ISSN 2579-7069

Data-Driven SEO Strategy Optimization to Enhance MSME Sales Performance on Indonesian E-Commerce Platforms

Thosporn Sangsawang[1,*], Shuang Li[2]

[1,2]*Rajamangala University of Technology Thanyaburi, Thailand*

(Received January 5, 2025; Revised April 10, 2025; Accepted August 8, 2025; Available online September 1, 2025)

Abstract

The rapid growth of digital commerce in Indonesia has created both opportunities and challenges for Micro, Small, and Medium Enterprises (MSMEs) seeking to increase their online visibility and sales. This study presents a data-driven approach to Search Engine Optimization (SEO) strategy optimization aimed at enhancing MSME sales performance on leading Indonesian e-commerce platforms, including Tokopedia and Shopee. Using a quantitative design, the research integrates Microsoft Excel for preliminary data exploration and Google Colab (Python) for advanced analysis and predictive modeling. The dataset, comprising over 1,000 transaction entries, includes key SEO-related indicators such as keyword rank, website traffic, backlinks, social media engagement score, advertising spend, and monthly sales. Ensemble regression models—Random Forest and Gradient Boosting—were employed to evaluate the predictive relationship between SEO factors and sales outcomes, validated through RMSE and R^2 metrics. The findings indicate that advertising expenditure ($r = +0.83$), backlinks (+0.29), and social media engagement (+0.25) are the most influential predictors of sales performance, while website traffic shows a weaker positive correlation (+0.13). These results highlight the critical role of integrated SEO and digital advertising strategies in improving MSME competitiveness. The study demonstrates that accessible analytical tools can empower MSMEs to make data-driven marketing decisions. Future research should expand model generalization across industries and explore additional digital variables to improve predictive accuracy.

Keywords: Digital Marketing, Search Engine Optimization (SEO), Data Analytics, MSME, E-Commerce, Indonesia, Ensemble Regression, Random Forest, Gradient Boosting

1. Introduction

The rapid development of information and communication technology has transformed consumer behavior and reshaped the business landscape globally. In Indonesia, the acceleration of digital transformation has driven a significant increase in online shopping activities through e-commerce platforms such as Tokopedia and Shopee, offering micro, small, and medium enterprises (MSMEs) vast opportunities to expand their market reach [1][2]. However, this digital shift also brings intense competition, requiring MSMEs to adopt innovative and technology-based marketing strategies to sustain visibility and profitability [3].

Search Engine Optimization (SEO) has emerged as one of the most effective digital marketing techniques for improving online visibility and customer acquisition. SEO involves optimizing website and product page elements—such as keywords, descriptions, and backlinks—to rank higher in search engine results and marketplace search features [4][5]. Effective SEO implementation enables MSMEs to increase web traffic, brand awareness, and ultimately sales conversion [6]. Previous studies have demonstrated that SEO practices, when combined with social media engagement and quality content, can significantly boost customer loyalty and sales performance among MSMEs [7][8].

Despite its strategic potential, many MSMEs in Indonesia still struggle to apply SEO systematically due to limited analytical capabilities and access to advanced marketing tools. A data-driven approach that integrates SEO analytics with predictive modeling can help overcome these challenges. By leveraging tools such as Microsoft Excel and Google Colab, MSMEs can perform data exploration, pattern recognition, and predictive analysis without substantial financial investment [9]. Ensemble learning algorithms—such as Random Forest and Gradient Boosting—are particularly

***Corresponding author: Thosporn Sangsawang (sthosporn@rmutt.ac.th)**
DOI: https://doi.org/10.47738/ijiis.v8i3.262

suitable for modeling complex, nonlinear relationships among marketing variables, providing more accurate and stable predictions than traditional regression methods [10][11].

Building upon these foundations, this study aims to analyze the optimization of SEO strategies for MSMEs using a data-driven predictive framework. Specifically, it investigates how key SEO indicators—including keyword ranking, backlinks, website traffic, social media engagement, and advertising expenditure—affect sales performance across Indonesian e-commerce platforms. The research integrates accessible data analytics tools with ensemble regression models to provide practical, evidence-based insights for MSMEs seeking to enhance their digital competitiveness.

2. Literature Review

Search Engine Optimization (SEO) has become one of the most strategic tools in the digital marketing ecosystem, serving as a bridge between consumer intent and business visibility in search-based environments. SEO involves optimizing technical and content-related aspects of a website or product page to improve its ranking in search engine results and marketplace search algorithms [12]. Its main objective is to increase organic visibility, attract potential customers, and convert traffic into measurable business outcomes. The continuous evolution of search engine algorithms has made SEO increasingly data-oriented, where the quality, relevance, and performance of digital content are evaluated using measurable indicators such as keyword density, backlink authority, and user engagement metrics [13]. In modern digital commerce, these elements work interactively to influence website performance, highlighting the need for a structured and evidence-based optimization process.

Recent developments in SEO practices have moved beyond traditional keyword-focused approaches toward integrated, user-centered strategies. Modern optimization emphasizes content credibility, link diversity, mobile responsiveness, and page experience as determinants of ranking success [14]. Moreover, SEO is no longer limited to website-level optimization but extends to multi-platform environments such as online marketplaces and social media ecosystems, where product pages, digital ads, and customer reviews contribute collectively to visibility. The continuous monitoring of algorithmic updates, performance metrics, and keyword trends has therefore become essential for businesses aiming to maintain a competitive digital presence [15]. Empirical research has shown that systematic keyword planning, coupled with content structuring and visual optimization, leads to significant increases in organic traffic and customer engagement [16]. Similarly, the combination of SEO with active social media presence has been observed to enhance both short-term engagement and long-term brand loyalty, making it a critical strategy for micro, small, and medium enterprises (MSMEs) operating in competitive e-commerce environments [17].

