P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v8i3.537 Vol. 8, No. 3. Juni 2026
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

PRODUCT SALES PREDICTION USING XGBOOST WITH FEATURE IMPORTANCE ANALYSIS FOR ADVERTISING MEDIA EVALUATION

Wilsen Grivin Mokodaser[-1], Tonny Irianto Soewignyo[-2], Fanny Soewignyo[-3]

[1]Informatika / Fakultas Ilmu Komputer / Universitas Klabat
wilsenm@unklab.ac.id[-1]
[2,3]Akuntansi / Fakultas Ekonomi dan Bisnis / Universitas Klabat
tonnysoewignyo@unklab.ac.id[-2], f.soewignyo@unklab.ac.id[-3]

Abstract

Product sales prediction plays a crucial role in supporting data-driven marketing strategies and optimizing advertising expenditures. Although previous studies have demonstrated the effectiveness of machine learning techniques for sales forecasting, most of them primarily focus on prediction accuracy and provide limited insights into the contribution of individual advertising channels to sales performance. This limitation reduces the interpretability and practical value of predictive models for business decision-making. Therefore, this study proposes a product sales prediction framework using Linear Regression as a baseline model and XGBoost Regression combined with Feature Importance Analysis for advertising media evaluation. The novelty of this study lies in integrating predictive modeling and interpretable analysis within a single framework, enabling both accurate sales prediction and the identification of influential advertising factors. Hyperparameter optimization and five-fold cross validation were employed to improve model reliability and robustness. Experimental results show that Linear Regression outperformed XGBoost, achieving an R^2 score close to 1.0, while XGBoost achieved an R^2 score of 0.953 with a mean cross-validation R^2 score of 0.950, indicating stable predictive performance. Feature Importance Analysis revealed that Affiliate Marketing was the most influential factor, followed by Billboards and Social Media. These findings contribute to marketing analytics by providing interpretable insights that support advertising budget optimization and more effective data-driven business decision-making.

Keywords: Marketing Analytics; XGBoost; Linear Regression; Feature Importance;

Abstrak

Abstrak- Prediksi penjualan produk memiliki peran penting dalam mendukung strategi pemasaran berbasis data dan optimalisasi pengeluaran iklan. Meskipun berbagai penelitian sebelumnya telah menunjukkan efektivitas teknik machine learning dalam peramalan penjualan, sebagian besar penelitian tersebut lebih berfokus pada akurasi prediksi dan masih memberikan keterbatasan dalam menjelaskan kontribusi masing-masing media periklanan terhadap hasil penjualan. Keterbatasan ini mengurangi nilai interpretabilitas dan manfaat praktis model prediksi dalam mendukung pengambilan keputusan bisnis. Oleh karena itu, penelitian ini mengusulkan kerangka prediksi penjualan produk menggunakan Linear Regression sebagai model dasar dan XGBoost Regression yang dikombinasikan dengan Feature Importance Analysis untuk mengevaluasi efektivitas media periklanan. Kebaruan penelitian ini terletak pada integrasi antara pemodelan prediktif dan analisis interpretabilitas dalam satu kerangka kerja sehingga mampu menghasilkan prediksi yang akurat sekaligus mengidentifikasi faktor-faktor yang paling berpengaruh terhadap penjualan. Optimasi hyperparameter dan validasi silang lima lipatan diterapkan untuk meningkatkan keandalan model. Hasil penelitian menunjukkan bahwa Linear Regression menghasilkan performa yang lebih baik dengan nilai R^2 mendekati 1, sedangkan XGBoost memperoleh nilai R^2 sebesar 0,953 dengan rata-rata nilai R^2 validasi silang sebesar 0,950 yang menunjukkan performa yang stabil. Analisis feature importance menunjukkan bahwa Affiliate Marketing, Billboards, dan Social Media merupakan faktor yang paling berpengaruh terhadap penjualan produk. Temuan ini memberikan kontribusi dalam bidang analitik pemasaran melalui penyediaan wawasan yang lebih interpretatif untuk mendukung optimalisasi anggaran iklan dan pengambilan keputusan bisnis berbasis data.

Kata kunci: Analitik Pemasaran; XGBoost; Linear Regression; Feature Importance;

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.537
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

INTRODUCTION

The development of digital technology has transformed modern marketing strategies into increasingly data-driven approaches. Various companies now utilize multiple advertising media (Endrawati Subroto et al., 2024), such as television, social media, Google Ads, billboards, affiliate marketing, and influencer marketing (Helen & Rusdi, 2023), to increase product sales (Adrian Hidayat, 2024). However, the growing investment in diverse marketing channels has created new challenges for companies, particularly in determining which promotional media are most effective in influencing product sales (Wardani, 2023). In the field of marketing analytics, understanding the effectiveness of advertising expenditures is essential for optimizing resource allocation and improving business performance (Wedel & Kannan, 2016). Errors in advertising budget allocation may result in inefficient spending and reduced business competitiveness. Therefore, analytical approaches capable of accurately predicting product sales while providing insights into the contribution of different advertising channels are increasingly needed.

