ISSN 2085-4579

Enhancing Sales Strategies In Prime Market Retail Business Using Tuned Gradient Boosting

Dudi Nurdiyansah[1], Raymond Sunardi Oetama[1], Iwan Prasetiawan[1]
[1] Department of Information System. Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia
raymond@umn.ac.id

Accepted on May 09th, 2024
Approved on May 29th, 2024

***Abstract*— In the retail sector, comprehending customer behavior and employing effective customer segmentation is pivotal for refining marketing strategies and augmenting profits. This study aims to use predictive modeling for customer segmentation at Prime Market, a prominent retail entity. We apply Gradient Boosting for customer classification. The research initially yields a classification error rate of 25.10%. However, this rate dramatically improves through meticulous parameter tuning, achieving an impressive accuracy of 91.4%. This refined model furnishes invaluable insights into Prime Market's customer segments, enabling the customization of marketing tactics and strategic business approaches. Armed with these insights, Prime Market can make data-driven decisions to enhance customer segmentation accuracy, better comprehend customer preferences, and pinpoint potential avenues for revenue growth. Leveraging advanced data analytics and predictive modeling empowers Prime Market to maintain a competitive edge and deliver its clientele a personalized, gratifying shopping experience.**

Index Terms*— *customer segmentation; data analytics; data mining; Gradient boosting; hyperparameter tunning

I. INTRODUCTION

In the fiercely competitive retail business industry, customer understanding and behavior analysis are fundamental for gaining a competitive edge and achieving success [1]. By delving deep into customer preferences, shopping habits, and needs, businesses can make well-informed, data-driven decisions that significantly impact their performance [2]. A primary advantage of comprehending customers is the ability to customize products and services to cater to their specific demands [3]. By analyzing customer data and identifying trends, companies can create products that resonate with their target audience, fostering customer satisfaction, loyalty, and repeat business. Additionally, customer behavior analysis empowers companies to develop effective marketing strategies [4]. Understanding the factors that drive customer purchases and their preferred shopping channels enables businesses to craft targeted and personalized marketing campaigns. This level of personalization ensures that the right message reaches the right customers at the right time, maximizing the impact of marketing efforts and driving sales.

Moreover, customer behavior analysis offers valuable insights for enhancing customer service [5]. Understanding customer pain points and preferences allows companies to proactively address issues and provide personalized and satisfactory experiences [6]. Happy and contented customers are likelier to become brand advocates, spreading positive word-of-mouth and attracting new customers.

Data mining is the systematic process of extracting valuable knowledge and insights from an extensive database or dataset [7], [8]. Data mining is integral to data analytics, contributing to understanding and utilizing data effectively for decision-making and problem-solving [9]. Data analytics is often employed in retail businesses to analyze data and optimize sales strategies [10], with one popular tool being SAS (Statistical Analysis System). SAS is widely used by statisticians to analyze data more effectively [11], calculating probabilities and statistics for customer numbers [12]. Additionally, SAS enables the generation of relevant visualizations from the analysis results [13].

This study contributes a unique solution to enhance sales strategies in retail business by applying Gradient Boosting with parameter tuning. While several previous studies in this area applied machine learning-based daily retail demand forecasting to examine past data and the impact of special days, such as weekends and holidays [14], forecasting in multi-channel retail using random forests and long short-term memory networks [15], marketing behavior evaluation in multi-channel retail using random forests and term memory networks, with machine learning applied to customer meta-combination brand equity analysis [16].

Prime Market is a supermarket retailer offering a diverse range of products and services across several categories, including Health and Beauty, Electronic Accessories, Home and Lifestyle, Fashion

Accessories, Food and Beverages, and Sports and Travel. Their goal is to provide a pleasant shopping experience and offer quality products at affordable prices to meet the daily needs of consumers in the Bekasi, Depok, and Tangerang areas. This research uses SAS Visual Analytics technology to understand Prime Market's customers better and optimize sales strategies.