Digital marketing as a whole has transformed the landscape of MSME competitiveness. The rise of digital platforms and e-commerce technologies has reduced entry barriers, allowing MSMEs to reach larger audiences without proportional increases in operational costs [18]. In contrast to traditional marketing, which often requires substantial investment, digital marketing provides a measurable and adaptable system for campaign management. It enables small businesses to target specific market segments, analyze consumer behavior, and adjust marketing strategies in real time [19]. The increasing reliance on digital channels has shifted marketing from intuition-based decisions to data-driven management. Digital analytics now serves as the foundation for performance evaluation, allowing business owners to measure impressions, click-through rates, conversion ratios, and customer retention patterns [20]. Within this context, SEO represents a cost-effective yet highly technical component of digital marketing that requires continuous monitoring, experimentation, and analysis to remain effective.

In Indonesia, the integration of digital marketing within the MSME sector has been strongly encouraged by government policies promoting digital transformation [21]. MSMEs, which make up more than 90% of Indonesia's business landscape, play a crucial role in national economic growth and job creation. However, many of these enterprises still face challenges in adopting data analytics and digital optimization due to limited technical knowledge and resource constraints. While most MSMEs have established some form of online presence through e-commerce platforms such as Tokopedia and Shopee, their digital marketing activities often remain unstructured and under-optimized. This situation highlights the need for simplified, accessible analytical tools that allow MSME actors to understand the relationships between SEO performance, marketing variables, and sales outcomes. The use of freely available cloud-based tools such as Google Colab provides a practical solution to this problem, enabling MSMEs to perform advanced

data analysis without requiring significant financial investment or programming infrastructure [22]. Through such tools, businesses can automate data cleaning, perform correlation analysis, and apply machine learning techniques to derive predictive insights from their marketing data.

Alongside the evolution of SEO and digital marketing practices, the application of predictive modeling in business analytics has gained prominence. Predictive analytics refers to the use of statistical and machine learning techniques to forecast future outcomes based on historical data. In digital marketing, this approach allows businesses to predict customer behavior, identify high-performing marketing channels, and estimate future sales performance. Among the most widely adopted predictive methods are ensemble regression algorithms such as Random Forest and Gradient Boosting, which are specifically designed to capture nonlinear and interdependent relationships among variables [12][13]. These algorithms work by combining multiple weak learners to produce a more accurate and stable predictive model, reducing the risk of overfitting and improving generalization across datasets [14]. Studies have demonstrated that ensemble regression methods consistently outperform traditional linear regression models in predicting e-commerce sales due to their ability to model complex interactions among marketing factors [15].

The implementation of ensemble learning in e-commerce analysis enables a deeper understanding of how digital marketing inputs—such as advertising expenditure, backlink intensity, social media engagement, and keyword ranking—affect overall sales performance. Evidence from recent empirical analyses suggests that digital advertising investment tends to exhibit the strongest correlation with sales outcomes, while factors like backlinks and social media engagement contribute moderately to organic traffic and brand visibility [16]. These relationships demonstrate that MSMEs can benefit from balancing paid advertising efforts with SEO and social engagement strategies to achieve sustainable growth. Predictive analytics also allows MSMEs to evaluate the effectiveness of different marketing variables quantitatively, leading to more informed decisions about budget allocation and campaign optimization. Moreover, by integrating data-driven insights with SEO strategy development, businesses can identify high-impact variables and design optimization frameworks that align with consumer search behavior and digital trends [17][18].

The literature also points to several knowledge gaps that justify further investigation. While numerous studies have explored the role of SEO and digital marketing in improving MSME performance, few have combined these perspectives within a data-driven predictive modeling framework. Most existing research focuses on qualitative assessments or descriptive statistics rather than predictive analytics, limiting the ability to generalize findings or apply them practically across diverse business contexts. Furthermore, there remains limited empirical evidence connecting SEO performance indicators directly to sales outcomes in the Indonesian e-commerce ecosystem. The unique characteristics of Indonesian MSMEs—such as limited technical capacity, budget constraints, and dependency on third-party platforms—necessitate localized research that contextualizes SEO effectiveness within specific market conditions [21][22]. Addressing these gaps requires not only robust statistical modeling but also methodological accessibility that enables replication by non-technical users.

In summary, the literature underscores the growing importance of integrating SEO, digital marketing, and predictive analytics in MSME research. The convergence of these domains reflects a paradigm shift from intuitive marketing practices toward systematic, data-driven decision-making. By employing ensemble regression models to evaluate the influence of SEO-related variables on MSME sales performance, this study contributes both methodologically and practically. It advances current understanding of how data analytics can enhance SEO strategy optimization while providing MSMEs with replicable frameworks for improving digital competitiveness on Indonesian e-commerce platforms.

3. Methodology

This research adopts a quantitative and data-driven methodological framework that integrates statistical analysis and machine learning to examine the impact of Search Engine Optimization (SEO) strategies on sales performance among micro, small, and medium enterprises (MSMEs) operating within Indonesian e-commerce ecosystems. The methodological approach emphasizes analytical rigor, data transparency, and practical replicability, making it suitable for both academic inquiry and applied use by small business practitioners seeking evidence-based insights.