In recent years, the application of machine learning in business and marketing has experienced rapid growth (Sandi Asmoro & Sriyono, 2025). Machine learning techniques are capable of identifying complex relationships among variables and generating accurate predictive models from large datasets. Among the various algorithms used for prediction, XGBoost (Extreme Gradient Boosting) has gained considerable attention due to its high predictive performance, computational efficiency, and ability to handle nonlinear relationships (Zundina Ulya et al., 2025). XGBoost employs an ensemble boosting technique that significantly improves prediction accuracy and incorporates regularization mechanisms to reduce the risk of overfitting (Zhang et al., 2022). Chen and Guestrin (Zhang et al., 2022), who introduced XGBoost, demonstrated that the algorithm consistently outperformed several conventional machine learning methods while maintaining scalability and efficiency. Consequently, XGBoost has been successfully applied in numerous domains, including forecasting, business analytics, customer behavior prediction, and sales prediction.

In addition to advanced machine learning algorithms, conventional statistical approaches such as Linear Regression remain widely used in sales prediction and marketing analytics because of their simplicity, interpretability, and effectiveness in modeling linear relationships among variables. Linear Regression provides a transparent mathematical representation that enables researchers and practitioners to understand the influence of input variables on prediction outcomes. Moreover, it is frequently employed as a baseline model in machine learning studies to evaluate whether more complex algorithms provide significant improvements in predictive performance. Therefore, incorporating Linear Regression as a benchmark model allows a more comprehensive assessment of the effectiveness of XGBoost Regression in predicting product sales and provides a clearer justification for selecting the most suitable model for advertising data analysis.

Several previous studies have reported promising results in sales forecasting and marketing analytics using machine learning approaches. However, most existing studies primarily focus on improving predictive accuracy while providing limited discussion regarding the influence of individual marketing variables on sales outcomes. As a result, business decision-makers often receive accurate predictions without sufficient explanation of the factors driving those predictions. This limitation reduces the practical value of predictive models for strategic marketing planning. In addition to obtaining sales forecasts, companies require insights into which advertising channels contribute most significantly to increasing product sales (Azura et al., 2025).

The issue of model transparency has become increasingly important with the widespread adoption of machine learning in business environments. Many predictive models operate as black boxes, making it difficult for users to understand the reasoning behind model decisions. Recent developments in Explainable Artificial Intelligence (XAI) emphasize the importance of interpretability in improving trust and supporting evidence-based decision-making. Lundberg and Lee (Lundberg & Lee, 2017) proposed SHAP as a unified framework for interpreting machine learning predictions by quantifying the contribution of individual features. Although interpretability has received growing attention in machine learning research, studies that simultaneously perform product sales prediction and advertising media effectiveness evaluation remain relatively limited. This gap highlights the need for approaches that combine predictive capability with interpretable business insights.

To address this research gap, this study proposes a product sales prediction framework using Linear Regression as a baseline model and

XGBoost Regression as the main prediction model, combined with Feature Importance Analysis. The comparison between the two models enables a comprehensive evaluation of predictive performance while providing justification for the use of more advanced machine learning techniques. Feature Importance Analysis is utilized to identify the contribution level of each advertising variable to the prediction results (NURADILLA et al., 2025). Through this approach, companies can determine the most effective marketing channels and better understand the relationship between advertising expenditures and product sales. Furthermore, interpretability analysis can support more accurate, efficient, and data-driven business decision-making processes (Khotimah et al., 2024).

The main contribution of this study is twofold. First, it demonstrates the effectiveness of XGBoost Regression for predicting product sales using advertising expenditure data. Second, it provides an interpretable evaluation of advertising media through Feature Importance Analysis, enabling organizations to identify the factors that most strongly influence sales performance. Unlike previous studies that primarily emphasize prediction accuracy, this research integrates sales prediction and advertising media evaluation within a single analytical framework. The findings are expected to support marketing budget optimization, improve strategic decision-making, and contribute to the development of transparent and reliable marketing analytics systems.

