Sales Prediction Analysis Based on SKU in Shopee Indonesia E-Commerce Using GBT (Case Study: Maschere Brand)

Amelia Citra Dewi[1], Arief Hermawan[2], Donny Avianto[3]
[1,2,3]*(Master of Information Technology, Universitas Teknologi Yogyakarta, Indonesia)*
[1]*ameliacitradewi@gmail.com*, [2]*ariefdb@uty.ac.id*, [3]*donny@uty.ac.id*

ABSTRACT: *The retail industry has been changed by the growth of e-commerce, requiring effective inventory control to satisfy customer needs. The Maschere brand, which focuses on fashion products, needs help with forecasting popular color trends and handling SKUs on the Shopee e-commerce platform. This research investigates how the Gradient Boosting Tree (GBT) algorithm can enhance sales forecasting by analyzing SKUs and color trends. The research analyzes sales data from 2021 with 17,543 entries, with a focus on improving stock levels to minimize both shortages and surplus inventory. The GBT model, assessed with MSE, RMSE, and R-squared metrics, shows excellent performance with extensive feature sets. Results from the experiment indicate that the model incorporating all features obtained an MSE of 0.258 and RMSE of 0.508, accounting for 57.3% of the variance in sales. Research shows that neutral colors, specifically the "ELEGANT" collection, are expected to have the highest sales, reflecting a consumer preference for these colors. The study emphasizes the possibility of e-commerce companies using machine learning to improve sales forecasting accuracy, optimize inventory control, and increase customer satisfaction. Future research should concentrate on enhancing feature engineering and improving real-time data processing to enhance predictive accuracy even more.*

KEYWORDS *-e-commerce, gradient boosting tree, predictions, sales, SKU*

I. INTRODUCTION

The development of information and communication technology has caused a significant transformation in many industrial sectors, including the trade sector. In particular, electronic commerce or e-commerce has experienced exponential growth, driven by changes in consumer behaviors, which now prefers to carry out transactions online [1]. E-commerce is viewed as an efficient way to promote, sell, and bundle online services, significantly contributing to customer identification, acquisition, and retention [2]. E-commerce is undergoing swift expansion, impacting every sector of the economy with the online sale of goods and services. Companies in this distribution channel are continually innovating to get more customers. This includes improving website interfaces for a smoother browsing experience, refining logistics for quicker deliveries, diversifying product ranges to offer competitive pricing, and launching promotions to stimulate purchases [3].

In this highly competitive industry, e-commerce companies are required to understand customer needs and preferences in real-time and predict future trends to remain relevant and successful. Predictive analytics, with its ability to process and analyze big data, provides opportunities for companies to identify previously invisible patterns, trends, and relationships, thereby enabling more informed and strategic decision-making [4].

In the context of e-commerce, one of the main challenges is stock and inventory management. Stock availability is crucial. Stock shortages on products that customers are interested in can result in significant lost sales opportunities, while excess stock on less popular products can lead to inventory buildup and unnecessary carrying costs. Overall business operational processes, as well as customer satisfaction can also be affected by the inability to manage stock efficiently. If there is no reliable mechanism to predict stock sales correctly, businesses are at risk of experiencing losses, ranging from loss of consumer trust to financial losses from lost revenue and excess storage costs [5].

Determining the right SKU to prepare is key to maximizing sales and reducing excess stock. This challenge is made more complex by the variability of consumer trends and preferences, especially in the aspect of rapidly changing color variations. Other external factors influence this, such as market trends, seasons, promotions, and so on. Therefore, an analytical method is needed that can accurately predict which stocks need to be provided to meet market demand [6].

The Maschere brand, which will be used as a case study in this research, is an entity that operates in the retail sector via an e-commerce platform in Indonesia, namely Shopee and Tokopedia platform. Operationally, Maschere has a series of products in the form of fabric masks and hijab which are divided into 9 types of color shades, with more than 40 derivative color variations which are completely marketed only online without physical shops/outlets. For example, Maschere divides hijab color types into red and blue which are referred to as parent SKUs, then from the “Parent SKU” it is further divided into derivative color variations starting from light red, dark red, dark blue and light blue, each of which has A separate SKU and is referred to as the “SKU Reference Number”.

