Comparative Analysis of Linear Regression Algorithms and Random Forest for Skincare Product Sales Prediction of Skincare Products on the Tokopedia Marketplace

Brian Colin Chang[a], Widodo Tjiawie[b], Jervis Patrick[c], Justhine Thorino[d], Delima Sitanggang[e]
[a,b,c,d,e]*Universitas Prima Indonesia*
Corresponding Author :
[e]*delimasitanggang@unprimdn.ac.id*

ABSTRACT
This study aims to evaluate and compare the performance of Linear Regression and Random Forest algorithms in predicting skincare product sales on the Tokopedia platform. Accurate sales predictions play a strategic role for businesses, particularly in pricing policy formulation, product positioning, and more efficient inventory management. The data used included 7,217 skincare products with six main variables, which were then expanded into 21 features through preprocessing and feature engineering stages, including product type extraction, main benefits, seller characteristics, location tier, and price position category. The model was developed and tested using a training and testing data split with a ratio of 80:20. The analysis results showed that the Random Forest algorithm performed better than Linear Regression. In the test data, Random Forest achieved an R^2 value of 0.2115 with an RMSE of 32,754.23 and an MAE of 12,073.59. In contrast, Linear Regression only obtained an R^2 of 0.0245 with an RMSE of 36,431.74 and an MAE of 14,688.99. Although the R^2 value is relatively low, the results are still realistic considering the complexity of e-commerce sales behavior, which is influenced by various external factors. Feature importance analysis shows that price and price per unit are the most dominant factors, followed by rating, number of product benefits, brand category, and product type. Based on these findings, Random Forest is recommended as a more suitable model to support skincare product sales predictions on Tokopedia, mainly due to its ability to capture non-linear patterns and interactions between features. This study is expected to contribute to the development of data-driven marketing strategies, particularly in pricing, multi-benefit product development, and skincare portfolio management on marketplace platforms.

Keywords : Sales Prediction, Random Forest, Linear Regression, E-commerce, Skincare Products.

INTRODUCTION

The beauty industry in Indonesia has experienced significant growth over the past decade. Public awareness of the importance of skin care has made skincare products a primary need, rather than just a lifestyle accessory (Ababil et al., 2022). According to data from the Ministry of Industry, skin care products account for approximately 32% of the total national cosmetics industry. This trend has further accelerated with the ease of access to e-commerce,

which has expanded the sales reach of various skincare brands, both local and international (Adi & Sudianto, 2022).

Marketplaces like Tokopedia have become one of the dominant platforms for distributing skincare products in Indonesia (Ashari & Sadiki, n.d.). As digital transactions and online shopping become easier, competition among sellers is getting tougher. Sellers need to understand consumer behavior patterns and predict sales levels to develop the right marketing strategies. However, fluctuations in demand, price variables, product ratings, customer reviews, and dynamic promotional trends make the sales prediction process complex and difficult to do manually (Dewi et al., 2022).

In this context, machine learning methods provide an effective solution for analyzing and predicting bulk sales data from marketplaces (Fitri, 2023). Regression models, such as Linear Regression and Random Forest Regression, are widely used in prediction studies because they are able to study the relationship between various variables that determine sales, such as price, ratings, and number of reviews, with a relatively high level of accuracy (Hamdanah & Fitrianah, 2021). Linear Regression offers advantages in interpretation and ease of analysis of linear relationships between variables, while Random Forest provides more stable performance for complex non-linear data (Hidayat et al., 2025).

Previous research has demonstrated the effectiveness of regression models in various business contexts. For example, research by Kusumayanti (2025) shows that comparing regression algorithms can increase production estimation accuracy by up to 20%, while similar research on e-commerce platforms reports a 25% increase in sales predictions using the Random Forest method compared to classic models (Krisnawati, 2022). Thus, the application of regression models to skincare sales data on Tokopedia is a relevant strategy for understanding market dynamics scientifically.