RESEARCH METHODS

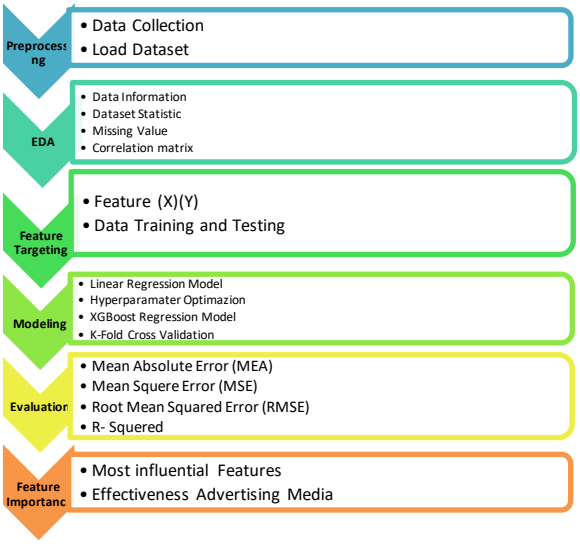

Figure 1. Research Stages

Figure 1 shows the research stages that will be carried out, where each stage consists of the following processes.

Data Collection

This study utilizes the Product Advertising Data dataset obtained from the Kaggle platform. The dataset consists of 300 observations with six independent variables representing advertising expenditures across different marketing channels, namely TV, Billboards, Google Ads, Social Media, Influencer Marketing, and Affiliate Marketing, and one dependent variable, Product_Sold. The dataset was selected because it represents a multi-channel advertising environment commonly used in modern marketing and is suitable for analyzing the relationship between advertising expenditure and product sales. Prior to model development, data preprocessing and exploratory data analysis were conducted to assess data quality. The results showed that the dataset contains no missing values, all variables are numerical, and the data are suitable for machine learning-based regression modeling.

Data Preprocessing

At the data preprocessing stage, data preparation is carried out before the dataset is used in the development of the machine learning model. This stage aims to ensure data quality so that the model can operate optimally. The preprocessing process includes reading the dataset from Google Drive, examining the data structure, checking for missing values, and ensuring that all attributes have appropriate data types. In addition, data separation is also performed to ensure that the dataset is clean, consistent, and ready for analysis and modeling processes. The preprocessing stage is very important because data quality greatly affects the performance and accuracy of the resulting prediction model.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is conducted to understand the characteristics and patterns contained in the advertising dataset (Da Poian et al., 2023). At this stage, descriptive statistical analysis is performed to examine the dataset characteristics, including the distribution, minimum values, maximum values, mean values, and standard deviations of each variable (Katyal et al., 2025). Furthermore, a correlation matrix is generated to visualize and analyze the relationships between advertising expenditure variables and the target variable, Product_Sold (Graffelman & de Leeuw, 2023). The EDA process provides valuable

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.537
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

insights into the dataset, helps identify potential data issues, and supports the selection of relevant features for developing an effective machine learning model.

Feature and Targeting

The feature and targeting stage is conducted by determining the independent variables (features) and dependent variable (target) that will be used in the machine learning model. The feature variables consist of advertising expenditure across various media channels, namely TV, Billboards, Google Ads, Social Media, Influencer Marketing, and Affiliate Marketing, while the target variable is Product_Sold. After the feature and target selection process, the dataset is divided into training and testing datasets using an 80:20 ratio. The training dataset is used to develop the prediction model, whereas the testing dataset is utilized to evaluate the model's performance on unseen data. To improve the reliability and robustness of the model evaluation, K-Fold Cross Validation is applied to the training dataset. This validation technique allows the model to be trained and validated across multiple data subsets, reducing evaluation bias and providing a more reliable estimate of model performance. This stage aims to ensure that the model can effectively learn the relationship between advertising expenditure and product sales while maintaining good generalization capability.

Modelling

At the XGBoost modelling stage, a prediction model is developed using the Extreme Gradient Boosting (XGBoost) algorithm. Linear Regression is employed as a baseline model because it is one of the most widely used regression techniques and provides a simple benchmark for evaluating predictive performance. This algorithm is selected because it has high predictive capability, can handle non-linear relationships between variables, and demonstrates strong performance in regression cases (Wiens et al., 2025). The model is built using several parameters such as the number of estimators, learning rate, and tree depth to improve prediction quality (Noorunnahar et al., 2023). Furthermore, the model is trained using the training data so that the system can learn the relationship patterns between advertising media and product sales. This stage represents the core of the research because it produces a machine learning-based prediction model used to predict the number of products sold. The XGBoost prediction model (Hameed et al., 2025) is expressed as:

$$\hat{y}i = \sum_{k=1}^{K} f_k(xi),\ f_k \in F$$

Symbol Description:

- $\hat{y}_i$ = prediction result for the i-th data
- K = number of decision trees
- f_k = k-th tree function
- x_i = i-th input data

This equation indicates that the prediction result is obtained from the summation of all decision tree outputs built sequentially.