The findings from this data analysis and customer segmentation will offer valuable insights to the management team. They will better understand customer preferences and needs, identifying opportunities to boost sales. Moreover, this information can be utilized for product development tailored to customer preferences, more efficient inventory management, and targeted marketing strategies—this research aims to enhance sales performance significantly. The organization can optimize sales strategies, provide an improved shopping experience, and achieve higher customer satisfaction by harnessing the power of data analysis and customer segmentation using SAS Visual Analytics. With a deeper understanding of customer preferences, shopping habits, and needs, the company can focus on developing suitable products, implementing more targeted marketing strategies, and enhancing customer service.

II. METHOD

A. Data

Data is collected from an open-source website, Kaggle, specifically the Prime Market dataset containing 17 columns and 1000 rows. Subsequently, the data is organized using SAS Studio and SAS Data Studio. Some techniques, such as data cleaning, standardization, and parsing, are employed at the data preparation stage.

B. Model Building

At the data modeling stage, Gradient Boosting is applied in SAS Visual Analytics to gain valuable insights and evaluate the performance and accuracy of each model. Afterward, Visualization is used to identify factors influencing sales.

Gradient Boosting is a robust machine-learning algorithm widely used for various classification and regression tasks [17]. It is based on ensemble learning, where multiple weaker learners, typically decision trees, are combined to create a robust predictive model. As shown in Figure 1, the algorithm works iteratively, and in each iteration, it adds a new decision tree to the ensemble, gradually refining its predictions. The main idea behind Gradient Boosting is to leverage the strengths of individual decision trees while compensating for their weaknesses. Each new decision tree is trained to correct the errors made by the previous ones, thus reducing the overall prediction errors. This sequential learning process sets Gradient Boosting apart from other ensemble methods.

C. Model Tuning

Hyperparameters play a crucial role in optimizing the performance of the gradient-boosting model [18]. These hyperparameters control various aspects of the training process, such as the number of trees in the ensemble, the learning rate that determines the contribution of each tree to the final prediction, and the depth of the individual decision trees. One of the critical advantages of Gradient Boosting is its ability to handle large and complex datasets effectively [19]. It can capture intricate relationships and non-linearities in the data, making it well-suited for tasks with high accuracy. However, due to its iterative nature, Gradient Boosting may require more computational resources [20] and longer training times than other algorithms [21]. Decision Trees, as the base learners in Gradient Boosting, are chosen for their simplicity and interpretability. They can handle both numerical and categorical data, making them versatile for a wide range of applications. By combining multiple decision trees in a weighted manner, Gradient Boosting can deliver highly accurate predictions even in noisy and complex data. In summary, Gradient Boosting is a powerful and flexible machine-learning algorithm that leverages the strengths of decision trees to achieve optimal classification results. Its ability to handle large datasets and model complex relationships makes it a popular choice for real-world applications, from customer churn prediction to image recognition and natural language processing.