This study aims to analyze sales forecasts for skincare products on Tokopedia by comparing Linear Regression and Random Forest algorithms. The resulting model is expected to provide an accurate predictive overview of sales potential and serve as a basis for data-driven business decisions. Additionally, the results of this study are expected to contribute to the scientific literature on the application of machine learning in sales analysis in digital marketplaces in Indonesia.

METHODS

Type of Research

This research is quantitative research with an approach of applying data analysis methods to build and compare sales prediction models for skincare products using Linear Regression and Random Forest algorithms (Kusumawardani et al., 2023). Sales data was collected through web scraping from the Tokopedia marketplace, covering variables such as price, number of ratings, reviews, store name, category, and total products sold. The research stages included data collection and cleaning, modeling using both algorithms, and model evaluation using the Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-Squared (R^2) metrics to assess prediction accuracy. The implementation was carried out using Python and the scikit-learn library (Nugraheny et al., 2023). This study aims to apply regression methods to

produce a prediction model that can support business decision-making in the marketplace, especially for skincare products.

Data Preprocessing

Data from the Tokopedia marketplace obtained through web scraping was first processed to make it ready for analysis and modeling. The preprocessing stage includes cleaning the data from duplication and input errors, handling missing values, and transforming variables to have the appropriate format and data type (Romadhon & Putra, 2024). Numeric variables, such as price and number of ratings, are normalized to equalize the scale, while categorical variables, such as product name, brand, and category, are converted to numeric format through encoding techniques. The data is then divided into a training set and a testing set with a ratio of 80:20 to ensure valid model evaluation (Rusdy, 2022). This process aims to produce high-quality data so that the Linear Regression and Random Forest algorithms can provide accurate and reliable sales predictions.

Application of the Linear Regression Algorithm

The Linear Regression method is used to model the linear relationship between the dependent variable, namely the number of products sold, and a number of independent variables such as product price, rating, and number of customer reviews (Sianturi et al., 2024). In mathematical form, the multiple linear regression model is written as follows.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \quad (1)$$

Where:

1. Y is the target variable in the form of a prediction of product sales.
2. β_0 is a constant or intercept that indicates the base value of$Y$ when all independent variables are zero.
3. β_i are the coefficients of each independent variableX_i that describe how much influence the variable has on$Y$.
4. ϵ is the error term representing the variability that cannot be explained by the variables$X$

The parametersβ_i are estimated using the least squares method so as to minimize the squared difference between the predicted values and the actual values.

Application of the Random Forest Algorithm

Random Forest Regression is an ensemble learning technique that utilizes a collection of decision trees to produce more stable and accurate predictions by combining the output of each individual tree (Tombeng & Ardian, 2021. The final prediction of this method,$\hat{Y}$, is calculated as the average of the prediction results in each tree as follows.

$$\hat{Y} = \frac{1}{M} + \sum_{m=1}^{M} T_m(X) \quad (2)$$

Where:

1. M indicates the number of trees in the forest.
2. $T_m(X)$ is the prediction value generated by the$m$ th tree based on the variable input$X$.

This method is capable of handling data with non-linear patterns and complex relationships between variables, thus providing an advantage over traditional linear regression in the context of dynamic e-commerce sales data.

Analysis and Evaluation of Results

After obtaining prediction results from the Linear Regression and Random Forest algorithms, model performance will be evaluated using the Mean Squared Error (MSE), Mean Absolute Error (MAE), and coefficient of determination (R^2) metrics. The analysis will include a comparison between the predicted values and actual sales values to assess each model's ability to handle marketplace data.

Furthermore, the advantages and limitations of each algorithm will be identified, including their capacity to model linear and non-linear relationships between variables. This quantitative comparison process aims to determine the method that shows higher accuracy and more stable predictions, so that it can be used as a basis for developing a Tokopedia data-based skincare product sales prediction system.

