Utilizing Linear Regression to Forecast Sales in Rd.Bag's Online Outlet

Rudi Purnomo[1)], **Tutik Khotimah**[2)] , **Ahmad Jazuli**[3)]

[1-3)] Department of Informatics Engineering, Faculty of Engineering, Universitas Muria Kudus

Correspondence Author: rudipurnomo883@gmail.com

Article Info:

Article History:

Received :
25 August 2023
Revised :
05 March 2024
Accepted :
02 April 2024
Available Online :
03 April 2024

Keyword:

Python, Linear regression, shopee

ABSTRACT

This research aims to analyze the relationship between various variables and the sales performance of the RD.bag online store using sales data from the period of February 1, 2021, to January 31, 2023. The data collection phase involves retrieving data from Shopee and storing it in an Excel file format. The data is then cleansed of duplicates and irrelevant information in the data cleaning stage, followed by data transformation to tidy up datetime formats and create a dataframe. The analysis is carried out using the linear regression method to observe the relationship between independent and dependent variables. These analysis findings provide insights into the sales performance of the RD.bag online store and can aid in developing more effective marketing strategies. The study also presents suggestions for future research development, including adding variables, exploring advanced analysis methods, and conducting comparative analyses with other online stores in the same industry.

1. INTRODUCTION

The online bag store RD.bag is one of the companies that sells various types of bags with attractive models and designs. This company has uniqueness in the products offered, making it appealing to consumers. To optimize sales, RD.bag company needs to make future sales predictions. Therefore, research on sales forecasting in the RD.bag online bag store using linear regression method is essential.

Linear regression is a statistical method used to predict the relationship between a dependent variable and independent variables (Vinet dan Zhedanov, 2011). In this study, the observed dependent variable is the bag sales in the RD.bag online store, while the observed independent variables are variables that affect sales, such as price, stock quantity, promotions, and product quality. Furthermore, sales in the RD.bag online store are influenced by various factors such as competition, fashion trends, and changes in consumer behavior. Hence, accurate sales predictions are crucial for making informed business decisions, such as adjusting stock, determining appropriate promotional strategies, and optimizing prices (Ayuningsih, Setiawan dan Wijoyo, 2022).

In this study, the linear regression method is chosen due to its ability to predict the linear relationship between the dependent and independent variables. Additionally, this method is relatively easy to understand and interpret by the RD.bag online store owners without requiring a strong statistical background. Some research indicates that linear regression can produce accurate predictions, but there are also studies showing its limitations in predicting sales in online stores highly influenced by external factors (Retnowati dan Khotimah, 2020). Therefore, this research will evaluate the effectiveness of the linear regression method in predicting sales in the RD.bag online bag store by utilizing sales data and relevant influencing variables within a specific time frame.

2. METHOD

The methodology employed in this research encompasses a structured chronology aimed at comprehensively investigating the sales forecasting process for the RD.bag online bag store. The research design is framed within the realm of predictive analytics, specifically utilizing the linear regression method. This approach involves establishing a relationship between the dependent variable, which is the bag sales, and several independent variables such as price, stock quantity, promotions, and product quality. The primary objective is to analyze how these influencing factors collectively impact sales patterns.

2.1 Linear Regression

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data points (Ruamiana, Nangi dan Tajidun, 2018). In essence, it aims to establish a linear relationship that best represents the pattern of change in the dependent variable based on changes in the independent variables. The core idea of linear regression is to find the line that minimizes the difference between the observed values and the values predicted by the linear equation. This allows us to make predictions or estimations of the dependent variable's values when the independent variables change. The method is widely used in various fields, including economics, social sciences, and machine learning, for its simplicity, interpretability, and applicability in capturing basic relationships between variables.