To increase sales and operational efficiency, Maschere faced the challenge of accurately predicting popular color trends and determining which SKUs to produce more of, especially in fast-moving products like hijab. This is also done to ensure that the available stock is in an efficient condition and can meet customer demand. Prediction errors can cause not only stock shortages and lost sales but also overstock, which leads to increased carrying costs and potential losses [7].

Based on the background and challenges faced by the Maschere brand, this research aims to explore the significant application of machine learning models in predicting sales. This research proposes the use of a Gradient Boosting Tree (GBT), a powerful and flexible machine learning method which has been proven effective in various predictive applications. By utilizing a series of decision trees in the learning process, this method can accommodate non-linear data and provide accurate predictions. Especially in e-commerce, the use of GBT can increase the accuracy in predicting sales based on SKU analysis and color trends [8].

Through in-depth analysis of SKUs and color trends, Maschere and other similar companies can better identify sales opportunities, make more informed decisions in optimizing stock management, improve operational efficiency, and ultimately increase customer satisfaction and company profits. Apart from that, this research was also carried out as a contribution to the field of data mining and is expected to become a basis for developing better prediction technology for the retail trade industry through e-commerce.

II. LITERATURE REVIEW

GBT is a machine learning technique that is included in the ensemble learning category, especially the boosting method. This technique is used to improve the prediction accuracy of other models, such as Decision Trees, and combine them into a more powerful model through an iteration process. The GBT algorithm gradually minimizes the loss function by adding new trees and correcting the prediction error of the entire model, measured based on the gradient of the loss function, such as the Root Mean Squared Error (RMSE). This process continues until a significant reduction in losses can no longer be achieved or a specified number of trees have been added. GBT has good enough flexibility so this method can be used for regression and classification tasks. Apart from that, GBT is also able to handle overfitting better, with techniques such as reducing the step size (learning rate shrinkage) and subsampling, so this method has more control over overfitting. In terms of effectiveness, GBT is known for its ability to produce accurate predictions, even on complex and high-dimensional datasets [9].

Even so, GBT still has limitations from several sides. First, GBT requires a longer training time compared to other machine learning methods, especially when the number of trees in the model is very large, so the computational complexity becomes more complicated. To maximize the performance of GBT and achieve optimal performance, it is necessary to carefully adjust the parameters, including the number of trees, tree depth, and learning rate [9].

In an effort to increase the accuracy of sales predictions, several studies have been conducted. A comparison of three algorithms at once, K-Nearest Neighbor, GBT, and Random Forest in predicting supermarket sales has been carried out. The results show that the Random Forest algorithm provides the best performance compared to the other two algorithms, with Gradient Boosted models tending to overfit the data set, and K-Nearest Neighbor, although fast, provides the lowest results among the three. The main factors contributing to prediction accuracy include supermarket type, product price, and the year the supermarket opened[10].

Other research has also applied the Gradient Boosting method to predict property prices, with parameters such as year of construction, number of floors, and nearby facilities. The research results indicate that the CatBoostRegressor model, which is based on Gradient Boosting, shows the best results in property price prediction by identifying important variables that influence prices [11].

In research on stock price prediction, GBT combines various natural language processing techniques such as TF-IDF, Word2Vec, CountVectorizer, and Doc2Vec for text data processing, as well as Principal Component Analysis (PCA) for dimensionality reduction has also been carried out. This method shows its effectiveness in improving prediction accuracy by utilizing information from social media and stock market news [12].

In a different study, an innovative multiclass classification approach using a Gradient Boosting Machine enriched with GBM-wFE feature engineering has also been carried out. This approach emphasizes the use of feature engineering to construct new features that improve classification accuracy [13].

In the context of e-commerce, the Extreme GBT algorithm has been used to predict sales during Black Friday[14]. The aim is to understand consumer purchasing behaviors and improve marketing strategies during promotional periods. From the experiments carried out, it was found that ensemble learning techniques, such as bagging and boosting, show significant performance in predicting the number of purchases, with results that can be improved through appropriate hyperparameter tuning and feature engineering techniques[14].

Predicting product purchases in the supply chain using machine learning techniques such as Distributed Random Forest and GBT has also been carried out. This aims to overcome demand uncertainty and optimize inventory management decisions to avoid losses that may arise due to product unavailability. The results show a 20% improvement in prediction model performance with this range approach compared to unprocessed data, indicating that the use of summarized features and oversampling techniques can significantly improve purchase prediction capabilities [15].