RESULTS AND DISCUSSION

Data Identification

In this study, the data used consisted of 7,217 skincare products collected from the Tokopedia marketplace, focusing on the facial care product category. The dataset included six original variables that reflected important product characteristics and consumer behavior on the e-commerce platform. These variables included store name, store location, product name, product price, number of units sold, and product ratings from customers. Each of these variables has a significant contribution in understanding the dynamics of skincare product sales on the marketplace and plays an important role in building an accurate sales prediction model.

Table 1 . Dataset Details

No	Store	Location	Product Name	Price (IDR)	Sold
1	Wardah Official	Jakarta	Wardah Micellar Water	39,000	2,450
2	Cetaphil Store	Surabaya	Cetaphil Cleanser	85,000	5,230
3	Emina Beauty	Bandung	Emina Acne Solution	45,000	1,230
4	Barber Daily	Jakarta	Barber Daily Foam	32,000	850
5	Innisfree	Medan	Innisfree Green Tea	72,000	3,400

...	...	...	...			...	...
7217	angeltacikskincare	Sidoarjo	Angel Lotion	Tacik	Body	150,000	100,000

Initial Data Description

Before further analysis, the dataset underwent data exploration (EDA) to understand the characteristics and distribution of the data. The results of the exploration showed the following characteristics of the dataset:

1. Number of data rows: 7,217 skincare products from various categories and sellers on Tokopedia
2. Number of original columns/variables: 6 (Store, Location, Product Name, Price, Sold, Rating)
3. Data retention: 100% - all 7,217 rows were retained without any deletions
4. Price range: IDR 100 to IDR 20,000,000 (from sample products to premium luxury skincare)
5. Rating range: 0 to 5.0 (1-5 star scale, including products not yet rated)
6. Sales Range: 0 to 500,000+ units sold
7. Missing Values: Store (1), Location (1), Sold (572), Rating (1,067)
8. Store Characteristics: Official Store, Regular Seller from various cities in Indonesia

This diverse data distribution provides a comprehensive representation of the Indonesian skincare market, covering products from various price positions, brand origins (local vs. international), and seller types. The decision to retain all 7,217 rows without removing the missing values was strategic because information about zero-sales products and products with incomplete data is still valuable for model learning.

Data Cleaning and Handling Missing Values

During the preprocessing stage, data cleaning was performed by filling in 0 for missing values in order to retain all available product information. This strategy was chosen because missing values in the Sold and Rating columns reflect the actual conditions in the marketplace, where products have not yet been sold or have not yet received reviews from consumers. This information is very valuable for training the model to understand the characteristics of new products. The results of data cleaning show:

1. Total initial data: 7,217 products
2. Missing value - Sold: 572 rows → FILL with 0 (product has not been sold)
3. Missing value - Rating: 1,067 rows → FILL with 0 (product has no rating)
4. Missing value - Store: 1 row, Location: 1 row → FILL with median/mode
5. Product Name: 0 missing values (complete data)
6. Price: 0 missing values (complete data)
7. Data retention: 100% - all 7,217 rows retained
8. FILL 0 strategy justified because 583 products (8.1%) with zero sales provide insight into market entry barriers

Table 2. Product Categorization Based on Sales Volume

Sales Category	Number of Products	Percentage	Description
Best Sellers (>100K)	48	0.7	Featured products with very high volume
Good Sellers (10K-100K)	504	7.0	Products that perform well in *the marketplace*
Low Sellers (<10K)	6,082	84.3	Products with low to moderate sales
No Sales (0 units)	583	8.1	New products or products that have not achieved *market fit*

Feature Engineering

Feature engineering is the process of adding new variables through the development of existing dataset variables. From the 6 original variables, 7 new variables were extracted and created through feature engineering to improve the quality and number of features used in modeling. After feature engineering and one-hot encoding, there were a total of 21 numerical features that were used to train both algorithms.