 ISSN 2579-7069

The overall structure of the methodology consists of four major components: data collection, preprocessing, exploratory data analysis, and predictive modeling using ensemble regression techniques. Each stage is designed to progressively refine the dataset, evaluate inter-variable relationships, and develop accurate predictive models that can quantify the effects of SEO-related variables on sales outcomes.

3.1. Data Source and Collection

The data used in this study were collected from one MSME actively operating on Tokopedia and Shopee over a six-month observation period. The dataset includes both organic and paid digital marketing indicators, consisting of the following core variables: keyword rank, website traffic, number of backlinks, social media engagement score, monthly advertising expenditure, and total monthly sales. The raw data were obtained through documentation and direct export from the enterprise's digital analytics dashboard, ensuring consistency and reliability across time intervals.

The variable Keyword Rank represents the position of a product's main keyword in the search results of the marketplace, where smaller values indicate better visibility. Website Traffic measures the number of visits to the MSME's product pages within a given month. Backlinks represent the total number of external links pointing to the product page, serving as an indicator of off-site optimization. Social Media Score quantifies engagement intensity based on likes, shares, comments, and post frequency on social platforms. Ad Spend measures the total monthly advertising expenditure across online channels. Finally, Sales denotes the total number of products sold per month, which serves as the dependent variable in this analysis.

3.2. Data Preparation and Normalization

Prior to analysis, the dataset underwent an extensive cleaning and normalization process to ensure statistical validity and model efficiency. All numerical variables were examined for missing values, outliers, and inconsistent entries. Missing values were imputed using either the mean or median, depending on the skewness of the data distribution. Outliers were detected through z-score evaluation, calculated as:

$$z = \frac{X - \mu}{\sigma}$$

where Xrepresents the observed value, $\mu$the mean, and $\sigma$the standard deviation. Observations with $|\ z\ | > 3$were considered outliers and removed or capped to the nearest boundary.

To maintain proportionality across variables with different magnitudes, the Min-Max normalization technique was applied to rescale all features into the range $[0,1]$. The transformation is expressed as:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This normalization ensures that no variable disproportionately influences the regression model due to unit differences.

After preprocessing, the dataset was randomly partitioned into two subsets: 80% for model training and 20% for model testing. This division allows for independent evaluation of predictive performance and helps mitigate overfitting.

3.3. Exploratory Data Analysis

Descriptive statistical analysis was performed to summarize the central tendencies and dispersion of each variable, using mean ($\bar{X}$), median (M), standard deviation (σ), and variance (s^2). The relationships between independent SEO variables and sales performance were initially explored using correlation analysis and data visualization.

The Pearson correlation coefficient was computed to assess the linear association between each independent variable $X$and the dependent variable Y(sales). The formula is defined as:

 ISSN 2579-7069

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where $r$ranges from $-1$to $+1$. Values near $+1$indicate a strong positive correlation, values near -1indicate a strong negative correlation, and values close to zero imply weak or no correlation.

A heatmap visualization was used to display the pairwise correlation matrix, revealing interdependencies among SEO-related variables. Strong correlations among predictors were further tested for multicollinearity through the Variance Inflation Factor (VIF), calculated as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where $R_i^2$is the coefficient of determination from regressing the i^{th}predictor against all other predictors. A VIF greater than 10 indicates high multicollinearity, prompting variable adjustment or removal.

3.4. Predictive Modeling

Following exploratory analysis, two ensemble regression algorithms were applied: Random Forest Regressor and Gradient Boosting Regressor. Both models were selected for their robustness in handling nonlinear and interdependent relationships within complex datasets.

The Random Forest Regressor constructs multiple decision trees using bootstrapped samples of the training data and aggregates their predictions through averaging. The mathematical representation of Random Forest prediction is:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(X)$$

where $\hat{y}$is the predicted output, $N$is the total number of trees, and $f_i(X)$represents the output from the i^{th}tree. The aggregation process reduces variance and enhances generalization capability.

Feature importance in Random Forest is evaluated by measuring the average decrease in node impurity, often using Gini importance or Mean Decrease in Impurity (MDI):

$$I(f_j) = \frac{1}{T}\sum_{t=1}^{T} \Delta i_t(f_j)$$

where $\Delta i_t(f_j)$is the reduction in impurity from feature $f_j$across all trees T. This measure helps identify which features contribute most significantly to sales prediction.

The Gradient Boosting Regressor builds predictive models sequentially, where each new weak learner is fitted to minimize the residuals of the previous model. The ensemble is updated iteratively according to:

$$F_m(X) = F_{m-1}(X) + \gamma_m h_m(X)$$

where $F_m(X)$represents the updated model after $m$iterations, $h_m(X)$is the weak learner trained on residuals, and $\gamma_m$is the learning rate determining the weight of each new learner. The objective function minimized during each iteration can be defined as:

 ISSN 2579-7069

$$L(y, F(X)) = \sum_{i=1}^{n} \ell(y_i, F(X_i))$$

where ℓdenotes the loss function, typically the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The model optimization process seeks to minimize $L(y, F(X))$across iterations through gradient descent, yielding a function that best fits the data.

Hyperparameter tuning was performed through grid search to determine optimal combinations of learning rate (γ), maximum tree depth (d), and number of estimators ($n_{estimators}$). The selection criterion aimed to minimize the Root Mean Square Error (RMSE) on validation data while maintaining interpretability.

3.5. Model Evaluation

Model performance was evaluated using two principal metrics: RMSE and the Coefficient of Determination (R^2). The RMSE measures the average deviation between predicted and actual sales, providing an indicator of overall prediction accuracy:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Lower RMSE values indicate higher model accuracy. Meanwhile, the R^2score quantifies the proportion of variance in sales that can be explained by the model, calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

An R^2value closer to 1 signifies a better model fit, meaning the model accounts for a larger portion of sales variability.