However, there has yet to be research that specifically discusses the use of the GBT algorithm in predicting sales based on product SKUs from e-commerce platform sales data, especially from Shopee sales data. The benefit of this research is to predict how many purchases will occur for an SKU identified as a color variation. It is hoped that it can provide information for the Maschere company and other similar companies in determining the amount of inventory in stock. Sales predictions for a product can also become a basis for Maschere in determining the company's marketing strategy.

III. METHODOLOGY

In this research, the methodology used includes several main steps. These steps are designed to ensure that the data collected and the analysis performed can provide accurate and useful insights for the benefit of the business. This methodology includes data collection processes, data preprocessing, feature selection, model training, and validation processes [16]. Fig. 1 explains the flow of this research. Each stage in this methodology has a critical role in ensuring the integrity and effectiveness of predictive analysis.

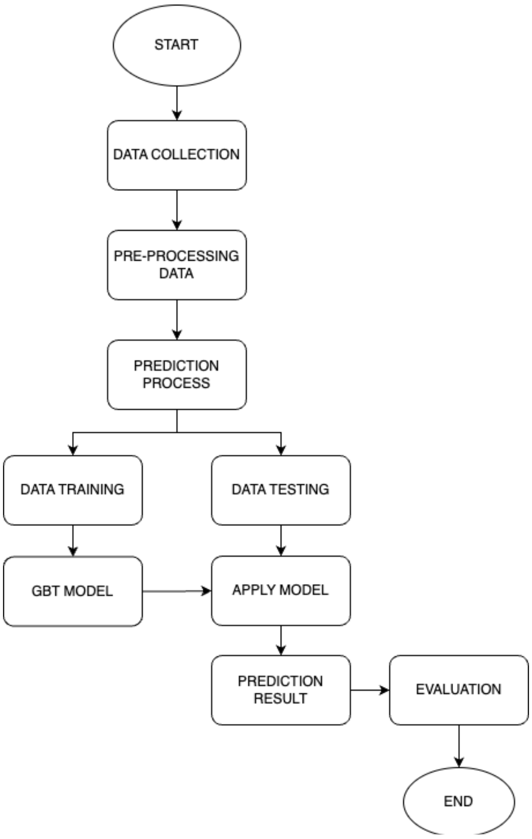

Fig 1. Research Flow Diagram

3.1 Data Collection

The data collection process is the first and crucial step in predictive analysis studies. In this research, data collection is carried out by ensuring the accuracy and completeness of the information. This process also considers aspects of data privacy and security in accordance with company regulations. After collection, data is stored in a structured format to facilitate further analysis.

This research will be carried out using Maschere sales data taken from one of the e-commerce platforms in Indonesia, namely Shopee. This information is taken directly from the Maschere seller dashboard on the e-commerce page, with the collaboration and permission of the Maschere brand.

The sales data taken is sales data for 2021, with a total of 17,543 data in .xls format. In the sales data, there are 44 features as displayed in Table 1, such as order number, order status, cancellation/return status, tracking number, shipping options, shipping time, payment time, payment method, parent SKU, product name, SKU reference number, variation name, original price, and so on.However, due to Shopee's customer data protection policy[17], the customer personal data feature, as stated in number 39-41, appears with hidden data such as “a******d”.

Table 1. Features in Sales Data of Maschere from Shopee Platform

No.	Features	No.	Features
1	Order Number	23	Shopee Discount
2	Order Status	24	Total Ordered Products
3	Cancellation Reason	25	Total Weight
4	Cancellation / Return Status	26	Voucher Sponsored by Seller
5	Tracking Number	27	Bundle Discount (Boolean: Y / N)
6	Shipping Option	28	Bundle Discount (Shopee discount)
7	Drop-off at Counter/Pickup	29	Bundle Discount (Seller discount)
8	Order Must be Shipped Before	30	Shopee Coin Discount
9	Shipping Time Set	31	Credit Card Discount
10	Payment Time	32	Shipping Cost Paid by Buyer
11	Payment Method	33	Estimated Shipping Discount
12	Parent SKU	34	Total Payment
13	Product Name	35	Estimated Shipping Cost
14	Reference SKU Number	36	Buyer's Notes
15	Variation Name	37	Seller’s Notes
16	Original Price	38	Username (hidden)
17	Discounted Price	39	Recipient's Name (hidden)
18	Quantity	40	Phone Number (hidden)
19	Return Quantity	41	Delivery Address (hidden)
20	Total Product Price	42	City/District
21	Total Discount	43	Province
22	Seller Discount	44	Order Completion Time