Table 3. New Variables from Feature Engineering

New Features	Data Type	Description	Characteristics
Brand Recognition	Categorical	Extraction of brand names from product names (Wardah, Cetaphil, etc.)	1,533 unique brands
Product Type	Categorical	Format type: Cleanser (20.1%), Gel (5.9%), Foam (4.4%), Other (66.5%)	8 categories
Benefit Keywords	Multiple Binary	Brightening (45%), Hydrating (20.3%), Acne (15.7%), Oil Control (6.7%), Anti-Aging (3.4%)	5 benefit types
Price Positioning	Categorical	Budget (<$36.5K), Mid-Range ($36.5K-$66K), Premium ($66K-$135K), Luxury (>$135K)	Balanced 25%
Seller Type	Categorical	Regular (92.2%), Official Store (5.9%), Wholesale (1.4%), Pharmacy (0.5%)	4 categories
Location Tier	Categorical	Tier 1 Metro (32.6%), Tier 2 Urban (28.6%), Tier 3 Regional (38.7%)	3 tiers
Price Per Unit	Numerical	Price per unit volume (Rp/ml) - Range Rp7 to Rp497,081	48.9% have data
Sales Category	Number of Products	Percentage	Description
Best Sellers	48	0.7	Featured products

(>100K)			with very high volume
Good Sellers (10K-100K)	504	7.0	Products that perform well in the marketplace
Low Sellers (<10K)	6,082	84.3	Products with low to moderate sales
No Sales (0 units)	583	8.1	New products or products that have not achieved market fit

Top 10 identified brands: Paket (294 products), Glad2Glow (290), Toner (249), Wardah (140), Kahf (132), Skintific (130), Cetaphil (116), Ms Glow (115), Nivea (109), and Marina (95). Insights from feature engineering show that the 'Brightening' benefit is the most common (45% of products), followed by 'Hydrating' (20.3%), indicating Indonesian consumers' market preference for skincare with brightening and skin hydration functions.

Feature Selection and Data Encoding

In this process, data encoding will be performed on the dataset that is not yet numerical so that the data can be applied to the algorithm that will be used. The features used in the modeling stage include 21 features resulting from feature engineering, consisting of 6 numerical features (Price, Rating, Rating_Count, Total_Benefits, Price_Per_Unit) and 15 categorical features that have been converted through one-hot encoding to ensure compatibility with the algorithm to be applied.

```
#pemilihan fitur
X = df_clean[[
    'Harga','Rating',  # Original  fitur
'Total_Benefits', 'Price_Per_Unit',  # Numerical engineered
]]

# Categorical features yang akan di-encode
categorical_features = ['Brand_Type', 'Product_Type', 'Price_Positioning', 'Seller_Type']

# One-Hot Encoding untuk categorical features
X_categorical = pd.get_dummies(df_clean[categorical_features], drop_first=True)

# menggabungkan numerikal dan kategorikal hasil encoding
X = pd.concat([X, X_categorical], axis=1)

# Handle NaN di Price_Per_Unit
X['Price_Per_Unit'].fillna(X['Price_Per_Unit'].median(), inplace=True)

# Target variable
y = df_clean['Terjual_Numeric']

print("\n🔍 Pembagian Data Training dan Testing...")
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=42
)
```

Figure1 . Encoding Process

Table 4. Descriptive Statistics of Numeric Features

Features	Count	Mean	Std Dev	Min	Max
Price	7,217	Rp136,083	Rp498,013	IDR 100	IDR 20,000,000
Rating	7,217	4.14	1.74	0.0	5.0

Sold_Numeric	7,217	10,918	44,745	0	500,000
Total_Benefits	7,217	0.91	0.92	0	5
Price_Per_Unit	7,217	Rp3,599	Rp19,200	IDR 7	IDR 497,081

Split and Training Dataset

The dataset is split into a training set and a testing set with an 80:20 ratio using a random split strategy with random_state=42 to ensure reproducibility. The training set consists of 5,773 samples, while the testing set consists of 1,444 samples. The distribution of target values (Sold_Numeric) in both sets is very balanced, with a training set mean of 11,381 units and a testing set mean of 9,071 units, indicating an effective split strategy for model generalization.