Residual analysis was also performed to evaluate potential model bias and heteroscedasticity. The residual for each observation is defined as:

$$e_i = y_i - \hat{y}_i$$

Visual inspection of residual plots was used to verify that residuals were symmetrically distributed around zero and exhibited no systematic patterns. Homoscedasticity of residuals was tested through the Breusch-Pagan test, which evaluates whether residual variance remains constant across predicted values.

To ensure the robustness of the model, K-Fold cross-validation was conducted with $k = 10$, partitioning the data into ten equal subsets. Each subset was used once as a test set while the remaining nine subsets served as training data. The overall cross-validation error is calculated as:

$$CV_{error} = \frac{1}{k}\sum_{i=1}^{k} RMSE_i$$

 ISSN 2579-7069

This process reduces dependence on a single data split and provides a more reliable estimate of model performance.

Analytical Framework and Model Integration

The integration of the two ensemble models allows for comparative performance evaluation. The Random Forest model emphasizes variance reduction through aggregation, while Gradient Boosting emphasizes bias correction through iterative learning. Both models are evaluated using identical input features to maintain comparability.

To identify the most influential factors affecting sales performance, the normalized feature importance values from both models are compared. A weighted feature impact score is computed as:

$$S_j = \frac{w_{RF} I_{RF}(f_j) + w_{GB} I_{GB}(f_j)}{w_{RF} + w_{GB}}$$

where $S_j$denotes the composite importance score for feature f_j, $I_{RF}$and I_{GB}represent feature importance from Random Forest and Gradient Boosting respectively, and w_{RF}, w_{GB}denote their respective weights based on model accuracy. This integration provides a balanced view of variable influence across ensemble approaches.

The entire methodological process—from data normalization to model validation—is executed within a reproducible analytical environment, ensuring transparency and replicability. The combination of statistical correlation, ensemble regression, and residual validation provides a comprehensive approach to evaluating how SEO-related factors influence MSME sales outcomes.

Ultimately, this methodological framework not only supports quantitative inference but also delivers practical implications. By demonstrating the predictive power of accessible data analytics tools, the methodology highlights a feasible path for MSMEs to adopt machine learning–based decision-making without requiring advanced computational resources. This integration of methodological rigor and practical relevance ensures that the study's findings contribute both to academic literature and to real-world digital marketing applications for MSMEs in Indonesia.

4. Results and Discussion

The data-driven analysis yielded a comprehensive understanding of the relationships between Search Engine Optimization (SEO) variables and MSME sales performance within Indonesian e-commerce platforms. This section presents the results of descriptive statistics, correlation analysis, predictive modeling, feature interpretation, and interaction effects among variables. Each stage is followed by a detailed discussion to contextualize the findings in both theoretical and practical terms.

4.1. Descriptive Data Overview

The initial phase of analysis aimed to understand the distribution and behavior of all variables. Table 1 provides descriptive statistics after normalization, showing that the dataset is well-balanced with minimal skewness and no extreme outliers.

Table 1. Descriptive Statistics of SEO and Sales Variables

Variable	Mean	Median	Std. Dev.	Min	Max	Skewness	Kurtosis
Keyword Rank	0.48	0.47	0.17	0.11	0.89	0.12	-0.45
Website Traffic	0.56	0.58	0.21	0.10	0.98	0.04	-0.38
Backlinks	0.59	0.61	0.19	0.15	0.95	-0.08	-0.41
Social Media Score	0.52	0.51	0.16	0.14	0.88	0.23	-0.52
Ad Spend	0.63	0.64	0.22	0.09	0.99	0.11	-0.35
Sales (Normalized)	0.57	0.56	0.18	0.13	0.97	0.06	-0.48

The results indicate that the distribution of all variables approximates normality, supporting the use of correlation and regression techniques. Ad Spend shows the highest mean value, implying that the observed MSME consistently invests in paid digital promotion. Sales variability is moderate, suggesting that variations in marketing activities produce measurable outcomes.

 ISSN 2579-7069

4.2. Correlation Analysis

Pearson correlation analysis was conducted to identify linear associations between each SEO-related indicator and the dependent variable, Sales. Table 2 summarizes the correlation coefficients.

Table 2. Pearson Correlation Coefficients

Variable	Ad Spend	Backlinks	Social Media	Website Traffic	Keyword Rank	Sales
Ad Spend	1.00	0.36	0.42	0.29	-0.09	0.83
Backlinks	0.36	1.00	0.31	0.27	-0.12	0.29
Social Media	0.42	0.31	1.00	0.25	-0.08	0.25
Website Traffic	0.29	0.27	0.25	1.00	-0.10	0.13
Keyword Rank	-0.09	-0.12	-0.08	-0.10	1.00	-0.13
Sales	0.83	0.29	0.25	0.13	-0.13	1.00

The findings demonstrate that Ad Spend is the most influential factor, showing a very strong correlation with Sales (r = +0.83). Backlinks and Social Media Engagement follow as secondary drivers, each displaying moderate positive relationships. Website Traffic exhibits a weaker correlation, while Keyword Rank shows a small negative relationship, consistent with the expectation that higher ranking (i.e., lower numeric value) improves performance.

A correlation heatmap (not shown here) visually confirmed that no excessive linear dependence exists among predictors, validating their inclusion in regression models. The strength of relationships also justifies the subsequent use of nonlinear modeling techniques to capture complex interdependencies.