The parent SKU is a unique identification given to each product sold, in this case referring to the color nuances of the hijab sold by Maschere. Meanwhile, the SKU reference number is an identification of color variations of product derivatives from the parent SKU. Maschere divides its products into 9 main SKUs as reflections of basic colors and more than 40 SKU reference numbers which reflect derivative colors from the base colors, displayed in Fig. 2.

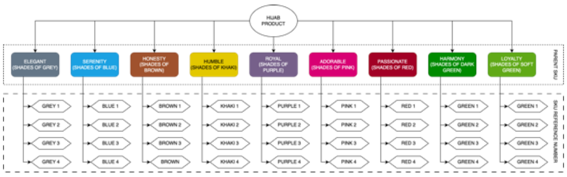

Fig 2. Mapping of Parent SKU and SKU Reference Numbers

In Fig. 2, we can see an example of parent SKU categorization, with the name “ADORABLE” as a name series for products with a pink base color and the SKU reference number with the label “SQUARE-AKAROA”, “SQUARE-PHARLAP” (in the figure referred to as Pink 1, Pink 2, etc.) which is a derivative of the spectrum or different shades of pink for hijab products sold at Maschere. Next, this data continues to the pre-processing stage to be further managed.

3.2 Pre-Processing Data

Sales data from the seller's dashboard will be pre-processed before being used in research. This stage is important, including cleaning data from unnecessary data or features. In this stage, features and data that are unnecessary or irrelevant to the research are removed from the sales data [18]. This is done to ensure that the data used is accurate and to reduce interference with research results [19].

In 2021 sales data, there were 17,543 sales records, which included sales of all products like fabric masks, hijab, and any other products that sold by Maschere, as well as sales with the status “cancelled” or “returned by buyer” data. After going through the pre-processing stage and excluding sales data that were cancelled or returned by customers, the data was reduced to 15,093 records.

In this research, sales predictions are only limited to Maschere hijab products.Furthermore, after taking into account sales that only focus on hijab products, the total remaining data is 9,985 records. These are the remaining data records for our research dataset.

3.3 Prediction Model

The stages of feature selection and model training must be carried out to create a robust and accurate predictive model. Careful selection of features focuses on variables that significantly influence product sales in e-commerce, especially for the Maschere brand. The GBT model is used because of its superiority in processing complex data and providing deep insights through its power to handle categorical and numerical features.

In this research, GBT model is carried out with several parameter optimizations. First, the number of trees tested was 30, 40, and 50 to understand the impact of increasing the number of trees on model accuracy and generalization ability while avoiding significant overfitting. Second, the experiment was conducted in two reliability conditions: reproducible (yes) and non-reproducible (no). Reproducible settings allow the experimenter to repeat consistent results, whereas non-reproducible settings introduce an element of chance that can affect the consistency of the results. Third, we set the maximum tree depth between 3 and 6 to assess the impact of increasing model complexity on performance, considering the risk of overfitting at greater depths. Finally, the number of bins tested was 5, 10, 15, 20, and 25, which were used to determine the data distribution at each node during tree construction.Increasing the number of bins allows for more detailed data exchange but increases computational time and may increase the risk of overfitting. This approach allows experiments to provide deep insight into the optimal configuration of the GBT model when analyzing complex data.

The model training process involves dividing the data into training and testing sets with a 90:10 split data composition, hyperparameter optimization through cross-validation, and evaluating the model using performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. The product sales predictions are projected for one year ahead. Model training is carried out carefully to ensure that the model can generalize well to new data, avoid overfitting, and provide accurate predictions for business decision-making.

Integrating feature selection and model

training is essential in building effective predictive systems. By focusing on the most informative features and applying appropriate machine learning techniques, this research seeks to provide a model that is not only capable of predicting product sales with a high degree of accuracy but also provides an understanding of what factors most influence sales in the e-commerce environment.Through this methodological approach, it can significantly contribute to Maschere's optimization of its sales and marketing strategies in the future.