Table 5 . Distribution of Training and Testing Datasets

Metrics	Training (80%)	Testing (20%)	Total
Number of Samples	5,773	1,444	7,217
Percentage	80%	20%	100%
Mean Sold	11,381	9,071	10,918
Sold Range	0 - 500K	0 - 500K	0 - 500K
Std Dev	45.620	41.234	44,745

Linear Regression Results

The Linear Regression model was developed using 5,773 data samples in the training stage and involved 21 predictor variables as input features with the following results.

```
#melatih model
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
print("✓ Model Linear Regression berhasil dilatih!")

print(f"\n Top 10 Fitur dengan Koefisien Terbesar (Linear Regression):")
feature_names = X.columns.tolist()
coef_df = pd.DataFrame({
    'Feature': feature_names,
    'Coefficient': lr_model.coef_
}).sort_values('Coefficient', key=abs, ascending=False).head(10)

for idx, row in coef_df.iterrows():
    print(f"  {row['Feature']}: {row['Coefficient']:.6f}")

print(f"\n  - Intercept: {lr_model.intercept_:,.2f}")

# Prediksi
y_pred_lr_train = lr_model.predict(X_train)
y_pred_lr_test = lr_model.predict(X_test)
```

Figure 2. Application of the Linear Regression Algorithm

Table 6 . Linear Regression Results

Evaluation Metrics	Training Set	Testing Set
R^2 Score	0.0357 (3.57%)	0.0245 (2.45%)
RMSE	45,650.46 units	36.431.74 units
MAE	16,542.62 units	14,688.99 units
MSE	2,083,964,677 .77	1,327,271,967 .47

The test results show that the R^2 value is 0.0245, which means that the model can only explain 2.45% of the variation in the sales variable. This low R^2 value indicates that the linear relationship between the predictor variables and sales is still very weak. This finding suggests that the actual relationship pattern is likely to be non-linear or involve variable interactions that cannot be captured by a simple linear model.

The five features with the largest coefficients in influencing sales are, in order: Product_Type_Mousse (15,282.23), Seller_Type_Regular_Seller (13,131.81), Product_Type_Gel (11,977.86), Total_Benefits (5,476.25), and Price_Positioning_Mid-Range (5,323.53). Positive coefficients indicate that an increase in these features is associated with an increase in sales. Conversely, negative coefficients—for example, in Product_Type_Scrub (-7,255.13)—indicate a downward effect on sales value.

Random Forest Results

The Random Forest model in this study was built using 100 decision trees, with the training process conducted on 5,773 data samples.

```
#Pelatihan model
rf_model = RandomForestRegressor(
    n_estimators=100,
    max_depth=15,
    min_samples_split=5,
    min_samples_leaf=2,
    random_state=42,
    n_jobs=-1
)
rf_model.fit(X_train, y_train)

print(f"\n Top 10 Fitur dengan Feature Importance Tertinggi (Random Forest):")
importance_df = pd.DataFrame({
    'Feature': feature_names,
    'Importance': rf_model.feature_importances_
}).sort_values('Importance', ascending=False).head(10)

for idx, row in importance_df.iterrows():
    print(f"  {row['Feature']}: {row['Importance']:.4f} ({row['Importance']*100:.2f}%)")

# Prediksi
y_pred_rf_train = rf_model.predict(X_train)
y_pred_rf_test = rf_model.predict(X_test)
```

Figure 3. Application of the Random Forest Algorithm

Table 7. Random Forest Results

Evaluation Metrics	Training Set	Testing Set
R^2 Score	0.5442 (54.42%)	0.2115 (21.15%)
RMSE	31.384.04 units	32.754.23 units
MAE	10,063.68 units	12,073.59 units
MSE	984,958,008.53	1,072,839,652.26

The Random Forest results obtained an R^2 value of 0.2115 on the test data, which shows a significant improvement in performance compared to the Linear Regression model. This value indicates that the model is able to explain 21.15% of the variation in sales, thus capturing non-linear patterns and complex interactions between features that cannot be modeled effectively by a linear approach. In addition, the RMSE value of 32,754 units and the MAE value of 12,074 units indicate an average prediction error rate that is still considered normal for e-commerce data characteristics that have a high inherent variance.