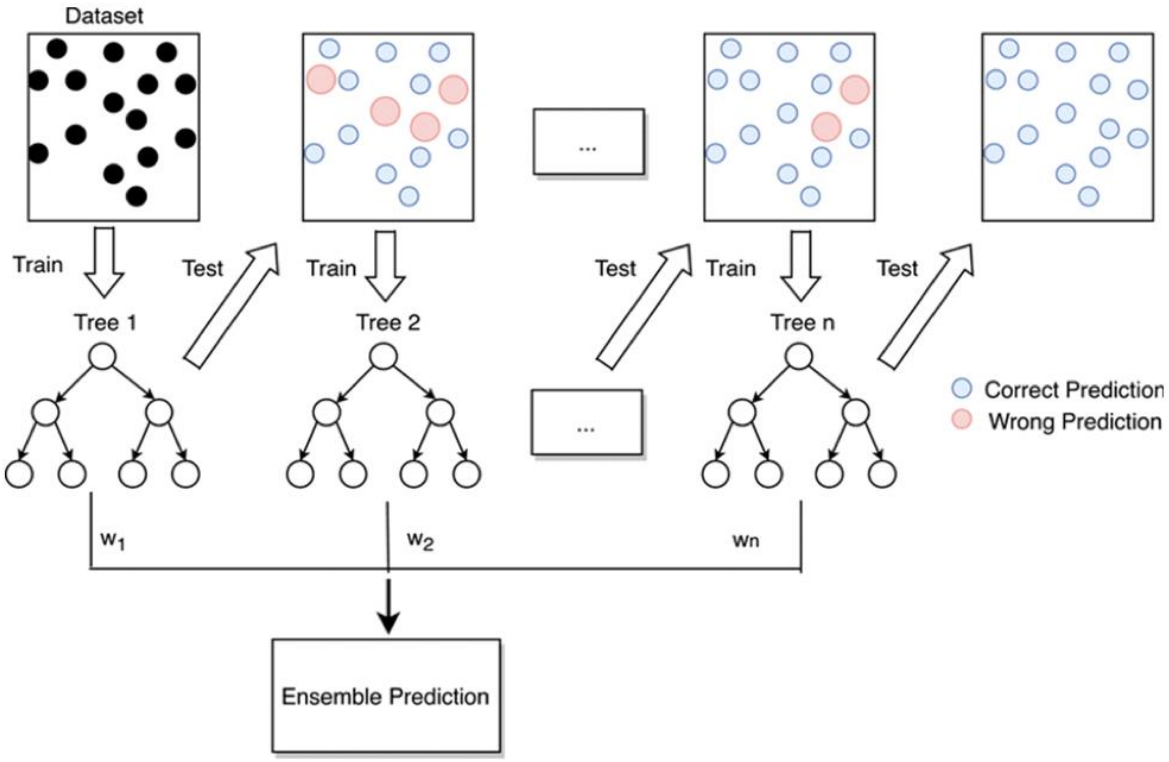

Fig. 1. Gradiant Boosting [22]

D. Model Evaluation

A confusion matrix is a crucial tool for assessing classification model performance, like the Gradient Boosting model in your example, by comparing its predictions to actual dataset labels. This matrix is typically a table with rows and columns representing actual and predicted classes. True Positives (TP) are correct optimistic predictions (e.g., identifying customer segments accurately), True Negatives (TN) are correct pessimistic predictions, False Positives (FP) are incorrect optimistic predictions (false alarms), and False Negatives (FN) are incorrect pessimistic predictions. These metrics are used to derive performance measures such as accuracy (correct predictions overall), Precision (true positives among positive predictions, indicating false positive avoidance), Recall (true positives among actual positives, capturing all positives), and F1 Score (harmonic mean of Precision and Recall, balancing precision-recall trade-offs).

III. Result and Discussion

Before the modeling process, the data was split into training and testing sets using a 70:30 partitioning ratio. This division was achieved through Simple Random Sampling, where the data was randomly partitioned into two predetermined subsets. This study compared five different models to analyze the Prime Market dataset. The model used is Gradient Boosting. This model is applied to predict the Response Customer Type based on predictors such as City, Gender, Payment, Quantity, Tax, Unit Price, COGS, and Rating. The modeling process was carried out to evaluate the performance of each model on the Prime Market dataset. The dataset was divided into training and testing sets in preparation for the modeling phase. This partitioning was carried out using a 70:30 ratio, where 70% of the data was allocated for training purposes, while the remaining 30% was set aside for testing the models.

Based on the Gradient Boosting algorithm, the variable importance analysis revealed that the "unit price" feature significantly contributes to decision-making. Following closely in the second position is the "rating" feature, while "tax," "payment," "quantity," and "city" features rank next in descending order of importance. These insights provide valuable information about the factors influencing customer churn decisions within the Prime Market dataset, enabling businesses to focus on crucial areas when implementing targeted retention strategies. The confusion matrix further illustrates the model's performance on the dataset. It shows 379 True Positives, 122 False Positives, 129 False Negatives, and 370 True Negatives. So, the accuracy of this specific dataset is 74.9%.

TABLE I. Experiment Setting

Options	Minimum	Maximum	Optimal
Number of Trees	50	100	100
Learning rate	0.1	0.9	0.75
Subsample Rate	0.5	0.9	0.5
Lasso	0	0	0
Ridge	1	1	1

