Estimating Peak Sales of Mechanical Store Item By Using Regression Tree Algorithm (Case Study at CV Kalbar Jaya Motor)

Andre Saputra[1*], Jimmy Tjen[2]

[1] Department of Digital Business, Faculty of Information Technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.
[2]Department of Informatics, Faculty of Information Technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.

Email:
[1] 22430142@widyadharma.ac.id
[2] jimmy.tjen@mathmods.eu

Abstract. This study aims to predict sales patterns for bearing items at CV Kalbar Jaya Motor using the regression tree method. The dataset consists of 4,029 bearing items from a total of 20,830 transactions. The primary objective is to identify the days and dates with the highest sales volumes to optimize stock management and inform marketing strategies. The analysis results demonstrate that the regression tree model predicts sales trends with an accuracy rate of 76.4%. The highest predicted sales occur on specific days and dates, influenced by factors like day, date, historical sales, and customer ordering patterns. For future research, it is recommended to explore ensemble learning method such as Random Forest or Gradient Boosting to enhance the accuracy and robustness of the predictions.

Keywords: Machine Learning; Prediction; Regression Tree; Bearings

Introduction

This study aims to enhance economic growth (SDG 8) by optimizing sales predictions for bearing items and fosters innovation in industry practices (SDG 9) through the application of advanced regression tree methods in data analysis and stock management. In recent years, technological advancements have significantly influenced the field of Machine Learning (ML), a domain that optimizes processes and minimizes human effort. As a branch of artificial intelligence, ML combines principles from statistics and computer science to create algorithms that enhance their performance by learning from data rather than relying solely on predefined instructions. Its applications encompass a wide range of areas, including speech recognition, image processing, localization, and text analysis. Fundamentally, ML involves the automatic enhancement of computational algorithms through experience and is widely recognized as a subset of artificial intelligence (Charbuty & Abdulazeez, 2021). Among its practical applications, sales forecasting or sales prediction utilizes historical data and statistical techniques to predict future outcomes (Pratiwi, Fauzi, Lestari, & Cahyana, 2024). One widely used Machine Learning (ML) techniques for prediction and decision making is the Decision Tree model. This modern approach analyzes domain specific data and constructs predictive models through systematic analysis. However, to address the challenges posed by large datasets, the Decision Tree algorithm often requires adaptation for distributed environments to achieve faster training times without compromising the accuracy of the resulting models (Patil & Kulkarni, 2019). Structurally, the Decision Tree facilities decision making using a tree, based on modeling approach that maps potential outcomes of interrelated choices. In statistical regression, it employs supervised learning techniques to represent decisions, incorporating elements such as input costs, potential outcomes, and utility. Furthermore, by training on monitored data, the Decision Tree model can effectively predict future conditions. Its applications extend to enhancing business processes by identifying critical elements or activities where improvements can have significant impacts (Sishi & Telukdarie, 2021).

There are few research papers that be the references for this paper that utilized this methodology, (Ozcan & Peker, 2023) This study utilized the CART algorithm to predict heart disease with 87% accuracy, rank key

features, and extract decision rules to aid clinical applications, benefiting both healthcare professionals and patients by simplifying diagnosis and reducing costs and time. (Huang & Lin, 2020) This study utilized a regression tree model to analyze 2017-2018 NBA game data, predicting individual player scores, total team scores, and game outcomes with 87.5% accuracy. (Yan, Zhou, Xiao, & Pan, 2023) This paper analyzes the factors affecting house prices, using a CART decision tree, based on molybdenum metal price algorithm model, and shows that with an average absolute error of 4.03% and a forecasting trend accuracy of 94.8%, the algorithm proves to be reliable. (Vinutha & Chandrika, 2021) This study employs the regression tree method to predict liver cirrhosis using a dataset of 435 records, demonstrating higher accuracy for female patients (MAE = 0.17, MSE = 0.03, RMSE = 0.17) compared to males, thereby proving its effectiveness in supporting the early diagnosis of liver disease.

This research employs the regression tree method to predict the highest sales levels of bearing items, aiming to identify the months in which peak sales occur based on historical data analysis. The dataset utilized in this study consists of sales data from CV Kalbar Jaya Motor, encompassing detailed information on dates, sales quantities, and other relevant variables necessary for the analysis.

The paper is organized as follows: in Method section, we describe how the dataset used in this paper was generated. In the Results and Discussion section, we detail the dataset that was tested using the Decision Tree algorithm and provide examples of the results obtained. Finally, in the Conclusion section, we summarize the progress made in this research, present the findings, and discuss how the data can be utilized in future studies.