The difference between the training R^2 (0.5442) and the testing R^2 (0.2115) with a difference of 0.3327 indicates mild overfitting, which is a common phenomenon in ensemble methods such as Random Forest. Despite this difference, the model's performance on the test data still shows fairly good generalization capabilities. These results illustrate that the model is able to maintain prediction stability when applied to data not included in the training process.

Comparison and Evaluation of Results

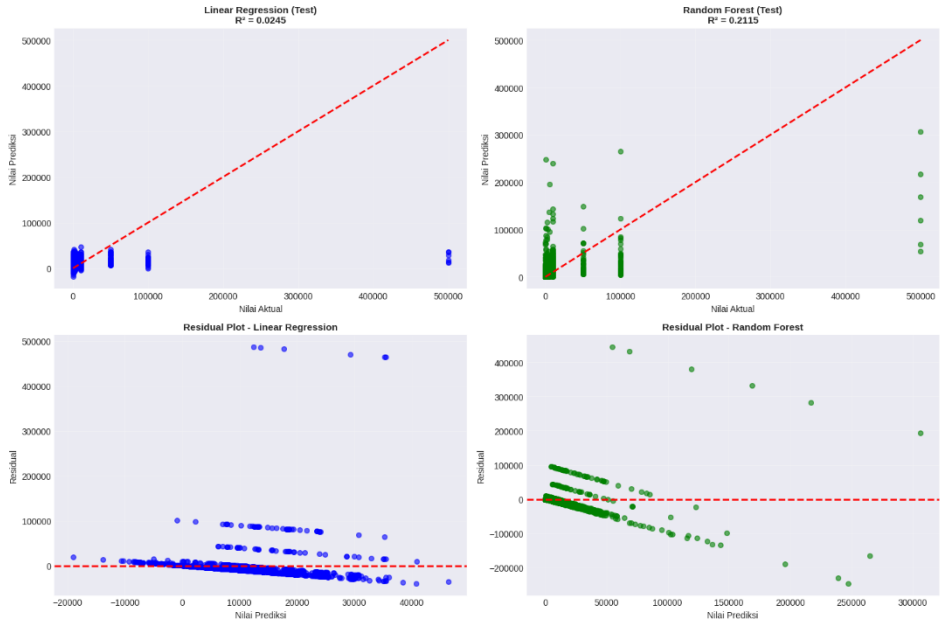

Figure 4. Scatter Plot

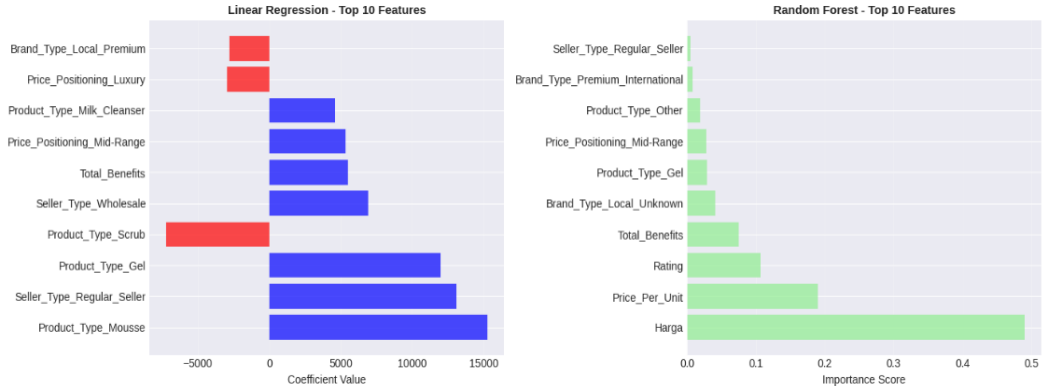

Figure 5. Visualization of Top 10 Features in Linear Regression and Random Forest

The visualizations in both figures provide an overview of the relationship between prediction quality and feature contribution in the Linear Regression and Random Forest models. In Figure 3.4, the scatter plot shows that Linear Regression produces a distribution of points far from the reference line, consistent with a very low R^2 value. This indicates that the linear relationship is unable to capture the variable characteristics of sales. In contrast, Random Forest shows prediction points closer to the diagonal, reflecting the model's ability to recognize non-linear patterns even though deviations are still visible in products with high sales.