Table I offers a concise overview of the options and corresponding parameter values utilized during the tuning process of the Gradient Boosting algorithm. Tuning is crucial in optimizing the model's performance and achieving the best possible results for predicting customer churn. The "Number of Trees" refers to the number of individual decision trees forming the ensemble in the Gradient Boosting algorithm. During tuning, the values for the number of trees varied between 50 and 100. Eventually, it was observed that the best performance was attained with 100 trees, suggesting that a larger ensemble contributed to superior predictive accuracy. The "Learning Rate" is a crucial parameter determining the contribution of each tree to the final ensemble prediction. Different learning rates were examined during tuning, ranging from 0.1 to 0.9. The optimal learning rate was 0.75, as it struck a balance between accuracy and computational efficiency. The "Subsample Rate" represents the fraction of the training data randomly sampled to train each tree in the ensemble. Different subsample rates were tested between 0.5 and 0.9 during the tuning process. The best performance was achieved with a subsampling rate of 0.5, implying that using a smaller portion of the data for each tree led to improved model generalization and reduced overfitting. The "Lasso" and "Ridge" parameters are associated with L1 and L2 regularization, respectively. During model training, these regularization techniques introduce penalty terms to the loss function to control model complexity and prevent overfitting. Lasso and Ridge were employed in this tuning process, with Lasso having a value of 0 and Ridge having a value of 1. The parameter tuning process identified the most effective configuration for the Gradient Boosting algorithm. The optimal combination of hyperparameter values, including the number of trees, learning rate, subsample rate, and regularization, yielded an outstanding accuracy of 91.4% and a significantly reduced misclassification rate of 0.0860. These parameter choices indicate the model's ability to effectively capture intricate patterns and relationships in the Prime Market dataset, making it a powerful tool for customer churn prediction and sales optimization.

TABLE II. CONFUSION MATRIX BEFORE AND AFTER TUNNING

Prediction	Before Tuning		After Tuning	
	Positive	Negative	Positive	Negative
Positive	379	122	459	42
Negative	129	370	44	455

As can be seen in Table II, Before tuning, the model's performance exhibited 379 true positive (TP) instances, indicating correct predictions of positive cases. However, there were 122 false positive (FP) instances where the model incorrectly classified negative cases as positive. Additionally, the model produced 129 false negative (FN) instances, indicating incorrect predictions of negative cases as positive. Furthermore, there were 370 true negative (TN) instances, representing accurate predictions of negative cases. After tuning, significant improvements were observed in the model's performance. The number of true positive (TP) instances increased from 379 to 459, indicating a better ability to classify positive cases correctly. The false positive (FP) instances decreased from 122 to 42, suggesting a reduction in misclassifying negative cases as positive.

Moreover, the false negative (FN) instances decreased from 129 to 44, indicating an improvement in predicting negative cases as unfavorable. Additionally, the number of true negative (TN) instances increased from 370 to 455, signifying more accurate predictions of negative cases. The tuning process enhanced model accuracy, with more correctly classified positive and negative instances. The reduction in false positive and false negative predictions demonstrated improved precision and recall. Consequently, the tuned model better distinguished positive and negative instances, making it a more reliable data classification and prediction tool. The model's predictions were correct for approximately 91.4% of the cases, indicating a substantial enhancement in predictive performance.

Figure 2 shows some significant variable importance from Rating to cogs. Analyzing the Rating variables reveals valuable insights into Prime Market's customer perceptions and preferences for different product categories. Among these categories, the Food and Beverages products stand out with the highest Rating of 7.11, indicating a high level of satisfaction and positive feedback from customers. This favorable Rating suggests that Prime Market's Food and beverage offerings have been well-received and likely meet or exceed customers' expectations. Following closely, Fashion Accessories products received a commendable rating of 7.09. It suggests that customers are also delighted with the selection and quality of fashion accessories offered by Prime Market. The positive feedback for this category demonstrates the effectiveness of the products in meeting customer needs and preferences.

On the other hand, Electronic Accessories and Sports and Travel products obtained ratings below 7, with scores of 6.92 and 6.91, respectively. These ratings may indicate room for improvement in these categories to better align with customer expectations and enhance overall satisfaction. In the Health and Beauty category, products received a satisfactory rating of 7.00. While this score reflects a generally positive reception, there may still be opportunities further to enhance product offerings and customer experiences within this category. Lastly, Home and Lifestyle products obtained a rating of 6.83. While the Rating is decent, Prime Market may have the potential to explore ways to improve product assortment and customer engagement in this category, aiming for higher levels of customer satisfaction. The insights

from analyzing the Rating variables can guide Prime Market's strategic decision-making. By focusing on product categories that received lower ratings, the company can identify areas for improvement and invest in enhancing the quality and appeal of those offerings. On the other hand, products in highly rated categories can be emphasized and leveraged to strengthen customer loyalty and attract new customers. Understanding customer perceptions through these ratings allows Prime Market to refine its product offerings continuously, tailor marketing strategies, and elevate the overall shopping experience to meet and exceed customer expectations.