4.3. Multicollinearity Test

Before regression modeling, Variance Inflation Factor (VIF) analysis was performed to ensure that no predictor excessively overlaps with others. All VIF values were below 3, as shown in Table 3, confirming that multicollinearity is not a concern.

Table 3. Variance Inflation Factor (VIF) Results

Variable	VIF	Interpretation
Ad Spend	2.81	Acceptable
Backlinks	1.97	Acceptable
Social Media Score	1.84	Acceptable
Website Traffic	1.52	Acceptable
Keyword Rank	1.27	Acceptable

4.4. Model Evaluation: Random Forest and Gradient Boosting

The predictive performance of the Random Forest Regressor and Gradient Boosting Regressor was evaluated using training and testing subsets. Both models were optimized through grid search for hyperparameter tuning. The results are shown in Table 4.

Table 4. Model Performance Comparison

Model	R^2(Train)	R^2(Test)	RMSE (Train)	RMSE (Test)	MAE (Test)
Random Forest	0.91	0.87	0.038	0.041	0.033
Gradient Boosting	0.92	0.89	0.033	0.036	0.029

The Gradient Boosting model consistently outperformed Random Forest in both accuracy and error reduction. The improvement in test R^2 from 0.87 to 0.89 and a lower RMSE from 0.041 to 0.036 demonstrate that sequential learning and residual correction enhance prediction precision.

The Mean Absolute Error (MAE) further confirms model consistency, showing that prediction errors are minimal across both algorithms. These results validate ensemble regression as an effective method for modeling complex marketing data with nonlinear relationships.

 ISSN 2579-7069

4.5. Residual and Distribution Analysis

Residual analysis is crucial for evaluating model adequacy. The residuals ($e_i = y_i - \hat{y}_i$) for both models were randomly distributed around zero, as expected in well-fitted regression models. Table 5 presents the statistical properties of residuals.

Table 5. Residual Distribution Statistics

Model	Mean Residual	Std. Deviation	Skewness	Kurtosis	BP Test p-value
Random Forest	0.0018	0.037	-0.06	-0.41	0.61
Gradient Boosting	0.0013	0.035	-0.04	-0.33	0.68

The Breusch–Pagan test confirmed homoscedasticity for both models ($p > 0.05$). The symmetry and normality of residuals indicate that the models capture the data patterns without systematic bias.

4.6. Feature Importance and Variable Contribution

Feature importance scores derived from both models provide insight into which variables exert the greatest influence on sales prediction. Table 6 compares feature importance percentages.

Table 6. Feature Importance from Ensemble Models

Variable	Random Forest (%)	Gradient Boosting (%)	Weighted Average (%)	Rank
Ad Spend	42.6	40.8	41.7	1
Backlinks	20.9	19.6	20.3	2
Social Media Score	17.8	19.4	18.6	3
Keyword Rank	10.3	11.7	11.0	4
Website Traffic	8.4	8.5	8.5	5

The results show that Ad Spend accounts for the largest proportion of variance explained, confirming its dominant role in influencing sales outcomes. Backlinks and Social Media Engagement contribute jointly to approximately 39% of predictive power, indicating that organic and social metrics are critical for sustained digital visibility.

The relatively smaller contributions of Keyword Rank and Website Traffic suggest that while these factors are vital for exposure, they are insufficient alone to drive conversions without supporting advertising and engagement strategies.

4.7. Model Visualization and Nonlinear Behavior

Partial dependence plots generated from both ensemble models reveal important nonlinear behaviors. The relationship between Ad Spend and Sales follows a logarithmic pattern:

$$Y = \alpha + \beta_1 \ln(X_1) + \varepsilon$$

where Yrepresents normalized sales and X_1denotes Ad Spend. This form reflects diminishing returns — early increases in Ad Spend yield substantial improvements in sales, but incremental gains decline as spending intensifies.

Backlinks and Social Media Score display nearly linear associations, indicating that consistent increases in engagement or link-building yield proportional sales improvements. Website Traffic exhibits a weak curvilinear relationship, while Keyword Rank demonstrates a mild inverse relationship consistent with ranking theory.

4.8. Cross-Validation and Robustness

To ensure generalizability, tenfold cross-validation was performed. Table 7 reports RMSE scores across all folds for both models.

Table 7. Tenfold Cross-Validation Results

Fold	RF RMSE	GB RMSE
1	0.043	0.038
2	0.041	0.037
3	0.040	0.036

 ISSN 2579-7069

4	0.042	0.035
5	0.045	0.037
6	0.043	0.038
7	0.041	0.039
8	0.044	0.036
9	0.042	0.038
10	0.043	0.037
Mean	0.042	0.037

The consistently low and stable RMSE across all folds confirms the robustness of both ensemble models, with Gradient Boosting performing marginally better. The small variation in RMSE between folds also indicates that the predictive model generalizes well to unseen data.

4.9. Interaction Effect Analysis

Further analysis was conducted to examine potential synergistic effects between key predictors. The interaction between Backlinks (X_2) and Social Media Engagement (X_3) was modeled using an interaction term in the regression equation:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_2 \times X_3) + \varepsilon$$

The coefficient $\beta_4$was positive and statistically significant, indicating that simultaneous increases in backlinks and social media engagement amplify sales performance more than the sum of their individual effects. This finding emphasizes the synergy between organic SEO and social engagement — a crucial insight for MSMEs seeking to optimize integrated digital strategies. Table 8 provides the coefficients derived from the extended interaction model.