Model Evaluation

The model evaluation stage is conducted to measure the performance of the Linear Regression and XGBoost model in predicting product sales. Model evaluation is performed using several regression metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared (R^2 Score). MAE is used to measure the average absolute prediction error (Robeson & Willmott, 2023), while MSE and RMSE are used to evaluate the model error level (Hodson, 2022). Meanwhile, the R^2 Score is used to measure the model's ability to explain data variation (Mahdi et al., 2025). In addition, visualization of prediction results between actual values and predicted values is conducted to observe the closeness of the model output to the actual data. The evaluation stage is important to ensure that the model has a high level of accuracy and reliability.

Feature Importance

The feature importance stage is conducted to analyze the level of influence of each feature on the XGBoost model prediction results (Kaneko, 2023). This analysis aims to identify which advertising media contribute the most to product sales. Feature importance values are obtained directly from the XGBoost model and then visualized using a bar chart to make them easier to understand. Through this stage, the research not only produces an accurate prediction model but also provides interpretation regarding the factors influencing product sales. This interpretability approach is highly important in supporting data-driven business decision-making and improving the transparency of machine learning models.

RESULTS AND DISCUSSION

Data Collection

The dataset used in this study was obtained from the publicly available Product Advertising Data dataset provided on Kaggle by Navjot Singh and available at Product Advertising Data Dataset. The dataset consists of 300 observations and seven numerical variables, comprising six advertising expenditure variables (TV, Billboards, Google Ads, Social Media, Influencer Marketing, and Affiliate Marketing) and one target variable, Product_Sold. Although the dataset contains a relatively limited number of observations, it provides complete and clean data without missing values, making it suitable for regression and machine learning experiments. Moreover, the dataset represents multiple advertising channels commonly used in digital marketing, allowing the evaluation of their influence on product sales. Therefore, the dataset is considered appropriate for developing and comparing predictive models using Linear Regression and XGBoost, as well as for conducting feature importance analysis to support advertising media evaluation

Data Preprocessing and Exploratory Data Analysis (EDA)

Based on the results of the Exploratory Data Analysis (EDA), the advertising dataset used in this study consists of 300 observation data with a total of 7 variables. These variables consist of six advertising media features, namely TV, Billboards, Google Ads, Social Media, Influencer Marketing, and Affiliate Marketing, along with one target variable, Product_Sold. The amount of data is considered sufficient for use in machine learning modeling processes based on regression algorithms such as XGBoost. The dataset structure indicates that the study focuses on the relationship between various marketing media and the number of products sold.

The descriptive statistical results show that the average advertising expenditure for each media channel ranges from 465 to 517. The TV variable has an average value of 517.43, while the average Product_Sold reaches 7031.52. These mean values indicate that companies generally allocate substantial advertising expenditures across various marketing media. In addition, the relatively high average product sales indicate that the dataset has a good sales scale for analysis using machine learning-based prediction methods.

The standard deviation values for each feature indicate a relatively high level of data variation. For example, the TV variable has a standard deviation of 288.11, while Product_Sold has a standard deviation of 1703.61. These values indicate that the data have a wide distribution around the mean, meaning there are considerable variations among observations. High data variation is very important in machine learning because it helps the model learn more complex relationship patterns and improves the model's generalization capability for new data.

The EDA results also show that each feature has a relatively wide value range. For example, TV advertising expenditure has a minimum value of 1.04 and a maximum value of 998.10, while Product_Sold ranges from 2259 to 12227. The large range of values indicates variations in advertising expenditure intensity and product sales across each observation. This condition allows the XGBoost model to learn the effects of various advertising expenditure levels on increases or decreases in product sales.

Quartile analysis shows the data distribution for each variable. For the Product_Sold variable, the first quartile (25%) is 5922, the median (50%) is 7051, and the third quartile (75%) is 8278. This indicates that most sales data are distributed around this range and that the data distribution is relatively stable. Quartile analysis also helps in understanding data dispersion and detecting possible outliers in the dataset before conducting the machine learning modeling process. Overall, the Exploratory Data Analysis results indicate that the advertising dataset has good quality for use in product sales prediction research using the XGBoost algorithm.