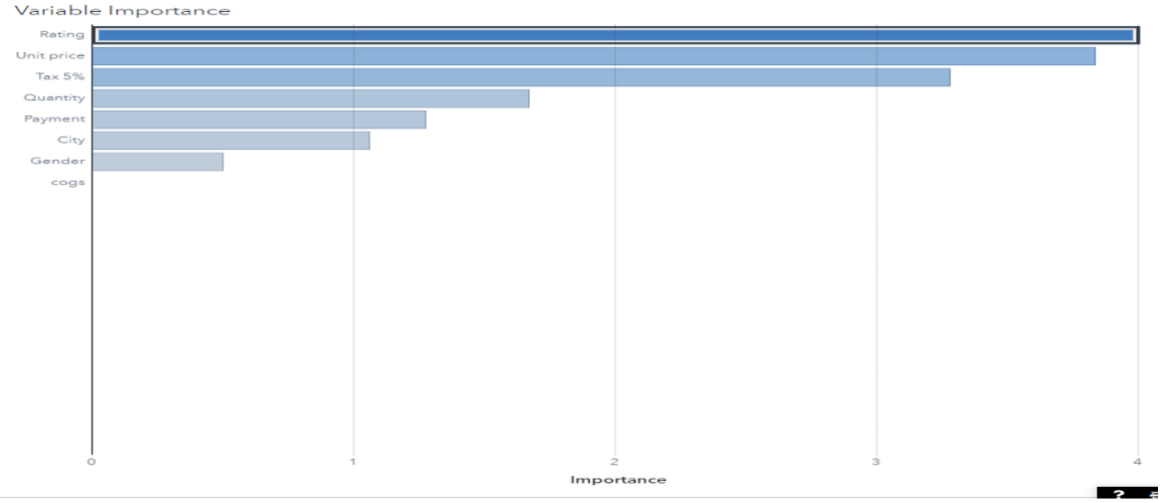

Fig. 2. Variable importancy

IV. Conclusion

Tuning several parameters was conducted to maximize the model's accuracy, resulting in an impressive accuracy of 91.4% or a misclassification rate of 0.0860. These results demonstrate that the Gradient Boosting model performs excellently in accurately classifying customers in the Prime Market. Additionally, the research identified several features that significantly contribute to customer segmentation decisions. For customer segmentation in the Prime Market, the Gradient Boosting model can be effectively employed with appropriately tuned parameters to achieve high accuracy. Moreover, the most influential feature is Rating. Food and Beverages received the highest Rating of 7.11, indicating high customer satisfaction. Fashion Accessories followed closely with a rating of 7.09, demonstrating positive feedback. Electronic Accessories and Sports and Travel received ratings below 7, suggesting areas for improvement. Health and Beauty received a satisfactory rating of 7.00, while Home and Lifestyle scored 6.83. These insights can guide strategic decision-making, allowing Prime Market to enhance products and tailor marketing strategies for improved customer satisfaction and loyalty.

Acknowledgement

We express our gratitude to Multimedia Nusantara University for the financial support and assistance provided by the faculty members.

References

[1] E. L. F. Zineb, R. Najat, and A. Jaafar, An intelligent approach for data analysis and decision making in big data: a case study on e-commerce industry, Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 7, 2021.

[2] M. M. I. Iliadi, Unlocking Customer Insights Through Service Analytics to Improve Customer Experience and Drive Business Success. University of Twente, 2023.

[3] A. Rosário and R. Raimundo, Consumer marketing strategy and E-commerce in the last decade: a literature review, J. Theor. Appl. Electron. Commer. Res., vol. 16, no. 7, pp. 3003–3024, 2021.

[4] A. Wibowo, S.-C. Chen, U. Wiangin, Y. Ma, and A. Ruangkanjanases, Customer behavior as an outcome of social media marketing: The role of social media marketing activity and customer experience, Sustainability, vol. 13, no. 1, p. 189, 2020.