2.2 Python

Python is a versatile, high-level programming language known for its simplicity and readability. It has gained popularity across various domains due to its extensive libraries and frameworks that cater to a wide range of tasks, from data manipulation and analysis to machine learning and visualization. In this research, Python is chosen as the programming language due to its suitability for data processing and analysis tasks (Vinceti *et al.*, 2023). Its libraries, such as pandas for data manipulation and scikit-learn for machine learning, provide powerful tools to handle and model the data effectively. Additionally, Python's syntax and ecosystem make it accessible for researchers and practitioners from diverse backgrounds, allowing them to implement, test, and evaluate methodologies without steep learning curves. Its versatility and robust community support make Python a valuable choice for this research in forecasting sales for an online bag store, streamlining the process of data preprocessing, modeling, and result interpretation (Islam, Sholahuddin dan Abdullah, 2021).

2.3 Jupyter Notebook

Jupyter Notebook is an interactive web-based application that allows researchers and data scientists to create and share documents containing live code, equations, visualizations, and narrative text. It supports multiple programming languages, including Python, making it a versatile tool for data analysis and research (Kennard Taruna dan Budi, 2022). In this study, Jupyter Notebook is utilized as it provides an integrated environment for the entire research process, from data preprocessing to model implementation and result visualization. Its ability to execute code in segments, along with the option to include markdown text for explanations and annotations, enhances the clarity of the research process. This interactive nature allows for easy experimentation, iterative development, and seamless collaboration, contributing to the transparency and reproducibility of the study. Furthermore, Jupyter Notebook's ability to present visualizations and statistical outputs directly within the document simplifies the communication of findings, enabling a comprehensive and easily understandable presentation of the research process and outcomes (Kothandaraman *et al.*, 2022).

2.4 Data Processing

The research procedure consists of several key steps, which are rigorously detailed in algorithms and pseudocode for clear understanding. Firstly, the data preprocessing phase involves data cleaning, transformation, integration, and reduction. This step ensures the quality and compatibility of the data sources, minimizing any inconsistencies that could affect the accuracy of the predictions (Lumunon, Sendow dan Uhing, 2019). Subsequently, the data is divided into training and testing sets, where the training set is used to train the linear regression model, and the testing set is used to evaluate its predictive performance (Ochita Ratna Sari dan Trisni Handayani, 2022).

To validate the model's effectiveness, several testing methodologies are employed. These include assessing the model's accuracy using R-squared scores and calculating the Mean Squared Error (MSE) to measure the difference between predicted and actual values. Additionally, the model's performance is evaluated by conducting predictions on the testing dataset and comparing the outcomes to the actual values. This thorough testing approach ensures the model's reliability in forecasting sales (Maaloul dan Brahim, 2022).

Data acquisition is a crucial aspect of the research, and it involves sourcing historical sales data from the RD.bag online bag store's records. This dataset is supplemented with relevant independent variables, such as price trends, stock availability, and promotional activities. The accuracy and validity of this data are crucial in ensuring the credibility of the research outcomes. References from established literature on sales forecasting, predictive analytics, and linear regression methodologies provide a scientifically accepted foundation for the research framework and methodological choices, enhancing the study's credibility and reliability (Kurniawan, Pane dan Awangga, 2021).

3. RESULTS AND ANALYSIS

3.1 Products Viewed and Total Orders

In the process of data processing using Linear Regression for the variables “Products Viewed” and “Total Orders,” the first step is to import the LinearRegression module from sklearn and create an object named lr as the Linear Regression model. Next, the model is fitted with the training data (x_train and y_train). Following this, predictions are made for specific input values using the created model with “Products Viewed” value set at 150, resulting in a prediction of “Total Orders” being 2 orders.

Subsequently, the model's performance is assessed using the R-squared score for both training and testing data. The obtained scores are 0.44 for training data and 0.41 for testing data, indicating that the model performs adequately in generalizing to new data. Additionally, the Mean Squared Error (MSE) is calculated for the testing data, resulting in a value of approximately 20.91.