Methods

This section describes a systematic approach to predicting when the highest sales levels occur in using regression tree methodology. The details of this method include Data Collection, Preprocessing Data, Data Process, Perception and Implementation. The following outlines the structure of this research work:

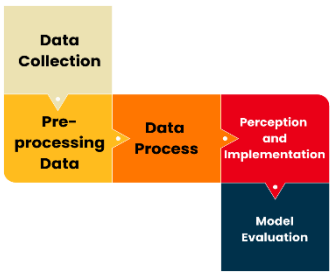

Figure 1. Workflow or Research Framework

1. **Data Collection**

   This study employs an experimental qualitative approach to analyze the sales performance of bearings, aiming to identify the days with the highest sales levels for these products. Primary data was collected and analyze from the historical sales records of CV Kalbar Jaya Motor, covering the period from January 2024 to June 2024. The study population comprised 20,830 sales transactions, from which a subset of 4,029 transactions involving bearings was selected as the sample data. CV Kalbar Jaya Motor, a supplier of Fuso and Truck spare parts based in Pontianak, West Kalimantan, Indonesia, serves as the focus of the research. In this study, the day with the highest sales of bearing is treated as the dependent variable, while the daily sales volume of bearings serves as the independent variable.

2. **Preprocessing Data**

   This research begins by collecting sales data from CV Kalbar Jaya Motor and organizing it into a structured table. The table comprises several columns: the first column contains the transaction date

(DD-MM-YYYY), the second column lists the item names, and the third column records the quantity of items sold (qty). To determine the interval between the training data and the prediction data, the lag matrix method is applied. The compiled data is then processed and transformed into a decision tree model, with 80% of the dataset allocated as training data and the remaining 20% as test data.

3. **Data Process**

   In the initial stage, the research begins by indexing all available items. Subsequently, a search is conducted for items based on the previously generated indexing results. Once an item is identified, its position is determined and numbered according to the search results. Date related information is then extracted from the item corresponding to its positional numbering. Following this, data related to the identified items is retrieved based on the predefined numbering. The process continues with the elimination of data associated with installment schemes to ensure the appropriateness of the analysis. After removing the installment related data, a data augmentation process is performed to facilitate the retrieval of sales data from previous days. The subsequent step involves dividing the data into training and test datasets, which are then utilized to generate prediction graphs and test graphs for further analysis. Finally, additional graphs are created to evaluate the significance of parameters within the model.

4. **Perception and Implementation**

   In contemporary business practice, understanding market dynamics is essential for strategic decision making and achieving a competitive advantage. A recent analysis using regression tree methodology revealed specific days namely Tuesday, Wednesday, Thursday, and Saturday as peak sales periods, with Wednesday exhibiting the highest sales volume. This insight derives from an evaluation of several month worth of sales data, demonstrating a significant correlation between particular weekdays and an increase in consumer purchasing behavior. The implementation of targeted marketing strategies based on these findings, such as customized promotions and inventory management, has the potential to enhance customer engagement and optimize stock levels. Furthermore, ongoing evaluation of sales data post implementation will enable businesses to continuously refine their strategies, ensuring alignment with evolving consumer preferences and maximizing sales performance.

5. **Evaluation Model**

   In this research, the accuracy formula is utilized to measure the performance of the trained model. This formula provides insight into how well the model can predict results compared to actual values, with higher accuracy values indicating better model performance. The equation is below (1).

$$Accuracy = 1 - \left(\frac{\sum(S_i - o_i^2)^2}{\sum o_i^2}\right) \text{ x } 100\% \quad (1)$$

Result and Discussion

In this study, sales data were analyzed to predict future periods with the highest sales levels, including identifying specific days and dates with the greatest sales. The prediction model achieved an accuracy of 76.4%. This level of accuracy suggests that the model performs effectively in predicting sales patterns based on the available input variables. At the initial stage, a Decision Tree Model was developed, as illustrated in Figure 2. In additionally, Figure 3. shows that the most importance parameter is the date variable.