The residual plot further reinforces these findings. Both models show widely dispersed and asymmetrical residuals, indicating the presence of external factors not represented in the features, as well as high natural variation in e-commerce sales data. Figure 3.5 then shows that Linear Regression is more influenced by categorical features through coefficient values, while Random Forest places numerical features such as Price, Price_Per_Unit, Rating, and Total_Benefits as the main determinants. This difference explains why Random Forest performs better: tree-based models are able to capture interactions between variables more flexibly than linear structures.

Table 8. Comparison of Results

Metrics	LR Train	LR Test	RF Train	RF Test
R^2 Score	0.0357	0.0245	0.5442	0.2115
RMSE	45.650	36.432	31.384	32.754
MAE	16,543	14,689	10,064	12,074
MSE	2.08B	1.33B	985 million	1.07B

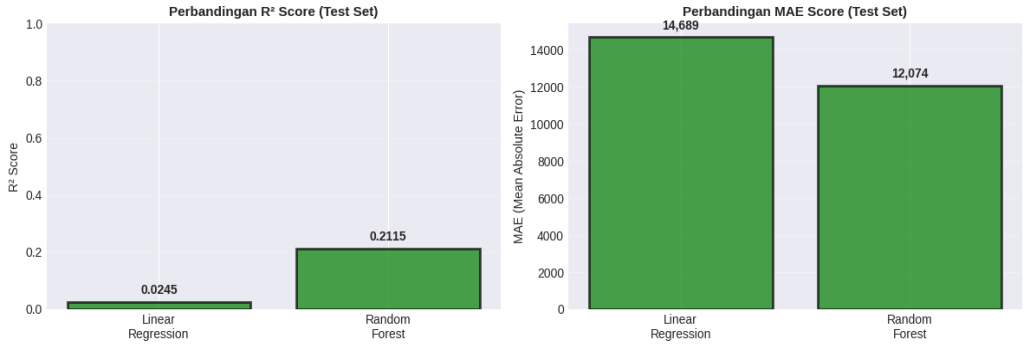

Figure 6. Comparison of Results

Testing set results show a clear difference in performance between the two algorithms. The Random Forest model achieved an R^2 value of 0.2115, much higher than Linear Regression with an R^2 of 0.0245, which is equivalent to an increase of around 765%. This improvement is also reflected in the error metrics, with an RMSE improvement of 10.1% and an MAE improvement of 17.8%.

The advantages of Random Forest can be traced back to its characteristics, such as its ability to model non-linear relationships, its resistance to outliers through the ensemble process, automatic feature selection through decision tree node splitting, and its flexibility in capturing pattern variations in heterogeneous market segments. Conversely, although Linear Regression has good interpretability, its predictive performance is less than optimal when used on e-commerce sales data, which has a complex structure and is influenced by many non-linear factors.

CONCLUSION

This study concludes that Random Forest is the most effective model for predicting skincare product sales on Tokopedia. Using a dataset containing 7,217 products and 21 features, Random Forest achieved an R^2 value of 0.2115 on the test data, meaning that this model can explain more than one-fifth of the sales variation. This result is much better than Linear Regression, which only produced an R^2 of 0.0245, indicating that the relationship between factors in skincare sales is non-linear and more appropriately modeled using ensemble-based algorithms.