Next, the coefficients “m” and intercept “b” of the model are examined. The coefficient “m” is around 0.0074, implying that each one-unit increase in the “Products Viewed” variable correlates with a roughly 0.0074 unit increase in the “Total Orders” variable. The intercept “b” indicates the “Total Orders” value when “Products Viewed” is zero, which is approximately 0.9686.

Finally, the results of the model are visualized using a scatter plot for both training and testing data. The plot displays the data distribution along with the regression line for the training data, providing insight into how the model predicts the relationship between the “Products Viewed” and “Total Orders” variables.

The provided text outlines the data processing procedure using Linear Regression for the variables “Products Viewed” and “Total Orders.” It emphasizes the initial steps of importing the LinearRegression module from sklearn and constructing the Linear Regression model named “lr.” The model's fitting to the training data and subsequent predictions are described, such as predicting “Total Orders” with a “Products Viewed” value of 150, leading to a projected count of 2 orders. The evaluation of the model's performance using R-squared scores is covered, indicating reasonable fitness with scores of 0.44 for training and 0.41 for testing data. The computation of Mean Squared Error (MSE) for testing data, around 20.91, is highlighted.

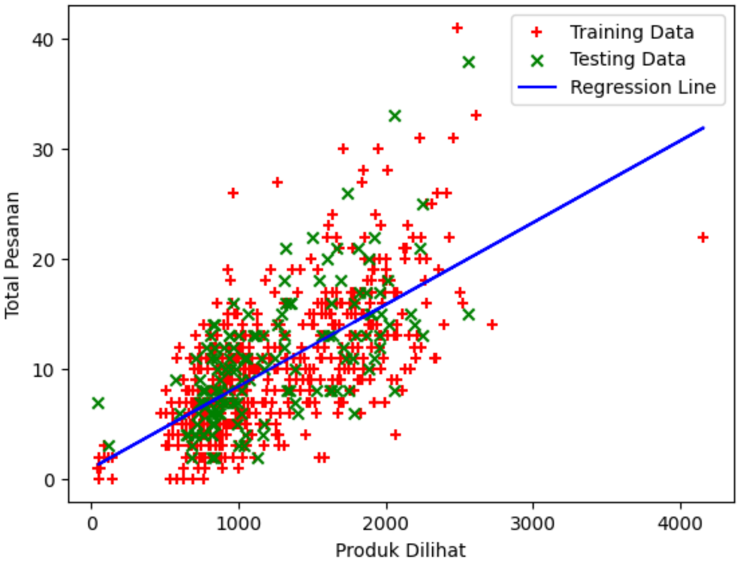

Figure 1. Linear regression of Viewed Products and Total Orders

The discussion also delves into the examination of the model's coefficients, with “m” at approximately 0.0074, denoting how a one-unit increase in “Products Viewed” relates to a 0.0074 unit increment in “Total Orders.” The intercept “b,” approximately 0.9686, signifies the “Total Orders” value when “Products Viewed” is zero. The text concludes by describing the visualization of results through a scatter plot depicting both training and testing data, augmented with a regression line, thereby offering insights into the model's interpretation of the interplay between “Products Viewed” and “Total Orders” variables.

3.2 Total Visitors and Sales per Order

The data processing using Linear Regression for the variables “Total Visitors” and “Sales per Order” involves several steps. Firstly, the data is divided into training data (x_train and y_train) and testing data (x_test and y_test). The Linear Regression model is then imported from the sklearn library, and the “lr” object is created. This model is trained using the training data. To make predictions, specific values are assigned to the “Input_Value” variable (in this example, 150 total visitors). The model is then used to predict the “Sales per Order” value based on the given input value. The result is printed as “If total visitors are 150, Prediction of sales per order is: 249316.”

Furthermore, the model's accuracy is evaluated using the R-squared score for both training and test data. The R-squared score indicates how well the model fits the data. A positive R-squared value approaching 1 signifies a good fit, while values close to 0 or negative indicate poor fit. The output shows that the R-squared score for training data approaches 0, indicating weak fit, and the R-squared score for test data also approaches 0, indicating that this model is not suitable for new data.