Feature and Targeting

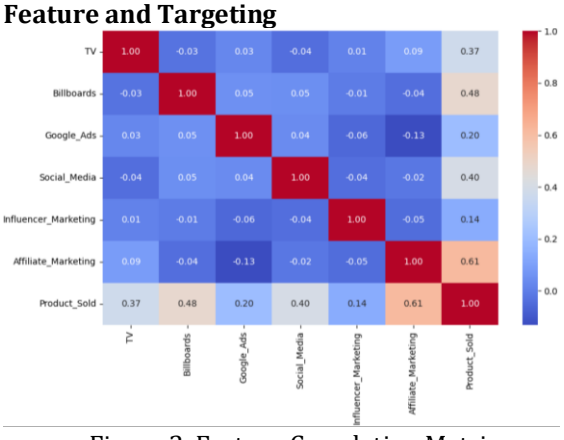

Figure 2. Feature Correlation Matrix

Figure 2 show the Correlation Matrix Heatmap results show the relationships among variables in the advertising dataset used in this study. The Affiliate_Marketing variable has the highest positive correlation with Product_Sold, with a value of 0.61, indicating that affiliate marketing is

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.537
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

the most influential factor in increasing product sales. In addition, Billboards and Social_Media also show relatively strong positive relationships with Product_Sold, with correlation values of 0.48 and 0.40, respectively, while TV has a correlation value of 0.37. Meanwhile, Google_Ads and Influencer_Marketing show lower positive relationships, with correlation values of 0.20 and 0.14. These results indicate that digital and conventional marketing media have different levels of influence on product sales.

In addition to the relationship with the target variable, the heatmap also shows that most features have low correlations with one another, indicating that there is no significant multicollinearity among variables. This condition demonstrates that the dataset has good quality for machine learning modeling using the XGBoost algorithm and supports the interpretability process in analyzing the influence of each advertising medium on the number of products sold.

After observing the correlations among features in the dataset, the next stage is determining the feature and target variables in the machine learning process. In the section X = Product_Sold, all columns except Product_Sold are used as features or input variables for the model to learn data relationship patterns, such as TV, Billboards, Google Ads, Social Media, Influencer Marketing, and Affiliate Marketing. Meanwhile, Y = Product_Sold is used to define Product_Sold as the target or output variable to be predicted by the XGBoost model. This stage is very important because machine learning models require a separation between input and output data in order to learn the relationship between advertising expenditure and the number of products sold. In addition to serving as the basis for the training and testing processes, the separation of features and targets also supports model performance evaluation and interpretability analysis using feature importance.

After determining the features, the training and testing datasets are defined. In this case, the dataset is divided into 20% testing data and 80% training data. In Python, X_train and Y_train are used to train the XGBoost model so that it can learn the relationship patterns between advertising media and product sales, while X_test and Y_test are used to evaluate the model's performance on previously unseen data.

Modelling

At the modeling stage, Linear Regression is first implemented as a baseline model to provide a reference for comparing prediction performance. Subsequently, hyperparameter optimization is performed using GridSearchCV to identify the optimal parameter combination for the XGBoost model and improve its predictive capability. After obtaining the best parameters, the XGBoost Regression model is developed as the main prediction model due to its ability to capture complex relationships among variables and achieve high prediction accuracy. Furthermore, K-Fold Cross Validation is applied to evaluate the robustness and generalization capability of the model by assessing its performance across multiple data partitions. This modeling approach enables a comprehensive comparison between conventional statistical methods and advanced machine learning techniques while ensuring the reliability and stability of the prediction results.

a. Linear Regression (Base Model)

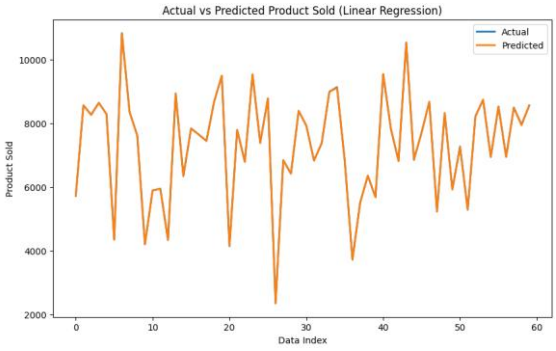

Figure 3. Linear Regression Result

Figure 3 illustrates the comparison between the actual product sales values and the values predicted by the Linear Regression model. It can be observed that the predicted values closely follow the pattern of the actual data across almost all observations. The high degree of overlap between the two curves indicates that the model is capable of accurately capturing the relationship between advertising expenditures and product sales. Although minor deviations are observed in several data points, the overall prediction trend remains highly consistent with the actual values. These results demonstrate that the Linear Regression model provides excellent predictive performance and is able to explain the variation in product sales effectively. The close agreement between actual and predicted values is further supported by the high coefficient of determination (R^2) and the low error values obtained during

model evaluation, indicating that Linear Regression is highly suitable for modeling the relationship between advertising media expenditures and product sales in this dataset.