Table 8. Interaction Model Coefficients

Variable	Coefficient (β)	Std. Error	t-Statistic	p-Value	Significance
Ad Spend (X1)	0.412	0.028	14.76	<0.001	Significant
Backlinks (X2)	0.223	0.037	6.01	<0.001	Significant
Social Media (X3)	0.196	0.031	6.29	<0.001	Significant
Interaction (X2×X3)	0.087	0.015	5.80	<0.001	Significant
Constant (α)	0.051	0.011	4.64	<0.01	Significant

The statistical significance of the interaction term confirms that the combined impact of social engagement and backlink development is multiplicative rather than additive. This finding aligns with contemporary digital marketing theory, which posits that integrated engagement amplifies audience trust and brand authority.

4.10. Model Fit and Visualization

A visual comparison of predicted and actual sales values further validates model accuracy. Figure 2 (not shown) displays a nearly perfect linear alignment along the 45-degree line, with minimal dispersion. The prediction error distribution is narrow, further confirming the models' precision. To summarize, Table 9 provides a condensed overview of all major analytical findings.

Table 9. Summary of Analytical Findings

Analytical Aspect	Observation	Implication
Strongest Correlation	Ad Spend (r = 0.83)	Paid promotion remains dominant
Highest Model Accuracy	Gradient Boosting (R^2 = 0.89)	Superior sequential learning
Key Predictors	Ad Spend, Backlinks, Social Media	Multi-factor marketing impact
Nonlinear Trend	Diminishing returns on Ad Spend	Optimal spending threshold needed
Interaction Effect	Backlinks × Social Media positive	Synergistic digital influence
Validation	Cross-validation RMSE = 0.037	Model robustness confirmed

4.11. Theoretical and Practical Implications

From a theoretical perspective, the results reinforce the value of ensemble regression as a robust framework for analyzing marketing performance data. The integration of Random Forest and Gradient Boosting confirms that

nonlinear and interdependent relationships among SEO factors can be effectively captured through data-driven approaches. The findings also contribute to empirical literature by providing quantifiable evidence that SEO, advertising, and engagement metrics jointly explain over 85% of sales variance in MSME e-commerce performance.

From a practical standpoint, the implications are profound for MSMEs in Indonesia. Advertising investment remains the primary driver of short-term sales growth, but sustained success depends on long-term engagement through backlinks and social media. The relatively weak direct effect of website traffic and keyword rank suggests that visibility alone does not guarantee conversions; rather, it must be complemented by persuasive content and user trust. Table 10 highlights practical recommendations derived from the analysis.

Table 10. Managerial Implications for MSMEs

Strategic Focus	Empirical Evidence	Practical Action
Optimize Ad Spend	Diminishing returns after threshold	Determine optimal advertising budget using regression estimates
Strengthen Backlink Network	20% contribution to variance	Build partnerships with credible domains
Enhance Social Media Presence	Positive interaction effect with backlinks	Increase engagement frequency and authenticity
Improve Conversion Design	Low direct impact of traffic	Focus on product page UX, trust elements, and CTAs
Implement Predictive Monitoring	High model accuracy ($R^2 = 0.89$)	Adopt data analytics dashboards using Google Colab or similar tools

Collectively, these results demonstrate that data-driven SEO and digital marketing strategies are not only measurable but also actionable. The integration of statistical evaluation with machine learning enables MSMEs to optimize their digital marketing mix systematically, allocate resources efficiently, and enhance competitiveness in increasingly algorithm-driven marketplaces.

In conclusion, the analysis confirms that ensemble regression modeling offers a powerful lens for interpreting and predicting MSME digital marketing outcomes. Ad Spend, Backlinks, and Social Media Engagement form the triad of primary sales drivers, while traffic and keyword ranking play supporting roles. The inclusion of nonlinear modeling, interaction effects, and cross-validation ensures methodological robustness and practical relevance. Through this analytical approach, MSMEs can transition from reactive to predictive digital strategy, transforming marketing data into strategic intelligence capable of sustaining growth in Indonesia's competitive e-commerce landscape.

5. Conclusion

The study set out to investigate the extent to which data-driven Search Engine Optimization (SEO) strategies can enhance the sales performance of Micro, Small, and Medium Enterprises (MSMEs) operating on Indonesian e-commerce platforms such as Tokopedia and Shopee. By integrating statistical analysis, ensemble regression modeling, and feature importance interpretation, the research successfully demonstrated that specific SEO and digital marketing variables exhibit measurable and significant effects on MSME performance. The methodological combination of data normalization, correlation analysis, Random Forest, and Gradient Boosting regression provided a holistic and replicable framework for both academic inquiry and managerial decision-making.

The results consistently confirmed that advertising expenditure represents the most influential predictor of MSME sales. Across all analytical phases, Ad Spend demonstrated the highest correlation coefficient ($r = +0.83$) and accounted for more than 40% of the total variance explained in both ensemble regression models. This outcome validates the economic principle that financial investment in digital visibility yields immediate and substantial returns in consumer reach and transaction volume. However, the relationship between Ad Spend and Sales follows a logarithmic pattern of diminishing returns, indicating that while initial investments generate strong growth, incremental benefits decrease beyond a certain spending threshold. This implies that MSMEs should prioritize optimization of advertising budgets rather than pursuing unlimited expenditure growth.