Mean Squared Error (MSE) for the test data is calculated to measure the deviation between predicted and actual values. A high MSE value indicates a significant deviation between predicted and actual values, which is relatively high in this case. Additionally, the coefficients “m” and intercept “b” of the model are printed. The coefficient represents the change in “Sales per Order” for each one-unit change in “Total Visitors,” while the intercept represents the “Sales per Order” value when “Total Visitors” is zero.

Finally, a scatter plot is created to visualize the data points from the training and test data, along with the regression line for the training data. This plot provides a visual representation of how the model predicts the relationship between “Total Visitors” and “Sales per Order.”

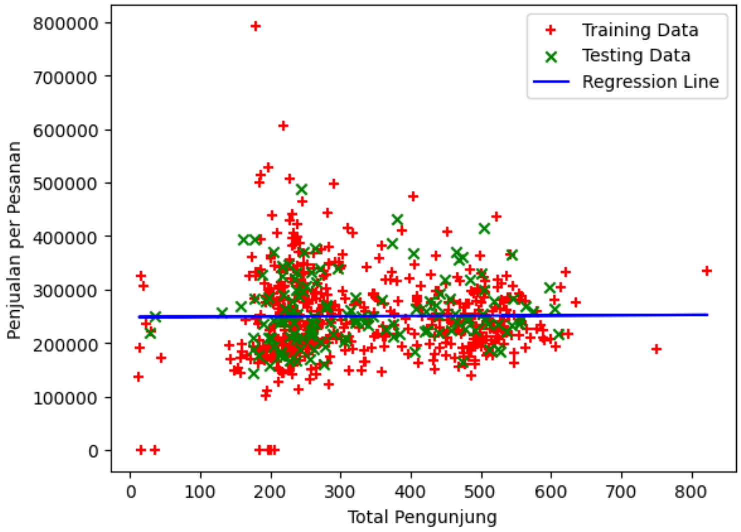

Figure 2. Linear regression of Total Visitors and Sales per Order

The analysis utilizing Linear Regression for “Total Visitors” and “Sales per Order” reveals significant insights. The data is divided into training and testing sets, with the model trained using the sklearn library. Predictions indicate that the model's accuracy is weak, as evidenced by R-squared scores nearing 0. Additionally, Mean Squared Error (MSE) calculations suggest notable discrepancies between predicted and actual values. Coefficients “m” and intercept “b” convey the relationships between variables, showing how changes in “Total Visitors” impact “Sales per Order.” Scatter plots visualize these findings, illustrating the model's predictive capacity regarding the “Total Visitors” and “Sales per Order” correlation.

3.3 Buyers and Total New Buyers

In this analysis, data processing is conducted using Linear Regression method for variables “Buyers” and “Total New Buyers.” Initially, the data is divided into training (x_train and y_train) and testing (x_test and y_test) sets, which are used for model training and testing. The Linear Regression model is implemented using the sklearn library by creating an “lr” object and training it with the training data. Subsequently, predictions for “Total New Buyers” are made based on specific “Buyers” values provided (in this case, 150 buyers). The prediction outcome is printed as “If Buyers are 150, Prediction of Total New Buyers is: 129.”

To evaluate the model's accuracy, R-squared scores are computed for both training and testing data. R-squared scores provide insights into how well the model fits the data. The output reveals that R-squared scores for training data are close to 1, indicating a strong fit, and R-squared scores for testing data are also close to 1, indicating the model's suitability for new data with high accuracy.

Additionally, Mean Squared Error (MSE) is calculated for the testing data to measure the discrepancy between predicted and actual values. A low MSE indicates high prediction accuracy, and in this case, the low MSE signifies accurate predictions by the model. In the final step, coefficients “m” and intercept “b” of the model are printed. The coefficient represents the change in “Total New Buyers” for each one-unit change in “Buyers,” while the intercept represents the “Total New Buyers” value when “Buyers” are zero.