b. Hyperparameter Optimization

Hyperparameter optimization was performed to obtain the optimal parameter combination for the XGBoost Regression model and improve its predictive performance. In this study, the optimization process was conducted using the GridSearchCV method with five-fold cross validation and the R^2 score as the evaluation metric. Several important hyperparameters were explored, including the number of estimators (n_estimators) with values of 100, 200, and 300, the learning rate (learning_rate) with values of 0.01, 0.05, and 0.1, the maximum tree depth (max_depth) with values of 3, 4, and 5, the subsampling ratio (subsample) with values of 0.8 and 1.0, and the column sampling ratio (colsample_bytree) with values of 0.8 and 1.0. GridSearchCV systematically evaluates all possible parameter combinations and selects the best configuration based on the highest average R^2 score obtained during cross-validation. This optimization process aims to enhance the prediction accuracy of the XGBoost model while reducing the risk of overfitting and improving its generalization capability.

c. XGBoost Regressor

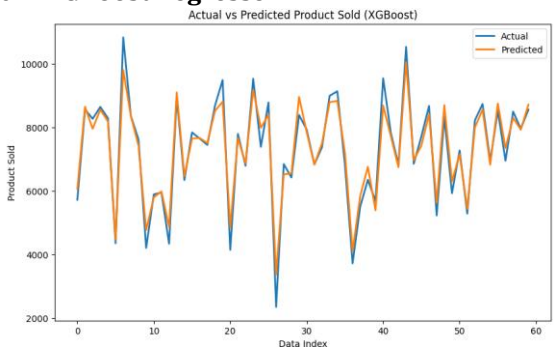

Figure 4. XGBoost Regressor Result

Figure 4 illustrates the comparison between the actual product sales values and the values predicted by the XGBoost model. It can be observed that the predicted values generally follow the overall pattern of the actual data, indicating that the model is capable of capturing the relationship between advertising expenditures and product sales. Although slight deviations are present at several observations, particularly at extreme values, the predicted curve remains closely aligned with the actual curve throughout most of the dataset. These results demonstrate that the XGBoost model provides strong predictive performance and is able to explain the variation in product sales effectively. The close agreement between actual and predicted values is further supported by the high coefficient of determination (R^2) and relatively low prediction errors obtained during model evaluation, indicating that XGBoost is suitable for modeling the relationship between advertising media expenditures and product sales.

d. K-Fold Cross Validation

```
===== CROSS VALIDATION =====
R² Score Each Fold
[0.95029489 0.95675084 0.95337758 0.9560737  0.9314025 ]
Mean R² : 0.949579902264054
Std R² : 0.009369561965977248
```

Figure 5. K-Fold cros validation result

Figure 5 illustrates the distribution of R^2 scores obtained from each fold during the five-fold cross validation process. The R^2 scores obtained from each fold are relatively consistent, ranging from 0.9314 to 0.9568, with an average R^2 value of 0.9496 and a standard deviation of 0.0094. The small variation among the folds indicates that the model exhibits stable performance across different data partitions and possesses good generalization capability. Furthermore, the consistently high R^2 values demonstrate that the XGBoost model is able to explain approximately 94.96% of the variance in product sales, confirming its effectiveness in capturing the relationship between advertising expenditures and sales outcomes. The low standard deviation also suggests that the model is not highly sensitive to changes in the training and testing subsets, thereby reducing the risk of overfitting and increasing the reliability of the prediction results. The use of K-Fold Cross Validation provides a more comprehensive evaluation than a single train-test split, ensuring that the model performance is not dependent on a particular data partition and increasing the reliability of the experimental results

Model Evaluation

Table 1. Evaluation Result

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.537
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

No	Evaluation Method	XGBoost	Linear Regreesion
1	Mean Absolute Error	280.40	7.0854
2	Mean Squared Error	133322.43	75.3659
3	Root Mean Squared Error	365.13	8.6814
4	R^2 Score	0.953	1.0

Table 1 presents the performance comparison between the XGBoost and Linear Regression models using four evaluation metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2 Score). The results indicate that the Linear Regression model achieved superior performance, obtaining an MAE of 7.0854, an MSE of 75.3659, and an RMSE of 8.6814, which are substantially lower than those obtained by the XGBoost model. In addition, Linear Regression achieved an R^2 score of approximately 1.0, whereas XGBoost obtained an R^2 score of 0.953. These findings suggest that Linear Regression was able to explain almost all of the variance in product sales and produce predictions that were much closer to the actual values. Although XGBoost demonstrated strong predictive capability with an R^2 score exceeding 95%, its prediction errors were considerably higher than those of Linear Regression. This result indicates that the relationship between advertising expenditures and product sales in the dataset is predominantly linear, allowing the simpler Linear Regression model to outperform the more complex XGBoost model. Furthermore, the comparison highlights the importance of using baseline models in machine learning studies, as more sophisticated algorithms do not necessarily guarantee better predictive performance.

