Implementation of Regression Tree Algorithm for Estimating Bread Sales (Case Study: Bysea Bites Pontianak)

Stevin Tandra[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: 22430186@widyadharma.ac.id[1], jimmy.tjen@mathmods.eu[2]

Abstract. This research aims to predict the sales pattern of Bysea Bites Bakery products in Pontianak, West Kalimantan, Indonesia, using sales data. By taking 1,150 samples, consisting of date (DD-MM-YYYY), product name, and quantity sold. Using training (80%) and testing (20%) subset data, the Regression Tree algorithm was applied to predict when wheat bread would experience high sales and to find the influential factors. The performance of the model was assessed using Root Mean Square Error(RMSE) and Normalized Root Mean Square Error(NRMSE). The accuracy of the model is 73.78%. The prediction shows that on the 3, 13, 19, and 21 are dates with high demand for wheat bread and high sales on Monday, Tuesday, Thursday, Friday, and Saturday. These insights allowed the bakery to optimize its inventory, and improve customer satisfaction. Future research can examine the influence of external factors on sales patterns.

Keywords: Machine Learning; Prediction; Regression Tree; Wheat Bread.

Introduction

This research aims to improve sustainable development goals by helping to optimize inventory to be more efficient(SDG 8). Effective marketing strategies are needed to help companies make decisions in today's business era. To produce a good marketing strategy, can be done by analyzing sales data. With us analyzing the data, we get information where this information can be used to make decisions that will be taken so that it can increase sales (Abalkanov M. M., 2023). One of the effective approach is the Regression Tree algorithm, this algorithm is relevant for predicting sales at bakeries (Catal, Kaan, Arslan, & Akbulut, 2019). By analyzing historical sales data, it is possible to develop models to predict which products will sell well (Platikanova M., 2022). With the data that has been processed can be used to support decision making that will be carried out by the company (Sishi, 2021).

Regression Tree is an algorithm in machine learning that can be used to create prediction models through existing data (Nie, 2021) (Liu, Wang, Huang, & Yin, 2020). This kind of decision tree was created especially to model the link between continuous input and output variables in regression issues (Eshankulov H., 2022). This algorithm has the advantages of being easy to understand, does not require data normalization, and good performance for non-linear data (Ao, Li, Zhu, Ali, & Yang, 2019). Few research publications that used this algorithm are used as references in this work, see e.g. (Fukui, 2023). This retrospective cross-sectional study examined COVID-19 patients who were diagnosed between January 1 and December 31,2020, at a university hospital. The study looked at clinical information such as symptoms, vital signs, lab result, and CT scan results to uncover predictors of COVID-19 pneumonia. Of the 221 patients, 160(72,4%) had pneumonia. (McNamara-Pittler, et al., 2024). 3,383 people with shoulder pain-including those with glenohumeral osteoarthritis(GHOA) and controls with other shoulder pathologies-participated in this cross-sectional study. 33 possible risk factors were evaluated in the study, and the most important ones were determined using

CART analysis. (Ghodsi, et al., 2022). This study investigated the association between polymorphism of five vitamin D receptor (VDR) genes (Apal, Bsml, Fokl, EcoRV, and Taql) and low bone density/osteopenia/osteoporosis in individuals with type 2 diabetes (T2D) using classification and regression tree (CART) algorithms. Data from 158 T2D participants were analyzed, considering factors such as age, BMI, Vitamin D deficiency, gender, and VDR gene polymorphisms. Age was the primary predictor of low bone density in all groups, followed by BMI, EcoRV polymorphism in women, and Taql polymorphism in men. The CART model demonstrated accuracy rates of 75.32% for both sexes. (Arif Rahman Hakim, 2023). This study aimed to classify male fertility levels using machine learning algorithms, specifically the Classification and Regression Tree(CART) algorithm, combined with the K-Fold Cross Validation method. The dataset, sourced from the UCI Machine Learning website, consisted of 100 data points with variables such as age, childhood diseases, accidents, surgical interventions, alcohol consumption, smoking habits, and more. The K-Fold Cross Validation method (K=1 to K=9) was used to assess the performance of the CART model, achieving an average accuracy of 98.70% for training data and 81.16% for testing data.

By using a regression tree, it is possible to predict when a product sells the most (Yang, Wang, Xu, Huang, & Tsui, 2020). So that by getting this information we can make decisions about how we will market the product, what strategies are suitable for increasing sales, or can restock products when the product is often sold. (Raizada S., 2021). This research is used to find out when sales will be predictably high, by getting information on when sales will be high, strategies can be determined to further increase sales (Christa, Suma, & Mohan, 2022).