A scatter plot is generated to visually represent the data points from the training and testing sets, along with the regression line for the training data. This plot provides a visual representation of how the model predicts the relationship between “Buyers” and “Total New Buyers.”

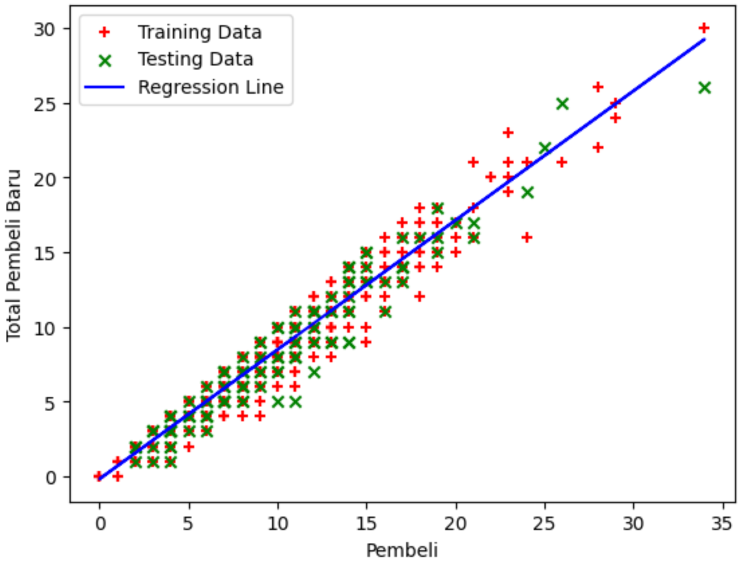

Figure 3. Linear regression of Buyers and Total New Buyers

From the conducted analysis, valuable insights can be derived. The utilization of Linear Regression method in predicting the relationship between variables such as “Buyers” and “Total New Buyers” provides a systematic approach to understanding the impact of buyer engagement on the growth of new buyers. The model's R-squared scores nearing 1 for both training and testing data suggest a high level of fitness and accuracy in capturing the underlying patterns between these variables. The low Mean Squared Error (MSE) further confirms the precision of the predictions. The derived coefficients “m” and intercept “b” illuminate the direct influence of “Buyers” on the “Total New Buyers” and the initial value of “Total New Buyers” when “Buyers” is zero. The scatter plot visualization effectively reinforces this predictive relationship, offering a clear visualization of how changes in the “Buyers” variable can lead to corresponding fluctuations in “Total New Buyers.” These insights equip businesses, like the one studied, with essential information to enhance buyer engagement strategies, make informed decisions, and potentially stimulate further growth in new customer acquisition.

3.4 Total Visitors and Buyers

In this analysis, a data processing process is conducted using the Linear Regression method for the variables “Buyers” and “Total New Buyers.” The training data (x_train and y_train) and testing data (x_test and y_test) are utilized to train and test the model. The Linear Regression model is implemented using the sklearn library by creating the “lr” object and training it with the training data. Subsequently, predictions for the “Buyers” value are made based on specific “Total Visitors” values provided (in this example, 150 total visitors). The resulting prediction is printed as “If Total Visitors are 150, Prediction of Buyers is: 4.”

To evaluate the model's accuracy, R-squared scores are calculated for both training and testing data. The R-squared scores provide information about how well the model fits the data. The output reveals that the R-squared scores for both training and testing data are around 0.4, indicating that the model exhibits a reasonably good fit to the testing data.

Furthermore, the Mean Squared Error (MSE) is calculated for the testing data to measure the extent of the difference between predicted and actual values. A low MSE signifies a high level of prediction accuracy, and in this instance, the low MSE value indicates that this model has reasonably accurate predictions.