IV. Results and Analysis

4.1 GBT Model Training and Testing

Implementing the GBT model in this research will be carried out using RapidMiner Studio software. Fig. 3 shows the overall process of using the GBT model to predict the number of products in subsequent sales. At this stage, the workflow begins by reading sales data, setting role data, and processing the optimization model that has been determined.

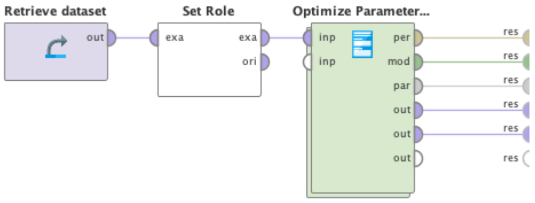

Fig 3. Implementation of the GBT Model on Maschere Sales Data

Fig. 4 shows the process of training and testing the GBT model. The training model is used on test data and evaluated through performance metrics at this stage. The "Store" operator stores the trained model for further use.

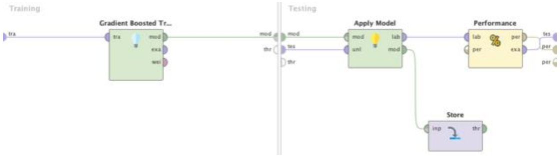

Fig 4. GBT Model Training and Testing

Next, Fig. 5 illustrates the cross-validation process used to evaluate the model, demonstrating how the model is tested repeatedly using different data segments. This ensures the model's performance is consistent and reliable across varied data sets.

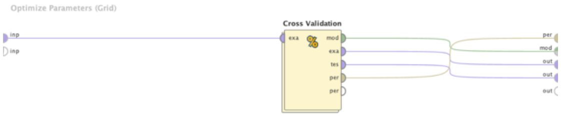

Fig 5. Model Evaluation through Cross-Validation

Several experiments will be carried out using different data features from sales data. In experiment 1, training and testing the GBT model used all features from sales data with a total of 44 features. In experiment 2, training and testing the GBT model used only 7 features that were considered the most relevant, including Payment Time, Parent SKU, SKU Reference Number, Variation Name, Original Price, Discount Price, and Quantity.

4.2 Model Performance Analysis

In this research, the GBT model is optimized with parameters to improve sales predictions based on SKU and color trends on e-commerce platforms. The model performance results from the two experiments are summarized in Table 2, in which "Score 1" represents the result for Experiment-1 with 44 data features, and "Score 2" represents the result for Experiment-2 with 7 data features.

Table 2. Metrics of GBT Model

Metrics	Score 1	Score 2
MSE	0.258	0.633
RMSE	0.508	0.796
R-squared	0.573	0.102
Mean Residual Deviance	0.258	0.634
Mean Absolute Error	0.129	0.213
Root Mean Squared Log Error	0.082	0.159

The Mean Squared Error (MSE) value in Experiment-1 is 0.258 and 0.633 in Experiment-2. This indicates that the error rate in model predictions in Experiment-1 is lower than in Experiment-2. The MSE value, which measures the average quadratic error between the prediction and the actual value in Experiment-1, shows the effectiveness of the model in estimating data with a high level of accuracy. The Root Mean Squared Error (RMSE) in Experiment-1 was 0.508, giving an idea of errors on the same scale as the target variable, emphasizing lower prediction errors and better performance compared to Experiment-2 with an RMSE value of 0.796.

Meanwhile, the R-Squared value in Experiment-1 was 0.573, indicating that 57.3% of the variability in the dataset could be explained by the model, emphasizing the relevance of the model to the predicted data. In Experiment-2, the R-Squared value was only 0.102, which can be interpreted as meaning that the model can only explain 10.2% of the total variability in sales data. This can also be seen from the Mean Residual Deviance value, which reflects the average error in predictions made by the model in Experiment-1, which is lower than in Experiment-2.

Apart from that, the Mean Absolute Error (MAE) value in Experiment-1 of 0.129 is also lower than Experiment-2 of 0.213, providing a clear picture of the lower prediction error in Experiment-1 compared to Experiment-2. The Root Mean Squared Log Error (RMSLE) value was 0.082 in Experiment-1, indicating that the model with a complete number of features had a better predictive ability, was more stable, and had a lower error rate than in Experiment-2, which only used 7 features.