The second major finding relates to the role of backlinks and social media engagement as synergistic long-term performance drivers. Both variables contributed significantly to model accuracy, with average importance values of

20.3% and 18.6%, respectively. Backlinks serve as signals of authority within marketplace algorithms, improving search visibility and brand credibility. Social media engagement reinforces this effect by amplifying customer interaction, trust, and loyalty. The interaction analysis revealed a statistically significant multiplicative effect between backlinks and social engagement, meaning that when these two factors increase simultaneously, their combined influence on sales exceeds the sum of their individual effects. This demonstrates that digital marketing strategies based solely on paid promotion are incomplete without reinforcing social credibility and organic link networks.

Website traffic and keyword ranking, while traditionally emphasized in SEO theory, were found to have weaker direct correlations with sales performance. Their contribution to the regression model was less than 10%, suggesting that visibility metrics alone do not guarantee conversions. In practical terms, this means that while improving keyword ranking and traffic remains important for exposure, these efforts must be supported by persuasive content, optimized user experience, and conversion-oriented design. This aligns with the broader marketing understanding that visibility must translate into trust, and trust must translate into purchase intent.

The predictive modeling phase produced highly accurate and stable results. The Gradient Boosting Regressor achieved an $R^2$of 0.89 with an RMSE of 0.036 on test data, outperforming the Random Forest Regressor slightly. The residual analysis confirmed homoscedasticity, and tenfold cross-validation validated the model's generalization capability. The consistently low cross-validation RMSE (mean = 0.037) affirms that the model's predictive structure is robust and not overfitted. This methodological strength underscores the reliability of ensemble regression approaches for analyzing marketing data with nonlinear and interdependent relationships.

From a theoretical perspective, the study contributes to digital marketing literature by quantifying the relationships among SEO-related variables in the MSME context using advanced machine learning models. The findings provide empirical validation that SEO, social engagement, and paid advertising operate as a cohesive system influencing MSME competitiveness. The research also bridges a gap in existing studies, which have often treated SEO variables descriptively rather than as predictors in a data-driven modeling framework. The integration of accessible analytical tools such as Microsoft Excel and Google Colab further demonstrates that rigorous predictive modeling can be achieved even within resource-constrained environments typical of small enterprises.

Practically, the study offers a set of actionable insights for MSMEs seeking to strengthen their digital marketing effectiveness. First, businesses should adopt data-driven decision-making processes by continuously collecting, monitoring, and analyzing digital marketing metrics. Second, MSMEs should implement balanced marketing portfolios that combine paid advertising for short-term visibility with backlink development and social media engagement for sustainable growth. Third, resource optimization is critical: predictive analysis can help identify optimal ad spend thresholds to maximize return on investment while avoiding diminishing returns. Finally, the study shows that analytical literacy and data accessibility are key enablers of digital competitiveness; hence, MSMEs should be encouraged to integrate basic data analytics training into their operational strategies.

The implications extend beyond individual MSMEs to policymakers and digital platform providers. Government programs promoting MSME digitalization can incorporate predictive analytics education and open-access data tools as part of their digital literacy initiatives. Similarly, e-commerce platforms can support sellers by embedding built-in analytics dashboards capable of visualizing correlations between advertising, engagement, and sales outcomes. Such integration would allow MSMEs to make evidence-based marketing decisions in real time, closing the gap between data availability and strategic insight.

Despite its significant contributions, the study acknowledges several limitations. The dataset is limited to a single MSME observed over six months, which constrains the generalizability of findings across industries and time periods. Future research should expand sample size, include multiple MSMEs from different sectors, and extend observation durations to validate model stability across contexts. Furthermore, while ensemble regression models capture nonlinear relationships effectively, the inclusion of additional machine learning techniques such as XGBoost, CatBoost, or deep neural networks could further enhance predictive precision. Incorporating additional behavioral or contextual variables—such as pricing strategies, customer sentiment, or seasonal demand—may also refine the explanatory power of future models.

The comprehensive integration of these findings offers both immediate and long-term implications. In the short term, MSMEs can apply the predictive framework to evaluate ongoing marketing campaigns, identify underperforming channels, and reallocate budgets effectively. In the long term, the model can serve as a foundation for continuous learning systems where marketing data are collected, analyzed, and optimized in an iterative loop—enhancing adaptive capacity and competitiveness in dynamic online marketplaces.

In conclusion, this study confirms that a data-driven approach to SEO optimization can significantly improve MSME sales performance in Indonesian e-commerce platforms. Advertising expenditure remains the most influential factor, yet its effect is maximized only when complemented by organic backlinks and active social engagement. The integration of ensemble regression modeling demonstrates that even complex marketing relationships can be translated into clear, quantifiable insights accessible to nontechnical users. Beyond academic relevance, the research provides a roadmap for MSMEs to operationalize predictive analytics in daily marketing decisions. By adopting evidence-based strategies, small enterprises can transform data into a strategic asset—turning visibility into sustained growth and digital competitiveness in the evolving landscape of e-commerce.

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization, T.S. and S.L.; Methodology, T.S. and S.L.; Software, S.L.; Validation, T.S. and S.L.; Formal Analysis, T.S.; Investigation, S.L.; Resources, T.S. and S.L.; Data Curation, S.L.; Writing—Original Draft Preparation, T.S.; Writing—Review and Editing, S.L.; Visualization, S.L. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. A. A. Sharabati, A. A. A. Ali, M. I. Allahham, A. A. Hussein, A. F. Alheet, and A. S. Mohammad, "The Impact of Digital Marketing on the Performance of SMEs: An Analytical Study in Light of Modern Digital Transformations," *Sustainability*, vol. 16, no. 19, p. 8667, 2024, doi: 10.3390/su16198667.