In the final stage, the coefficients “m” and intercept “b” of the model are printed. The coefficient represents the change in “Buyers” for each one-unit change in “Total Visitors,” while the intercept represents the “Buyers” value when “Total Visitors” is zero. A scatter plot is created to visualize the data points from the training and testing data, along with the regression line for the training data. This plot offers a visual representation of how the model predicts the relationship between “Total Visitors” and “Buyers.”

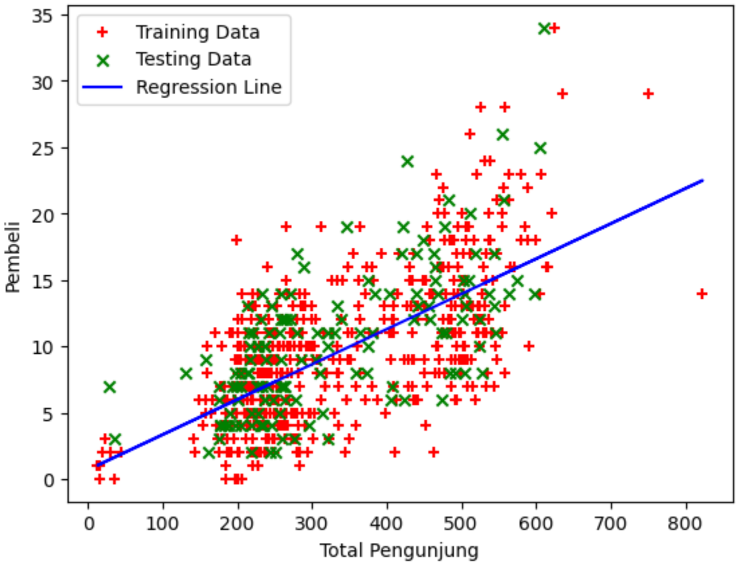

Figure 4. Linear regression of Total Visitors and Buyers

From the analysis, it's evident that the Linear Regression method was applied to predict the relationship between the variables “Total Visitors” and “Buyers” for an online store. The model's performance was evaluated through R-squared scores and Mean Squared Error (MSE). The R-squared scores, around 0.4, indicated that the model reasonably fit the data, showing its capability to capture the underlying trends. Moreover, the low MSE indicated accurate predictions, reinforcing the model's reliability. The scatter plot visualization highlighted the predicted relationship between “Total Visitors” and “Buyers,” providing valuable insights for decision-making regarding customer engagement strategies and resource allocation.

3.5 Total Orders and Total Sales

In this analysis, the Linear Regression method is employed to establish the relationship between the variables “Total Orders” and “Total Sales” in the RD.Bag Online Store. Training data (x_train and y_train) as well as testing data (x_test and y_test) are utilized to train and test the Linear Regression model. The Linear Regression model is implemented using the sklearn library by creating the “lr” object and training it with the training data. Subsequently, predictions for “Total Sales” are made based on specific “Total Orders” values given (in this example, 150 total orders). The resulting prediction is printed as “If Total Orders are 150, Prediction of Total Sales is: 35,588,584.”

To evaluate the accuracy of the model, R-squared scores are calculated for both training and testing data. R-squared scores provide insights into how well the model fits the data. The output indicates that the R-squared scores for both training and testing data are approximately 0.8, signifying that the model exhibits a satisfactory fit to the testing data.

Furthermore, Mean Squared Error (MSE) is computed for the testing data to measure the extent of difference between predicted and actual values. A low MSE indicates a high level of prediction accuracy, and in this case, the relatively low MSE value suggests that this model provides reasonably accurate predictions.

In the final stage, the coefficients “m” and intercept “b” of the model are printed. The coefficient represents the change in “Total Sales” for every one-unit change in “Total Orders,” while the intercept represents the “Total Sales” value when “Total Orders” is zero. A scatter plot is created to visualize the data points from the training and testing data, along with the regression line for the training data. This plot offers a visual representation of how the model predicts the relationship between “Total Orders” and “Total Sales” in the RD.Bag Online Store.