This performance comparison emphasizes the importance of feature selection and utility in building predictive models. Models that use more comprehensive datasets and features show significantly better performance, emphasizing the fact that the additional information held by these features allows the model to make more accurate and efficient predictions. The lack of critical features in Experiment-2 reduces the model's ability to capture the complexity and variability in sales data, causing model performance to decrease significantly compared to Experiment-1.

Table 3. Optimal Parameters for GBT Model

Parameters	Score 1	Score 2
Number of Trees	50	50
Reproducible	no	yes
Maximal Depth	4	4
Number of Bins	15	5

In Table 3, the best parameter for the number of trees in both experiments was 50 trees. This shows that this number of trees is considered the most optimal. Although the use of many trees can improve model accuracy, it also affects computing time. In Experiment-1, the computing time required to train the model was 13 minutes and 18 seconds. Meanwhile, in Experiment 2, even though both used 50 trees, the computing time required was only 2 minutes 14 seconds. The best maximum depth in both experiments was 4. This depth limits the model from being too complex, which can help avoid overfitting. This depth indicates an effort to keep the model simple enough but still deep enough to capture important patterns in the dataset.

Meanwhile, the best parameter Number of Bins in both experiments shows significant differences. In Experiment-1, the number of bins used was 15, while in Experiment-2, it was only 5 bins. The larger number of bins in Experiment-1 allows the model to partition the data in more detail during the learning process. Meanwhile, the use of fewer bins in Experiment-2 could be caused by lower model complexity due to the smaller number of features used, and this resulted in a decrease in accuracy in model training.

4.3 Prediction Performance Analysis

Evaluation of the GBT model prediction results was carried out using the 10-Fold Cross Validation approach. In this context, the data is divided into ten parts, where each part is used as a testing set while the rest is used as a training set. This approach ensures that the model is thoroughly tested on all available data, resulting in more stable and reliable performance estimates. Below is an in-depth analysis of the evaluation performance metrics resulting from this process. Table 4 displays the performance prediction results of the GBT model.

Table 4. Performance Vector of GBT Model

Metrics	Score 1	Score 2
Squared Error	0.282	0.698
RMSE	0.464	0.762
Correlation	0.977	0.134
R-squared	0.956	0.024
Predictive Average	1.113	1.12

The Squared Error value in Experiment-1 was 0.282, lower than in Experiment-2, which was 0.698. This shows that the prediction performance in Experiment-1 is closer to the actual value and has a lower error rate. This is also reinforced by the RMSE value, which is also lower in Experiment-1, namely 0.464. The significant difference in the

RMSE value, which provides a measure of the magnitude of the error in the same units as the predicted value, explains that Experiment-1 is superior to Experiment-2.

The correlation coefficients between the two experiments also reveal a clear difference: Experiment-1 has a very high correlation of 0.977, while Experiment-2 has a very low correlation of 0.134. This measure shows the extent to which the actual data and the anticipated values match linearly. The model is able to capture and forecast the underlying patterns in the dataset, as evidenced by the high correlation in Experiment-1, which highlights a strong linear relationship. On the other hand, low correlation in Experiment-2 indicates that there may be a discrepancy between the model's predictions and the actual values. This discrepancy could be the result of overfitting, inadequate feature selection, or missing external variables during the modeling process.

R-squared supports this interpretation, whereas in Experiment-1, the R-squared value of 0.956 is superior to Experiment-2, which was only 0.024. This value indicates that the model can explain 95.6% of the variability in the response variable. This high value indicates excellent model performance. The Predictive Average shows a small difference between the two experiments: 1.113 in Experiment-1 and 1.120 in Experiment-2. A comparison of these two results revealed that Experiment-1 had better predictive quality and dependability than Experiment-2, even though this metric by itself would indicate comparable average predictive outcomes.

Overall, the application of the GBT model to Maschere's e-commerce sales data on Shopee shows accurate prediction results when using all the features in the dataset. This model effectively captures and predicts patterns in sales data. On the other hand, models that use fewer data features are unable to provide accurate prediction results. Continuous adjustment and evaluation of models in response to varying data conditions is an inherent challenge in predictive modeling in the e-commerce space.