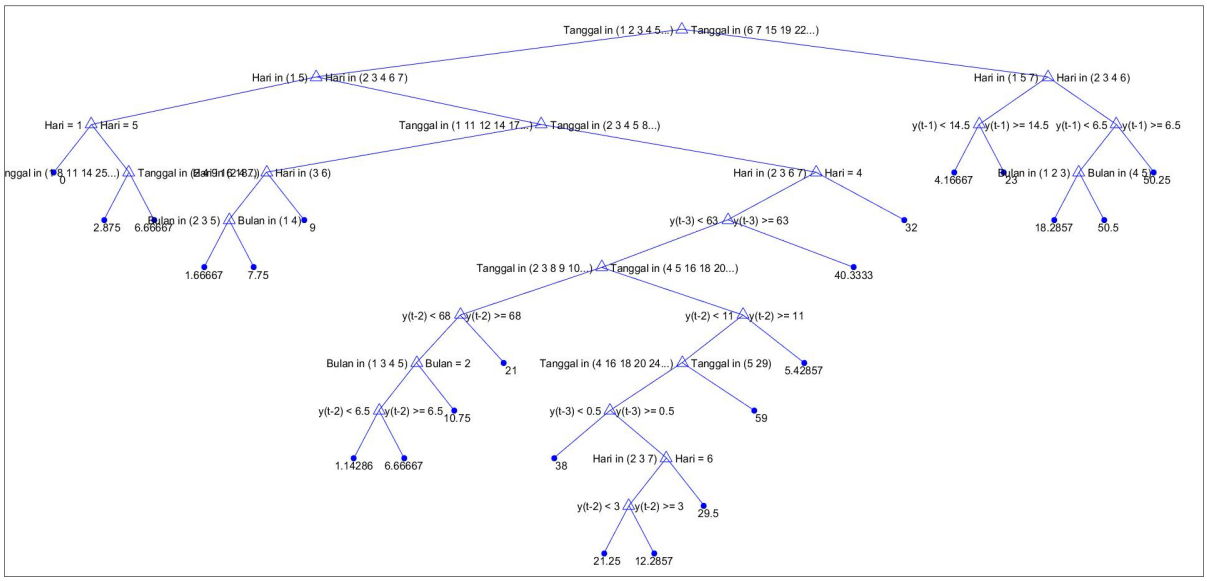

Figure 2. Decision Tree Calculation Results

The nodes with a sales prediction value of 50.5 include certain days, namely Tuesday, Wednesday, Thursday, and Saturday, as well as the 6th, 7th, 15th, 19th, 22nd, 27th, and 30th of April and May. Predictions on these dates and months show conditions that match reality. This node represents the highest sales value because the conditions leading to this node reflect the optimal combination of parameters that affect demand. These parameters include the specific day of the week, the value of the previous sales break, and the month. For example, Wednesday appears to be the day with the highest sales rate, due to its position in the middle of the work week, when business activity is usually at its peak. This is particularly relevant for customers in the industrial sector, who often place regular orders in the middle of the week. Therefore, CV Kalbar Jaya Motor can effectively manage bearing stock to avoid shortages on these specific days or dates. In addition, the date parameter shows a significant influence in this study, as it is directly related to a specific order cycle or schedule that affects the sales pattern.

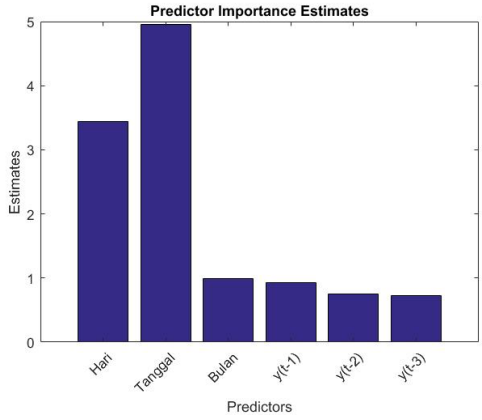

Figure 3. Importance Parameter

The importance parameter in Figure 3. highlights that the date parameter makes the most significant contribution in reducing sales variability by effectively partitioning the data. Factors such as seasonal customer behavior, structured ordering schedules, and periodic promotions or maintenance on specific dates may underlie the importance of this parameter in analysis. In summary, the combination of parameters: date, day of the week, and historical sales value resulted in a significant impact culminating in the node with the highest sales prediction.

Conclusion

This study successfully analyzed sales data to predict periods with the highest sales levels, identifying specific days and dates associated with optimal sales performance. The Decision Tree model developed in the research achieved an accuracy rate of 76.4%, demonstrating its effectiveness in predicting sales patterns based on the input variables provided.

The findings indicate that the node with the highest predicted sales value (50.5) is characterized by an optimal combination of parameters, including specific days (e.g., Wednesday, which falls in the middle of the workweek), lag values of historical sales, and specific months (April and May). Additionally, date parameters significantly contribute to the analysis due to their direct correlation with order cycles or schedules that influence sales patterns. Key factors such as seasonal customer behavior, structured ordering schedules, and periodic promotions or maintenance underpin the significance of these parameters in reducing sales variability. Collectively, the combination of date, day of the week, and historical sales value parameters was found to have a substantial impact, enabling the identification of periods with the highest predicted sales levels.

References

Anshori, M. Y., Herlambang, T., Asy'ari, V., Katias, P., Firdaus, A. A., & Arof, H. (2024). Forecasting Occupancy Rate using Neural Network and Decision Tree at Hotel X. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 49--62.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of applied science and technology trends*, 20--28.

Huang, M.-L., & Lin, Y.-J. (2020). Regression tree model for predicting game scores for the golden state warriors in the national basketball association. *Symmetry*, 835.

Jijo, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 20-28.

Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, 2772-4425.

Patil, S., & Kulkarni, U. (2019). Accuracy Prediction for Distributed Decision Tree using Machine Learning approach. *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 1365-1371.

Pratiwi, S. A., Fauzi, A., Lestari, S. A., & Cahyana, Y. (2024). Prediksi Persediaan Obat Pada Apotek Menggunakan Algoritma Decision Tree. *KLIK: Kajian Ilmiah Informatika dan Komputer*, 2381--2388.

Sishi, M., & Telukdarie, A. (2021). The application of decision tree regression to optimize business processes. *Proceedings of the International Conference on Industrial Engineering and Operations Management Sao Paulo, Brazil*, 48--57.

Vinutha, M., & Chandrika, J. (2021). Prediction of liver disease using regression tree. *International Association of Online Engineering*.

Yan, J., Zhou, Q., Xiao, Y., & Pan, B. (2023). A Comparative Study Of CART Algorithm For Forecasting. *2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA)*, 124-131.