The focus of this research is to use data to create a bakery's top sales prediction model. A Regression Tree algorithm will be used to assess historical sales data, including external variables such as holidays and promotions (Phyo P.P., 2021). This method will give shop owners insight into the elements that affect sales growth in addition to assisting them in establishing production plans.

The paper is structured as follows: in section Method, we describe how the setup dataset used in this paper was generated; in section Result and Discussion, we describe the dataset that was tested using the Regression Tree algorithm and provide examples of the result that were tested; and in section Conclusion, we describe the progress made in this research, present the findings, and provide estimates of how the data can be used in future studies.

Methods

This section discusses the prediction of best-selling sales using regression trees from the Bysea Bites store. In this solution, the steps taken in the methodology are data collection, data preprocessing, model building, and interpretation. The structure of this research project is described below.

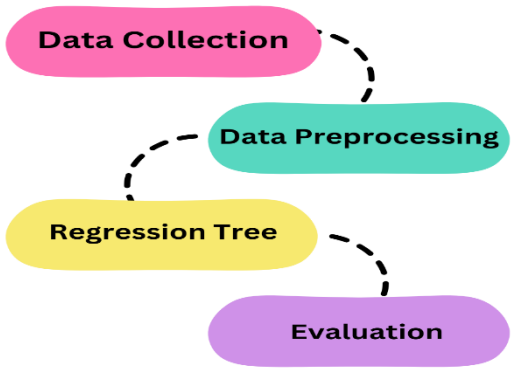

Figure 1. Process of The Research Framework

1. **Data Collection**

   This study uses bakery sales data used to predict when the most sales of a product are sold, then making predictions based on the average target value in each subset. The object used in this data is the Bysea Bites bakery, located in Pontianak, West Kalimantan, Indonesia, with the number of samples used 1,150. This study took data from the bakery Bysea Bites by taking the date(DD-MM-YYYY), product name, and thirdly indicating the quantity.

2. **Data Preprocessing**
   There is data preprocessing, which is data from the sum of several wheat bread transactions on the same date added together. The reason i choose wheat bread is because wheat bread is Bysea Bites best seller and is the most sold bread product.
3. **Regression Tree**
   With the preprocessed data, the Regression Tree algorithm is used. This method of machine learning creates a model that resembles a tree and uses input data to forecast sales, one of which is to predict when the highest sales of a product will be. The process is carried out by selecting the desired product, then retrieving the existing date information, after that removing the outlier data, then doing data augmentation, followed by creating a tree structure that is used to forecast sales. The dataset is divided into two subsets: training data(80%) and test data(20%).
4. **Evaluation**
   Root Mean Square Error(RMSE) and Normalized RMSE(NRMSE) are performance measures. By comparing the anticipated sales figures with the actual sales from the test dataset, the accuracy of the model can be assessed, with accuracy determined as follows: accuracy = (1 – NRSME)*100%.The variables that have the greatest impact on sales projections are determined by interpreting the generated model. Important determinants are emphasized, including product type, seasonality, and marketing. The model guides future product stocking selections and assists in determining which bakery products are most likely to have large sales volumes. A prediction is then created by adding the sales data and its average value, which includes the average anticipated quantity of bread sales. After then, a huge number of leaf and branch nodes will be made in order to determine which day of the week or month-represented by a number-is best for replenishing. After obtaining the results, it can be used to help the company to make the right decisions on when to make the right stocking, and when to provide attractive promotions.