4.4 Sales Prediction Results

Using the GBT model, it is predicted that the highest sales of Maschere products will be reviewed based on the parent SKU (color shades) and SKU reference number (hijab product color) from the order of best-selling to least-selling as listed. Table 5 shows that the highest sales are predicted to come from the parent SKU “ELEGANT” series, with predicted sales reaching 2,356 pcs from this product series. It also shows that products with the parent SKU “ELEGANT”, which are hijab with neutral color shades, have high demand in the market. Complementing the parent SKU “ELEGANT”, the largest contribution of hijab color variations from this series comes from the SKU reference number “SQUARE-SPACEBLACK”, with predicted sales of this specific color reaching 1,083 pcs. This shows that the black hijab variation is one of the color choices most popular with customers.

Table 5. Sales Prediction Results with GBT Model

Parent SKU	Qty	SKU Reference Number	Qty
ELEGANT	2356	SQUARE-SPACEBLACK	1083
ROYAL	2209	SQUARE-TURKISHROSE	258
SERENITY	1858	SQUARE-NAVY	708
HONESTY	1792	SQUARE-SORELBROWN	321
ADORABLE	941	SQUARE-PHARLAP	118
HUMBLE	672	SQUARE-SOFTAMBER	163
HARMONY	485	SQUARE-HIMALAYA	143
LOYALTY	474	SQUARE-SPANISHGREEN	224
PASSIONATE	397	SQUARE-BURGUNDY	248

Although the highest sales prediction is for neutral colors, several other parent SKUs, such as “ROYAL” and “SERENITY” also have significant sales numbers, although lower than “ELEGANT” series.

It can be noticed that sales of other hijab products outside of neutral colors significantly contribute to sales of hijabs, including parent SKUs “ROYAL” and “HONESTY”. In these two main SKUs, the majority of predicted best-selling hijab variations based on the SKU reference number are hijab with soft / light colors, such as “SQUARE-TURKISHROSE”, which is light purple and “SQUARE-SORELBROWN”, which is a lightbrown color.

Based on these results, the analysis results can be used to increase the stock of hijab in basic colors such as black and navy to meet high market demand. In addition, hijab production is prioritized for the main SKUs “ROYAL”, “HONESTY”, and “ADORABLE”, which have a significant sales contribution. Hijab color variations, dominated by soft or light colors, also need to be increased in production, while promotions can be increased to maximize sales of dark color hijab. From the prediction results, this information can be used to plan more effective sales and marketing strategies for the Maschere brand.

V. Conclusion

The research used the Gradient Boosting Tree (GBT) method to predict sales based on SKU and color trends in e-commerce, especially for the Maschere brand. With a focus on SKU-based analysis of the Maschere brand on the Shopee Indonesia platform, this research highlights the potential of the Gradient Boosting Tree (GBT) model for sales prediction in the e-commerce industry. The research findings indicate that the GBT model exhibits noteworthy predictive accuracy when fine-tuned using extensive feature sets. Nevertheless, as the number of features decreases, the model's effectiveness does as well, highlighting the significance of broad data use in predictive analytics.

The results show that although the GBT model is capable of handling complicated data structures, the size and depth of the dataset affect its predictive power. With a full feature set, the model yielded an MSE of 0.258 and an RMSE of 0.508, which accounted for 57.3% of the variation in sales data. This demonstrates how, given enough data, the GBT model can identify complex patterns in sales trends.

The research also pinpoints important sales patterns for the Maschere brand, namely the increased desire for items with muted color schemes, such as those in the "ELEGANT" series. According to this insightful observation, increasing stock levels for color variations that are in great demand can improve sales performance and customer happiness. This information offers useful information for optimizing inventory and marketing strategies.

However, this research also underscores the GBT model's limitations, specifically its computational complexity and the necessity for significant feature selection and parameter tuning. Further research might focus on fine-tuning the model, investigating advanced feature engineering techniques, and incorporating other data sources to improve forecast accuracy. More research into adaptive learning models and real-time data processing could yield forecasts that are more responsive and dynamic.

In conclusion, the use of the GBT model in e-commerce sales forecasts demonstrates encouraging results, enabling a solid framework for inventory management and strategic planning. Businesses like Maschere can enhance their operational efficiency and profitability and more adeptly traverse the competitive e-commerce landscape by resolving their limitations and leveraging their strengths.