[5] C. France, D. Grace, J. Lo Iacono, and J. Carlini, Exploring the interplay between customer perceived brand value and customer brand co-creation behaviour dimensions, J. Brand Manag., vol. 27, pp. 466–480, 2020.

[6] M.-C. Chiu, J.-H. Huang, S. Gupta, and G. Akman, Developing a personalized recommendation system in a smart product service system based on unsupervised learning model, Comput. Ind., vol. 128, p. 103421, 2021.

[7] N. N. Wilim and R. S. Oetama, Sentiment Analysis About Indonesian Lawyers Club Television Program Using K-Nearest Neighbor, Naïve Bayes Classifier, And Decision Tree, IJNMT (International J. New Media Technol., vol. 8, no. 1, pp. 50–56, 2021, doi: 10.31937/ijnmt.v8i1.1965.10.31937/ijnmt.v8i1.1965

[8] R. Nainggolan, F. A. T. Tobing, and E. J. G. Harianja, Sentiment; Clustering; K-Means Analysis Sentiment in Bukalapak Comments with K-Means Clustering Method, IJNMT (International J. New Media Technol., vol. 9, no. 2, pp. 87–92, 2022.

[9] Y. Pan and L. Zhang, A BIM-data mining integrated digital twin framework for advanced project management, Autom. Constr., vol. 124, p. 103564, 2021.

[10] M. M. Mariani and S. F. Wamba, Exploring how consumer goods companies innovate in the digital age: The role of big data analytics companies, J. Bus. Res., vol. 121, pp. 338–352, 2020.

[11] E. Durner, Effective analysis of interactive effects with non-normal data using the aligned rank transform, ARTool and SAS® university edition, Horticulturae, vol. 5, no. 3,

p. 57, 2019.
[12] R. V McCarthy et al., Applying predictive analytics. Springer, 2022.
[13] D. Y. Kulkarni and L. di Mare, Virtual gas turbines part ii: an automated whole-engine secondary air system model generation, in Turbo Expo: Power for Land, Sea, and Air, American Society of Mechanical Engineers, 2021, p. V02CT34A034.
[14] J. Huber and H. Stuckenschmidt, Daily retail demand forecasting using machine learning with emphasis on calendric special days, Int. J. Forecast., vol. 36, no. 4, pp. 1420–1438, 2020.
[15] S. Punia, K. Nikolopoulos, S. P. Singh, J. K. Madaan, and K. Litsiou, Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail, Int. J. Prod. Res., vol. 58, no. 16, pp. 4964–4979, 2020.
[16] Z. Xu, G. Zhu, N. Metawa, and Q. Zhou, Machine learning based customer meta-combination brand equity analysis for marketing behavior evaluation, Inf. Process. Manag., vol. 59, no. 1, p. 102800, 2022, doi: 10.1016/j.ipm.2021.102800.10.1016/j.ipm.2021.102800
[17] A. Malinin, L. Prokhorenkova, and A. Ustimenko, Uncertainty in Gradient boosting via ensembles, arXiv Prepr. arXiv2006.10562, 2020.
[18] H. Huang, R. Jia, X. Shi, J. Liang, and J. Dang, Feature selection and hyper parameters optimization for short-term wind power forecast, Appl. Intell., pp. 1–19, 2021.
[19] W. Liu, H. Fan, and M. Xia, Credit scoring based on tree-enhanced Gradient boosting decision trees, Expert Syst. Appl., vol. 189, p. 116034, 2022.
[20] R. Shwartz-Ziv and A. Armon, Tabular data: Deep learning is not all you need, Inf. Fusion, vol. 81, pp. 84–90, 2022.
[21] R. Sun, G. Wang, W. Zhang, L.-T. Hsu, and W. Y. Ochieng, A gradient boosting decision tree based GPS signal reception classification algorithm, Appl. Soft Comput., vol. 86, p. 105942, 2020.
[22] T. Zhang et al., Improving convection trigger functions in deep convective parameterization schemes using machine learning, J. Adv. Model. Earth Syst., vol. 13, no. 5, p. e2020MS002365, 2021.