Feature Importance

Table 2. Feature importance result

No	Feature	Importance
1	Affiliate Marketing	0.376328
2	Billboards	0.231083
3	Social Media	0.153095
4	TV	0.125247
5	Google Ads	0.073181
6	Influencer Marketing	0.041065

Table 2 presents the feature importance scores obtained from the XGBoost model, which indicate the relative contribution of each advertising medium to product sales prediction. The results show that Affiliate Marketing is the most influential feature, with an importance score of 0.3763, suggesting that it has the greatest impact on the model's prediction results. Billboards rank second with an importance value of 0.2311, followed by Social Media with a score of 0.1531. Meanwhile, TV and Google Ads contribute moderately to product sales prediction, with importance scores of 0.1252 and 0.0732, respectively. Influencer Marketing exhibits the lowest importance value of 0.0411, indicating that its contribution to the prediction model is relatively limited compared to the other advertising channels. These findings suggest that Affiliate Marketing and Billboards are the most effective advertising media in influencing product sales within the dataset. Moreover, the feature importance analysis provides valuable insights for business decision-makers by enabling more efficient allocation of marketing budgets and supporting data-driven advertising strategies.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study proposed a product sales prediction framework using Linear Regression as a baseline model and XGBoost Regression combined with Feature Importance Analysis for advertising media evaluation. The dataset used in this study consisted of 300 observations with six advertising expenditure variables, namely TV, Billboards, Google Ads, Social Media, Influencer Marketing, and Affiliate Marketing, and one target variable, Product_Sold. The data preprocessing stage confirmed that the dataset contained no missing values and that all attributes were numerical, making the data suitable for machine learning analysis. Furthermore, Exploratory Data Analysis (EDA) and correlation analysis provided an initial understanding of the relationships among variables and supported the model development process.

At the modeling stage, Linear Regression and XGBoost were implemented and compared using several evaluation metrics. The results showed that Linear Regression achieved superior performance, with an MAE of 7.0854, an MSE of 75.3659, an RMSE of 8.6814, and an R^2 score close to 1.0. Meanwhile, XGBoost obtained an MAE of 280.40, an MSE of 133322.43, an RMSE of 365.13, and an R^2 score of 0.953. These findings indicate

that the relationship between advertising expenditures and product sales in the dataset is predominantly linear, allowing the simpler Linear Regression model to outperform the more complex XGBoost model.

Hyperparameter optimization using GridSearchCV and five-fold cross validation demonstrated that the XGBoost model achieved a mean R^2 score of 0.9496 with a standard deviation of 0.0094, indicating stable performance and good generalization capability. In addition, Feature Importance Analysis revealed that Affiliate Marketing was the most influential variable with an importance score of 0.3763, followed by Billboards (0.2311) and Social Media (0.1531), whereas Influencer Marketing exhibited the lowest contribution. These results provide valuable insights for evaluating advertising media effectiveness and support more efficient and data-driven marketing decision-making.

Suggestion

Future studies are recommended to employ larger and real-world datasets collected from companies or e-commerce platforms to improve the generalizability of the findings. In addition, other machine learning algorithms, such as Random Forest, Support Vector Regression, and deep learning models, may be investigated to provide a more comprehensive comparison of predictive performance. Moreover, more advanced explainable artificial intelligence (XAI) techniques, such as SHAP and LIME, can be incorporated to provide deeper insights into the contribution of individual advertising variables and further enhance the transparency and interpretability of prediction models.

REFERENCES

Adrian Hidayat. (2024). Strategi Periklanan Terbaru Food & Beverage (F&B) Di Dunia Digital Di Asia Tenggara Dan Indonesia Juga Manfaatnya Bagi Kedua Belah Pihak. *EMABI : EKONOMI DAN MANAJEMEN BISNIS*, *3*.

Azura, D., Reksi, P., Kurniawan, B., & Susandri, S. (2025). Model Prediksi Penjualan Berbasis XGBoost – SHAP untuk Decision Support dalam Arsitektur TOGAF Prosiding Semnas 2025 Sekolah Tinggi Teknologi Dumai. *Prosiding Semnas 2025 Sekolah Tinggi Teknologi Dumai Dumai*, *1*(2), 343–357.