Analysis of the important features in Random Forest reveals that price and value for money play a dominant role in influencing sales, followed by review volume and ratings as indicators of consumer trust. Product characteristics, such as form and number of benefits offered, also contribute significantly, while brand, seller type, and market positioning factors complete the determinant structure that shapes sales performance. These findings confirm that purchasing behavior in e-commerce is influenced by a combination of functional aspects, consumer perceptions, and marketing strategies.

The research methodology, which retained all data through zero imputation and feature engineering by developing 21 derived attributes, proved to support improved model performance. This approach enabled the model to learn from products without sales or reviews, thereby improving its generalization ability for new products. Overall, this study shows that the use of algorithms capable of capturing non-linear interactions, coupled with

the selection of features relevant to the market context, is an important step in building reliable sales prediction models in the e-commerce environment.

REFERENCES

Ababil, O. J., Wibowo, S. A., & Zahro, H. Z. (2022). Application of linear regression method in predicting liquid vape sales at the Pandaan Vapor Store based on a website.

Adi, R. M. S., & Sudianto, S. (2022). Prediction of food commodity prices using the long short-term memory (LSTM) algorithm. Building of Informatics, Technology and Science (BITS), 4(2). https://doi.org/10.47065/bits.v4i2.2229

Ashari, M. L., & Sadiki, M. (n.d.). Prediction of time series sales transaction data using LSTM regression.

Dewi, S. P., Nurwati, N., & Rahayu, E. (2022). Application of data mining for predicting sales of best-selling products using the k-nearest neighbor method. Building of Informatics, Technology and Science (BITS), 3(4), 639–648. https://doi.org/10.47065/bits.v3i4.1408

Fitri, E. (2023). Comparative analysis of linear regression, random forest regression, and gradient boosted trees regression methods for house price prediction. Journal of Applied Computer Science and Technology (JACOST), 4(1), 2723–1453. https://doi.org/10.52158/jacost.491

Hamdanah, F. H., & Fitrianah, D. (2021). Analysis of the performance of linear regression algorithms with generalized linear models for sales prediction in micro, small, and medium enterprises. Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI), 10(1), 23. https://doi.org/10.23887/janapati.v10i1.31035

Hidayat, R., et al. (2025). Implementation of the random forest regression algorithm to predict production sales in supermarkets. SIMKOM, 10(1), 101–109. https://doi.org/10.51717/simkom.v10i1.703

Krisnawati, W. (2022). The effect of the special event “one memorable day with Oh Sehun” on the brand image of Whitelab skincare (Survey of visitors to the special event “one memorable day with Oh Sehun”) (Thesis, Faculty of Communication Sciences, Muhammadiyah University Jakarta).

Kusumawardani, N., Afandi, M. R., & Riani, L. P. (2023). Demand forecasting analysis using the linear exponential smoothing method (Study on Fendy Batik products, Klaten).

Nugraheny, D., Indrianingsih, Y., Kurniawan, S., & Sunaryo, H. (2023). Prediction of minimum revenue targets for Dasimoen florist flower shop using web-based multiple linear regression methods. Angkasa: Jurnal Ilmiah Bidang Teknologi, 15(1), 76. https://doi.org/10.28989/angkasa.v15i1.1592

Romadhon, D. P., & Putra, R. E. (2024). Application of deep learning methods using CNN-based recommendation algorithms in GOLS e-commerce applications (Case study: PT. Cipta Giri Sentosa). Journal of Informatics and Computer Science, 5.

Rusdy, A. A. (2022). Application of linear regression methods in predicting drug supply and demand: Case study of point of sales application. Bulletin of Islamic Information and Technology Systems, 3(2), 121–126.

Sianturi, C. J. M., Sinaga, M. D., Sembiring, N. S. B., Ginting, E., & Potensi Utama, U. K. (2024). Multiple linear regression method in web-based product sales prediction. Journal of Informatics, Management and Computers, 16(1).

Tombeng, M. T., & Ardian, Z. (2021). Supermarket sales prediction using a deep learning approach. Cogito Smart Journal, 7(1).