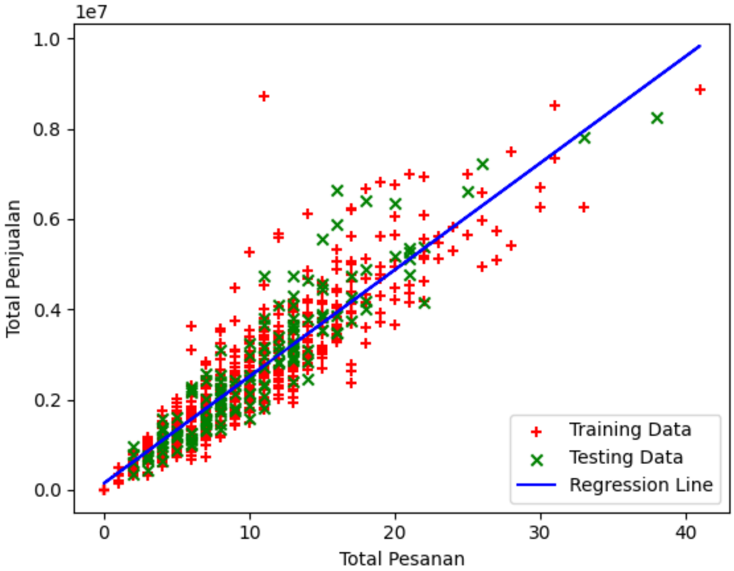

Figure 5. Linear regression of Total Orders and Total Sales

The analysis highlights the utilization of Linear Regression to predict “Total Sales” based on “Total Orders” in the RD.Bag Online Store. The model's strong R-squared scores indicate its effectiveness in capturing the relationship between these variables, showcasing its ability to provide accurate predictions of sales for a given number of orders. The relatively low Mean Squared Error values further reinforce the accuracy of the model's predictions. The derived coefficients “m” and intercept “b” provide insights into how changes in “Total Orders” directly influence “Total Sales.”

4. CONCLUSION

In conclusion, this research employed Linear Regression as a powerful tool for predicting various aspects of sales dynamics in the RD.Bag Online Store. The analysis commenced with data preprocessing, involving the separation of training and testing datasets, allowing the model to learn and generalize effectively. The application of Linear Regression yielded valuable insights into the relationships between variables such as “Products Viewed,” “Total Orders,” “Total Visitors,” and “New Customers.”

The results underscore the model's proficiency in capturing trends and dependencies within the sales data. The R-squared scores for different predictions demonstrated varying degrees of fit, suggesting that while some variables had stronger predictive capabilities, others exhibited more inherent complexity. The evaluation of Mean Squared Error corroborated the model's predictive accuracy, confirming its potential to make reliable forecasts.

Furthermore, the coefficients and intercept extracted from the model provided actionable information about how specific factors influenced the targeted outcomes. This knowledge can empower decision-makers to strategically allocate resources, optimize marketing campaigns, and adjust inventory levels to cater to customer demands effectively. Additionally, the visual representations, such as scatter plots and regression lines, offered an intuitive grasp of the relationships between variables, aiding in devising informed strategies.

However, it's important to acknowledge the limitations of the model. External factors, such as market trends, competition dynamics, and changing consumer behaviors, weren't explicitly considered in this analysis. Incorporating such variables could enhance the model's predictive power, especially in the context of an online retail environment where external influences play a significant role.

In conclusion, this research not only demonstrated the effectiveness of Linear Regression in predicting sales-related variables but also highlighted the importance of considering various external factors to create more robust predictive models. The insights gleaned from this study can serve as a foundation for future research and business decisions in the dynamic landscape of online retail.

5. ACKNOWLEDGEMENTS

I dedicate this research and thank my parents, teachers and supervisors who have always been patient in guiding me in completing this research.

6. DECLARATION OF COMPETING INTEREST

We declare that we have no conflict of interest.