Da Poian, V., Theiling, B., Clough, L., McKinney, B., Major, J., Chen, J., & Hörst, S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. *Frontiers in Astronomy and Space Sciences*, *10*(May), 1–17. https://doi.org/10.3389/fspas.2023.1134141

Endrawati Subroto, D., Nurmiati, A. S., Supriatna, E., Khaldun, F., & Endah Fajariana, D. (2024). Sosialisasi Penggunaan Advertising Sosial Media Sebagai Langkah Peningkatan Digital Marketing Pada Home Industry. *Jurnal Pengabdian Kepada Masyarakat Nusantara*, *5*(1), 1509–1517. https://doi.org/10.55338/jpkmn.v5i1.3012

Graffelman, J., & de Leeuw, J. (2023). Improved Approximation and Visualization of the Correlation Matrix. *American Statistician*, *77*(4), 432–442. https://doi.org/10.1080/00031305.2023.2186952

Hameed, M. M., Masood, A., Hamid, A., Elbeltagi, A., Razali, S. F. M., & Salem, A. (2025). Forecasting monthly runoff in a glacierized catchment: A comparison of extreme gradient boosting (XGBoost) and deep learning models. *PLoS ONE*, *20*(5 May), 1–29. https://doi.org/10.1371/journal.pone.0321008

Helen, E., & Rusdi, F. (2023). Komunikasi Pemasaran Salmonbyesther Menggunakan Media Sosial sebagai Media Periklanan. *Kiwari*, *2*(3), 444–451. https://doi.org/10.24912/ki.v2i3.25877

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Kaneko, H. (2023). Interpretation of Machine Learning Models for Data Sets with Many Features Using Feature Importance. *ACS Omega*, *8*(25), 23218–23225. https://doi.org/10.1021/acsomega.3c03722

Katyal, A., Sharma, P. K., & Kannan, M. (2025). Exploratory Data Analysis (EDA) on Undergraduate Data Science Students Through R Programming. *Research Square*, *Icd*, 1–18. https://www.researchsquare.com/article/rs-7422204/v1

Khotimah, K., Yudistira, F., & Ardiansyah, M. (2024). Efisiensi Deep learning untuk Analisis Data dan Pengambilan Keputusan. *Jurnal Insan Peduli Pendidikan (JIPENDIK)*, *2*(2), 79–82.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Advances in Neural Information Processing Systems, *30*.

Mahdi, W. A., Alhowyan, A., & Obaidullah, A. J. (2025). Intelligence analysis of drug nanoparticles delivery efficiency to cancer tumor sites using machine learning models. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-024-84450-9

Noorunnahar, M., Chowdhury, A. H., & Mila, F. A. (2023). A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. *PLoS ONE*, *18*(3 March), 1–15. https://doi.org/10.1371/journal.pone.0283452

NURADILLA, S., SADIK, K., SUHAENI, C., & SOLEH, A. M. (2025). Klasifikasi Halaman SEO Berbasis Machine Learning Melalui Mutual Information dan Random Forest Feature Importance. *MIND Journal*, *10*(1), 114–129. https://doi.org/10.26760/mindjournal.v10i1.114-129

Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLoS ONE*, *18*(2 February), 1–8. https://doi.org/10.1371/journal.pone.0279774

Sandi Asmoro, A., & Sriyono, S. (2025). Peran Machine Learning dalam Pengambilan Keputusan Manajerial di Industri Fintech: Studi Kasus pada Perusahaan Startup. *Journal of Accounting and Finance Management*, *6*(3), 997–1003. https://doi.org/10.38035/jafm.v6i3.2041

Wardani, S. (2023). Analisis Strategis Komunikasi Pemasaran Dalam Meningkatkan Kinerja Ekonomi Perusahaan. *Jurnal Ilmiah Manajemen Profetik*, *1*(2), 76–80. https://doi.org/10.55182/jimp.v1i2.424

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121.

Wiens, M., Verone-Boyle, A., Henscheid, N., Podichetty, J. T., & Burton, J. (2025). A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications. *Clinical and Translational Science*, *18*(3). https://doi.org/10.1111/cts.70172

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6). https://doi.org/10.1177/15501329221106935

Zundina Ulya, F., Khomsah, S., Annisa Ferani Tanjung, N., & Korespondensi, P. (2025). Perbandingan Algoritma Xgboost Dan Lstm Untuk Memprediksi Harga Bitcoin Berdasarkan Harga Harian, Sentimen, Dan Google Trends Index Comparison of Xgboost and Lstm Algorithms To Predicte Bitcoin Price Based on Daily Price, Sentiment, and Google Trends Inde. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *12*(6), 2355–7699. https://trends.google.com/