu, Z., Jiang, P., De Bock, K. W., Wang, J., Zhang, L., & Niu, X. (2024). Extreme gradient

boosting trees with efficient Bayesian optimization for profit-driven customer churn prediction. Technological Forecasting and Social Change, 198. https://doi.org/10.1016/j.techfore.2023.122945

Malik, R., Sharma, A., & Chaudhary, P. (2025). Augmenting Customer Retention Through Big Data Analytics.

Murtiningrum, A. D., Darmawan, A., & Wong, H. (2022). The adoption of electric motorcycles: A survey of public perception in Indonesia. Journal of Cleaner Production, 379. https://doi.org/10.1016/j.jclepro.2022.134737

Noviandy, T. R., Idroes, G. M., & Hardi, I. (2025). Integrating explainable artificial intelligence and light gradient boosting machine for glioma grading. Informatics and Health, 2(1), 1–8. https://doi.org/10.1016/j.infoh.2024.12.001

Nyadzayo, M. W., & Khajehzadeh, S. (2016). The antecedents of customer loyalty: A moderated mediation model of customer relationship management quality and brand image. Journal of Retailing and Consumer Services, 30, 262–270. https://doi.org/10.1016/j.jretconser.2016.02.002

Omotehinwa, T. O., Oyewola, D. O., & Dada, E. G. (2023). A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis. Healthcare Analytics, 4. https://doi.org/10.1016/j.health.2023.100218

Omotehinwa, T. O., Oyewola, D. O., & Moung, E. G. (2024). Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease. Informatics and Health, 1(2), 70–81. https://doi.org/10.1016/j.infoh.2024.06.001

Panhalkar, A. R., & Doye, D. D. (2022). Optimization of decision trees using modified African buffalo algorithm. Journal of King Saud University - Computer and Information Sciences, 34(8), 4763–4772. https://doi.org/10.1016/j.jksuci.2021.01.011

Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. International Journal of Intelligent Networks, 4, 145–154. https://doi.org/10.1016/j.ijin.2023.05.005

Saha, D., & Manickavasagan, A. (2021). Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. In Current Research in Food Science (Vol. 4, pp. 28–44). Elsevier B.V. https://doi.org/10.1016/j.crfs.2021.01.002

Schweitzer, E., & Aurich, J. C. (2010). Continuous improvement of industrial product-service systems. CIRP Journal of Manufacturing Science and Technology, 3(2), 158–164. https://doi.org/10.1016/j.cirpj.2010.04.002

Sharda, R., Delen, D., Turban, E., Aronson, J. E., Liang, T.-P., & King, David. (2018). Business intelligence, analytics, and data science : a managerial perspective (4th Edition). Pearson Education.

Subroto, A. (2020). Predicting Customer Churn in the Indonesia Telecommunications Industry Using Big Data. NOVA Publisher.

Sun, K. K., He, S. Y., & Thøgersen, J. (2022). The purchase intention of electric vehicles in Hong Kong, a high-density Asian context, and main differences from a Nordic context. Transport Policy, 128, 98–112. https://doi.org/10.1016/j.tranpol.2022.09.009

Thangeda, R., Kumar, N., & Majhi, R. (2024a). A neural network-based predictive decision model for customer retention in the telecommunication sector. Technological Forecasting and Social Change, 202. https://doi.org/10.1016/j.techfore.2024.123250

Thangeda, R., Kumar, N., & Majhi, R. (2024b). A neural network-based predictive decision model for customer retention in the telecommunication sector. Technological Forecasting and Social Change, 202. https://doi.org/10.1016/j.techfore.2024.123250

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning

techniques. Results in Control and Optimization, 14. https://doi.org/10.1016/j.rico.2023.100342

Wang, J., Lai, X., Zhang, S., Wang, W. M., & Chen, J. (2020). Predicting customer absence for automobile 4S shops: A lifecycle perspective. Engineering Applications of Artificial Intelligence, 89. https://doi.org/10.1016/j.engappai.2019.103405

Zhang, W., Gu, X., Hong, L., Han, L., & Wang, L. (2023). Comprehensive review of machine learning in geotechnical reliability analysis: Algorithms, applications and further challenges. In Applied Soft Computing (Vol. 136). Elsevier Ltd. https://doi.org/10.1016/j.asoc.2023.110066

Zhao, R., Hong, L., Ji, H., Zhang, Q., Zhang, S., Li, Q., & Gong, H. (2025). Decision tree based parameter identification and state estimation: Application to Reactor Operation Digital Twin. Nuclear Engineering and Technology. https://doi.org/10.1016/j.net.2025.103527