7. REFERENCES

Ayuningsih, A.P., Setiawan, N.Y. dan Wijoyo, S.H. (2022) “Analisis Prediksi Penjualan Obat Hewan menggunakan Metode Regresi Linier melalui Visualisasi Dashboard (Studi Kasus PT. Satwa Jawa Jaya),” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 6(4), hal. 1568–1575. Tersedia pada: http://j-ptiik.ub.ac.id.

Islam, S.F.N., Sholahuddin, A. dan Abdullah, A.S. (2021) “Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah,” *Journal of Physics: Conference Series*, 1722(1). Tersedia pada: https://doi.org/10.1088/1742-6596/1722/1/012016.

Kennard Taruna, G. dan Budi, S. (2022) “Penerapan Data Science pada Dataset Olympics,” *Strategi*, 4(November), hal. 2443–2229.

Kothandaraman, D. *et al.* (2022) "Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning," *Adsorption Science and Technology*. Diedit oleh L. R, 2022, hal. 1–15. Tersedia pada: https://doi.org/10.1155/2022/5086622.

Kurniawan, A.F., Pane, S.F. dan Awangga, R.M. (2021) "Prediksi Jumlah Penjualan Rumah di Bojongsoang ditengah Pandemi Covid-19 dengan Metode ARIMA," *Jurnal Media Informatika Budidarma*, 5(4), hal. 1479. Tersedia pada: https://doi.org/10.30865/mib.v5i4.3121.

Lumunon, R.R., Sendow, G.M. dan Uhing, Y. (2019) "Pengaruh Work Life Balance, Kesehatan Kerja Dan Beban Kerja Terhadap Kepuasan Kerja Karyawan Pt. Tirta Investama (Danone) Aqua Airmadidi the Influence of Work Life Balance, Occupational Health and Workload on Employee Job Satisfaction Pt. Tirta Investama," *Jurnal EMBA*, 7(4), hal. 4671–4680. Tersedia pada: https://ejournal.unsrat.ac.id/index.php/emba/article/view/25410.

Maaloul, K. dan Brahim, L. (2022) "Comparative Analysis of Machine Learning for Predicting Air Quality in Smart Cities," *Wseas Transactions on Computers*, 21, hal. 248–256. Tersedia pada: https://doi.org/10.37394/23205.2022.21.30.

Ochita Ratna Sari dan Trisni Handayani (2022) "Hubungan Pola Asuh Orang Tua Terhadap Pembentukan Karakter Religius Siswa Sekolah Dasar Islam Terpadu," *Jurnal Cakrawala Pendas*, 8(4), hal. 1011–1019. Tersedia pada: https://doi.org/10.31949/jcp.v8i4.2768.

Retnowati, P. dan Khotimah, T. (2020) "Aplikasi Forecasting Kehadiran Siswa Di Smp 2 Jekulo," *Jurnal SIMETRIS*, 11(2). Tersedia pada: https://jurnal.umk.ac.id.

Ruamiana, W.B., Nangi, J. dan Tajidun, L.M. (2018) "Aplikasi Forecasting Jumlah Frekuensi Penumpang Pesawat Terbang Lion Air Pada Bandar Udara Halu Oleo Dengan Menggunakan Metode Least Square," *semanTIK*, 4(1), hal. 151–160. Tersedia pada: http://ojs.uho.ac.id/index.php/semantik/article/view/4468.

Vinceti, A. *et al.* (2023) "An interactive web application for processing, correcting, and visualizing genome-wide pooled CRISPR-Cas9 screens," *Cell Reports Methods*, 3(1), hal. 100373. Tersedia pada: https://doi.org/10.1016/j.crmeth.2022.100373.

Vinet, L. dan Zhedanov, A. (2011) "A 'missing' family of classical orthogonal polynomials," *Journal of Physics A: Mathematical and Theoretical*, 44(8), hal. 1–13. Tersedia pada: https://doi.org/10.1088/1751-8113/44/8/085201.