Result and Discussion

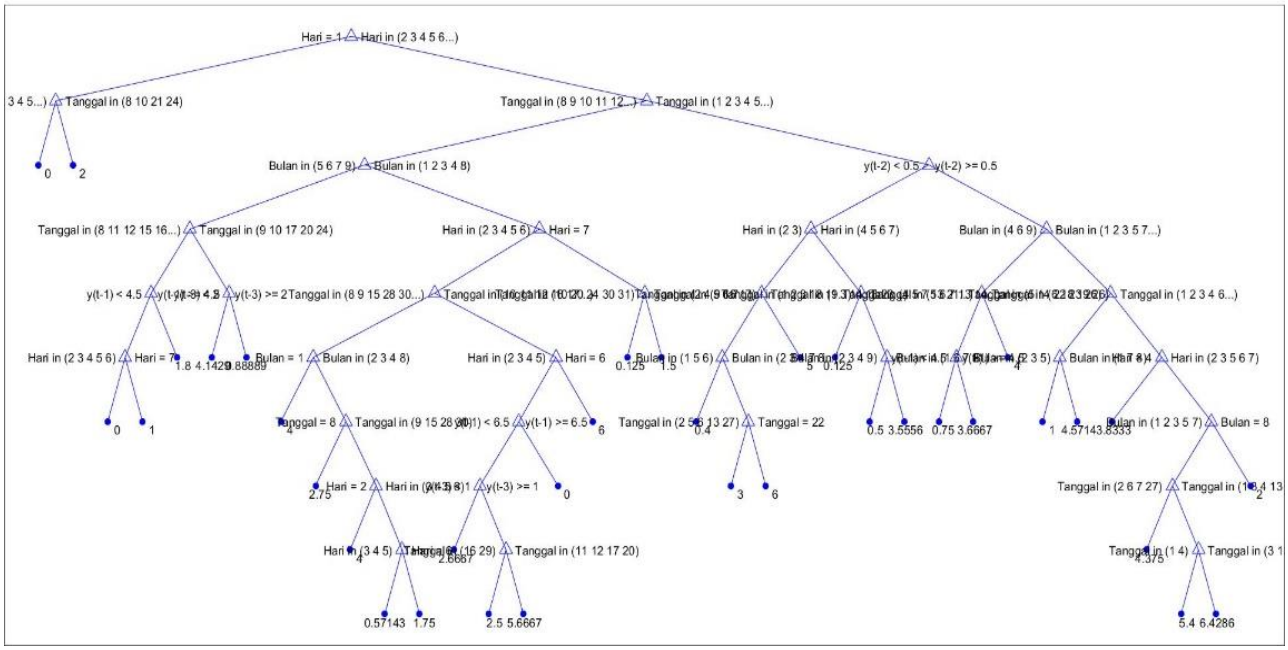

Figure 2. Decision Tree Calculation Results

In this study, sales information can be used to forecast future times when it is appropriate to perform daily, weekly, or monthly restocking, with this prediction model's accuracy value is 73.78%. The Decision Tree is

generated as in Figure 2. For example, the model predicts that there will be four dates when a lot of wheat bread is sold, namely the 3, 13, 19, and 21, indicating that wheat bread is high with a value of 6,4 units. This is consistent with the weekly restocking schedule. By getting information about when the dates will be crowded, strategies can be made to further increase sales; namely preparing more stock, providing attractive discounts on dates that are predicted to be busy, and organizing production time so that on busy dates there is enough time to produce so as not to reduce the quality of the product if it is produced in a hurry.

The prediction also found that the highest sales occurred on Monday, Tuesday, Thursday, Friday, and Saturday, which is influenced by weekdays so for people who want to have instant breakfast which makes the data generated wheat bread sales will be high on weekdays. So from the analysis, it is found that the time of sale, especially the day and date, can affect the high sales that occur. This prediction model also provides information about sales patterns at Bysea Bites, namely in the early and late weeks of the month showing higher sales patterns.

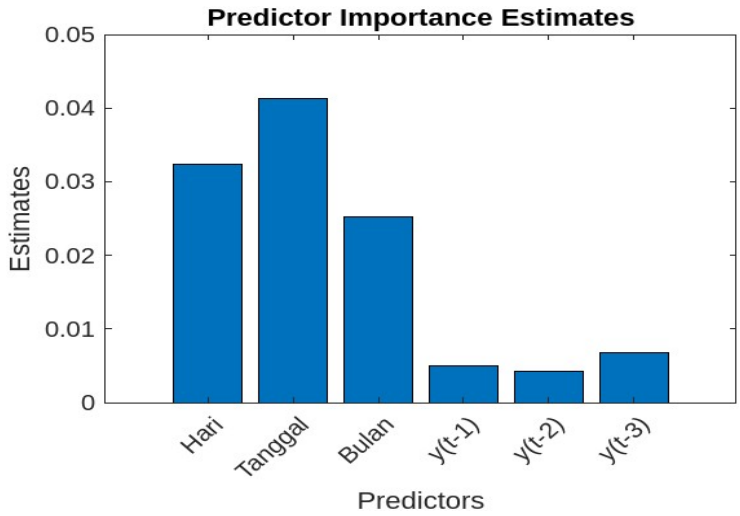

Figure 3. Parameter Importance Results

The Parameter Importance is generated as in Figure 3, based on the parameters generated it is found that it turns out that the sales of the past few days have slight no to effect while those that do are days and dates. This can be caused by the shopping habits of certain customers who shop at the beginning of the week (Monday) or to prepare for the end of the week (Saturday), dates can be caused by the receipt of salaries such as the 3, 13, 19, and 21 are dates where it is still the middle of the month.