References

[1] S. Habib and N. N. Hamadneh, “Impact of Perceived Risk on Consumers Technology Acceptance in Online Grocery Adoption amid COVID-19 Pandemic,” *Sustainability*, vol. 13, no. 18, p. 10221, Sep. 2021, doi: 10.3390/su131810221.

[2] P. Pathmanathan *et al.*, “The Benefit and Impact of E-Commerce in Tourism Enterprises,” *2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pp. 193–198, 2021, doi: 10.1109/ICSCEE50312.2021.9497947.

[3] S. Iraten, F. Bouguedour, M. Bouchetara, M. Zerouti, and H. Mahmoudi, “Factors Determining the Adoption of Online Shopping by the Algerian Consumer,CASE: Jumia Company,” *MER*, vol. 7, no. 1, pp. 33–48, Mar. 2022, doi: 10.24818/mer/2022.02-03.

[4] D. Agnani, A. Bavkar, S. Salgar, and S. Ahir, “Predicting E-commerce Sales & Inventory Management using Machine Learning,” *ITM Web Conf.*, vol. 44, p. 03040, 2022, doi: 10.1051/itmconf/20224403040.

[5] N. Jain and T. F. Tan, “M-Commerce, Sales Concentration, and Inventory Management,” *Manufacturing and Service Operations Management*, vol. 24, no. 4, pp. 2256–2273, 2022, doi:

https://doi.org/10.1287/msom.2021.1071.

[6] T. Tang, "Analysis and Demand Forecasting Based On e-Commerce Data," in *2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, May 2023, pp. 64–68. doi: 10.1109/ICAIBD57115.2023.10206072.

[7] C. S. Ibrahima, J. Xue, and T. Gueye, "Inventory Management and Demand Forecasting Improvement of a Forecasting Model Based on Artificial Neural Networks," *j. of manag. sci. & eng. res.*, vol. 4, no. 2, pp. 33–39, Aug. 2021, doi: 10.30564/jmser.v4i2.3242.

[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Statist.*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[10] R. Odegua, "Applied Machine Learning for Supermarket Sales Prediction," 2020.

[11] N. Fedorov and Y. Petrichenko, "Gradient Boosting–Based Machine Learning Methods in Real Estate Market Forecasting:," in *Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)*, Ufa, Stavropol, Khanty-Mansiysk, Russian Federation: Atlantis Press, 2020. doi: 10.2991/aisr.k.201029.039.

[12] J. Yang, C. Zhao, H. Yu, and H. Chen, "Use GBDT to Predict the Stock Market," *Procedia Computer Science*, vol. 174, pp. 161–171, 2020, doi: 10.1016/j.procs.2020.06.071.

[13] R. M. Nabi, S. Ab. M. Saeed, and H. Harron, "A Novel Approach for Stock Price Prediction Using Gradient Boosting Machine with Feature Engineering (GBM-wFE)," *KJAR*, vol. 5, no. 1, pp. 28–48, Apr. 2020, doi: 10.24017/science.2020.1.3.

[14] N. Duong-Trung, D. Tan, T.-D. Luu, and H. Huynh, "Black Friday Sale Prediction via Extreme Gradient Boosted Trees," Jun. 2019. doi: 10.15625/vap.2019.0007.

[15] S. Islam and S. H. Amin, "Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques," *J Big Data*, vol. 7, no. 1, p. 65, Dec. 2020, doi: 10.1186/s40537-020-00345-2.

[16] A. A. Maryoosh and E. M. Hussein, "A Review: Data Mining Techniques and Its Applications," *IJCSMA*, vol. 10, no. 3, pp. 1–14, Mar. 2022, doi: 10.47760/ijcsma.2022.v10i03.001.

[17] "Perubahan Tampilan Informasi Pembeli | Pusat Edukasi Penjual Shopee Indonesia." Accessed: Jan. 16, 2024. [Online]. Available: https://seller.shopee.co.id/edu/article/14910

[18] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, p. 652801, Mar. 2021, doi: 10.3389/fenrg.2021.652801.

[19] M. Arhami and M. Nasir, *Data Mining - Algoritma dan Implementasi*. Andi, 2020.