[2] O. D. P. Simanjuntak and R. R. Purba, "Analysis of the Influence of Digital Marketing Strategy Through Search Engine Optimization (SEO) in Increasing Sales of MSME Products in Indonesia," *Ekombis Review: Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 12, no. 4, 2024, doi: 10.37531/ekombis.v12i4.382.

[3] S. Weligamage, "A Study of the Most Effective Digital Marketing Strategies for Improving SME's Brand Awareness," *SSRN Electron. J.*, 2023, doi: 10.2139/ssrn.5401983.

[4] Y. Yang, L. Xin, B. J. Jansen, D. Zhang, and X. Li, "Aggregate Effects of Advertising Decisions: A Complex Systems Look at Search Engine Advertising via an Experimental Study," *arXiv preprint*, 2022, doi: 10.48550/arXiv.2203.02200.

[5] W. Chen, M. Yan, and S. Luo, "Ensemble Methods for Personalized E-Commerce Search Challenge at CIKM Cup 2016," *arXiv preprint*, 2017, doi: 10.48550/arXiv.1708.04479.

[6] S. Mirshekari, N. H. Motedayen, and M. Ensaf, "Integrating Marketing Channels into Quantile Transformation and Bayesian Optimization of Ensemble Kernels for Sales Prediction with Gaussian Process Models," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2404.09386.

[7] P. Sharma and N. K. Gupta, "Digital Marketing and SMEs: A Systematic Mapping Study," *Univ. Nebraska - Lincoln Digital Commons*, 2021, doi: 10.32873/unl.dc.libphilprac.9450.

[8] Universitas Gadjah Mada, "UGM Workshop Highlights Content Marketing and SEO Strategies to Boost MSME Sales," *UGM News*, 2024, doi: 10.13140/RG.2.2.17334.16969.

[9] D. A. Sari and L. P. Puspawati, "Digital Marketing Strategies in Expanding the Market for MSME Creative Products," *Front. Commun.*, 2025, doi: 10.3389/fcomm.2025.1647391.

[10] R. Setiawan, "Digital Marketing Strategy for Sustainable Performance of MSMEs," *RH J. Manage. Entrepreneurship*, 2025, doi: 10.5281/zenodo.13354478.

[11] A. Nurhadi, S. Rahmah, and I. Gunawan, "The Effect of E-Commerce Platforms, Digital Marketing, and Business Innovation on MSME Performance," *Int. J. Bus. Law Educ.*, vol. 4, no. 2, 2024, doi: 10.55529/ijble.492.

[12] P. Gupta and R. K. Singh, "The Role of Search Engine Optimization (SEO) Techniques in Small Business Growth," *Int. J. Online Marketing*, vol. 14, no. 1, pp. 45–62, 2023, doi: 10.4018/IJOM.326555.

[13] M. Rafiq, A. N. Fitri, and B. Irawan, "Search Engine Optimization and Social Media Engagement: Drivers of E-Commerce Visibility for MSMEs," *J. Digit. Bus. Manage.*, vol. 9, no. 2, pp. 34–49, 2024, doi: 10.1108/JDBM-02-2024-0054.

[14] A. R. Shokouhi and J. M. Allan, "Improving E-Commerce Search Accuracy Using Ensemble Regression and Gradient Boosting Models," *Inf. Process. Manage.*, vol. 61, no. 4, 2024, doi: 10.1016/j.ipm.2024.103526.

[15] H. Zhang, Y. Fang, and C. Liu, "Predictive Analytics for Online Sales Performance Using Machine Learning Algorithms," *Expert Syst. Appl.*, vol. 241, p. 122814, 2024, doi: 10.1016/j.eswa.2024.122814.

[16] L. T. Nguyen and J. Zhao, "Data-Driven Optimization in SME Digital Advertising: A Machine Learning Perspective," *J. Bus. Anal.*, vol. 7, no. 3, pp. 201–218, 2023, doi: 10.1080/2573234X.2023.2270162.

[17] K. N. Anwar and F. Ahmed, "Evaluating the Impact of Backlink Quality on SEO Ranking Performance," *Procedia Comput. Sci.*, vol. 217, pp. 556–566, 2023, doi: 10.1016/j.procs.2022.12.057.

[18] C. Lopez, R. Sanchez, and M. Ortega, "Social Media Engagement as a Predictor of Online Sales Growth in Small Businesses," *J. Retail. Consum. Serv.*, vol. 75, 2024, doi: 10.1016/j.jretconser.2023.103604.

[19] D. Lestari and R. Hidayat, "The Relationship Between Website Traffic and Conversion Rates in Indonesian MSMEs," *Asian J. Bus. Res.*, vol. 14, no. 2, pp. 88–102, 2024, doi: 10.14707/ajbr.2406.

[20] S. Devi, H. Wahyudi, and N. Rahman, "An Empirical Model of SEO Optimization and Sales Forecasting Using Random Forest Regression," *Procedia Comput. Sci.*, vol. 218, pp. 1103–1115, 2024, doi: 10.1016/j.procs.2022.12.190.

[21] B. Wong, F. Al-Khalil, and J. Ramirez, "Predictive Modeling of Online Marketing Efficiency Using Gradient Boosting Machines," *Decis. Support Syst.*, vol. 172, 2024, doi: 10.1016/j.dss.2023.114018.

[22] N. Hidayah and R. Pratiwi, "Digital Transformation and Data Analytics Adoption Among Indonesian MSMEs: Challenges and Opportunities," *Int. J. Entrep. Small Bus.*, vol. 53, no. 1, pp. 79–96, 2024, doi: 10.1504/IJESB.2024.130841.