From this analysis, we can stock more wheat bread in that period. The resulting schedule can be a valuable resource for companies in strategizing the right time to restock. In addition, this schedule can also be a reference for increasing product sales and maintaining customer satisfaction.

Conclusion

In this study, the Regression Tree algorithm is used as an analytical tool to determine customer purchasing patterns of wheat bread products from Bysea Bites based on sales data. This insight is then used to find out when wheat bread products sell a lot in the daily and weekly timeframes, with this prediction model's accuracy value is 73.78%. By getting this information, it can be used to make the right time selection when to re-supply stock or reduce stock. The resulting prediction is that the highest sales of wheat bread occur on the 3, 13, 19, and 21, and this prediction finds that Monday, Tuesday, Thursday, Friday, and Saturday are days when products are sold a lot. This prediction model also provides information in the early and late weeks of the month showing higher sales patterns at Bysea Bites. It was also found that days, dates, and promotions can affect the sales levels. Based on this analysis, the regression tree is able to provide a specific and appropriate benefits for Bysea Bites, where using the predicted schedule can be used in the future to optimize inventory management, marketing, and sales strategies that can help reduce waste and also meet customer demand consistently in order to increase customer satisfaction.

References

Abalkanov M. M., A. G. (2023). The role of artificial intelligence and machine learning in business intelligence. *Bulletin of Shakarim University. Technical Sciences* .

Ao, Y., Li, H., Zhu, L., Ali, S., & Yang, Z. (2019). The linear random forest algorithm and its advatanges in machine learning assisted logging regression. *Journal of Petroleum Science and Engineering*.

Arif Rahman Hakim, D. M. (2023). Performance Analysis of Classification and Regression Tree (CART) Algorithm in Classifying Male Fertility Levels with Mobile-Based. *Tech-E*.

Catal, C., Kaan, E., Arslan, B., & Akbulut, A. (2019). Benchmarking of regression algorithms and time series analysis techniques for sales forecasting. *Balkan Journal of Electrical and Computer Engineering*.

Christa, S., Suma, V., & Mohan, U. (2022). Regression and decision tree approaches in predicting the effort in resolving incidents. *International Journal of Business Information Systems*.

Eshankulov H., M. A. (2022). Regression based on decision tree algorithm. *Universum:Technical sciences*.

Fukui, S. a. (2023). A Predictive Rule for COVID-19 Pneumonia Among COVID-19 Patients: A Classification and Regression Tree (CART) Analysis Model. . *Cureus*.

Ghodsi, M., Larijani, B., Roshani, S., Mohammadi Amoli, M., Razi, F., Keshtkar, A. A., . . . Mohajerani Tehrani, M. (2022). An application of CART algorithms for detection of an association between VDR polymorphisms and reduced bone density in individuals with type 2 diabetes: a population-based cross-sectional study. *Journal of Biostatistics and Epidemiology*.

Liu, Q., Wang, X., Huang, X., & Yin, X. (2020). Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data. *Tunnelling and Underground Space Technology*.

McNamara-Pittler, E. N., Prakash, R., Atem, F. D., Pathak, R., Liu, W., Khazzam, M., & Jain, N. B. (2024). Risk Factor Prediction and Categorization for Glenohumeral Osteoarthritis: A Classification and Regression Tree (CART) Analysis. . *American Journal of Physical Medicine \& Rehabilitation*.

Nie, P. a. (2021). Prediction of home energy consumption based on gradient boosting regression tree. *Energy Reports*.

Phyo P.P., J. C. (2021). Daily Load Forecasting Based on a Combination of Classification and Regression Tree and Deep Belief Network. *IEEE Access*.

Platikanova M., Y. A. (2022). Dependence of body mass index on some dietary habits: An application of classification and regression tree. *Iranian Journal of Public Health*.

Raizada S., S. J. (2021). Comparative Analysis of Supervised Machine Learning Techniques for sales forecasting. *International Journal of Advanced Computer Science and Applications* .

Sishi, M. 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*.

Yang, F., Wang, D., Xu, F., Huang, Z., & Tsui, K.-L. (2020). Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model. *Journal of Power Sources*.