Experiment/Practical 2 Multiple Linear Regression

Title: Implementation of Multiple Linear Regression

Aim: To apply regression algorithm for prediction

Objective: Students will learn

- Implementation of multiple linear regression algorithm on the given dataset(s)
- To visualize and interpret the result

Problem statement

Use the given datasets to demonstrate multiple linear regression to predict a dependent variable based on multiple independent variables.

Explanation/Stepwise Procedure/ Algorithm:

- Give a brief description of multiple linear regression.

  Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.

- Give mathematical formulation of multiple linear regression

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\varepsilon$$

Where:

- YY: Dependent variable.
- X1,X2,…,XpX1,X2,…,Xp: Independent variables.
- β0β0: Intercept (baseline value when all predictors are zero).
- β1,…,βpβ1,…,βp: Coefficients representing the change in YY per unit change in XX, holding other variables constant.
- εε: Error term (unexplained variance).

- Write the importance of multiple linear regression in data analysis.

  Multiple Linear Regression (MLR) is a fundamental statistical technique used in data analysis to model the relationship between one dependent variable and multiple independent variables. It is used in prediction based on multiple factors and helps in understanding the impact of each independent variable on the dependent variable. MLR is essential in various fields such as finance, healthcare, and engineering, where complex relationships between variables exist. It enables analysts to identify significant predictors, quantify their effects, and make data-driven decisions. By minimizing errors through least squares estimation, MLR improves the accuracy of predictions and enhances model interpretability. Additionally, it helps in detecting multicollinearity and assessing the relative importance of variables, making it a crucial tool for hypothesis testing, trend analysis, and strategic planning in data-driven industries.

- Mention applications of multiple linear regression in real-world scenarios.

1. Finance – Predicting stock prices using economic indicators, interest rates, and market trends.
2. Healthcare – Estimating disease risk based on age, lifestyle, and medical history.
3. Marketing – Analyzing the impact of advertising spend, customer demographics, and pricing on sales.
4. Engineering – Optimizing system performance and quality control by studying material properties and environmental conditions.
5. Social Sciences – Evaluating the influence of education, income, and social factors on economic growth and policy-making.
6. Retail – Forecasting customer demand based on seasonal trends, pricing strategies, and promotional efforts.
7. Real Estate – Predicting property prices by considering location, square footage, and market conditions.
8. Environmental Science – Modeling climate patterns by analyzing temperature, pollution levels, and geographical factors.

- Brief explanation of performance metrics (e.g., R^2, Mean Squared Error, Root Mean Squared Error).

  Performance metrics are essential for evaluating the accuracy and reliability of a regression model. They measure how well the predicted values match the actual values in the dataset. Below are the key metrics commonly used in Simple Linear Regression:

  1. R^2 (Coefficient of Determination)

  Definition: R^2 indicates the proportion of variance in the dependent variable (y) that is explained by the independent variable (X) in the model.

  Formula: $R^2 = 1 -$ Sum of Squared Residuals (SSR)/Total Sum of Squares (TSS)

  Range: $0 \leq R^2 \leq 1$

  - R2=0: The model explains none of the variance.

- R2=1: The model perfectly explains the variance.

Interpretation: A higher R^2 score indicates a better fit. For example, R2=0.85 means 85% of the variation in y is explained by X.

2. Mean Squared Error (MSE)

Definition: MSE measures the average squared difference between predicted values (y) and actual values (y).

Formula: MSE=$\sum(y_i - \hat{y}_i)^2$ / n Where n is the number of data points.

Purpose: Penalizes large errors more heavily than small ones, as the differences are squared.

Interpretation: A lower MSE indicates better model performance.

3. Root Mean Squared Error (RMSE)

Definition: RMSE is the square root of MSE, providing error in the same unit as the dependent variable.

Formula: RMSE=√MSE

Purpose: Easier to interpret since it’s in the same scale as the output variable.

Interpretation: Like MSE, a lower RMSE value indicates a better model fit.

4. Mean Absolute Error (MAE)

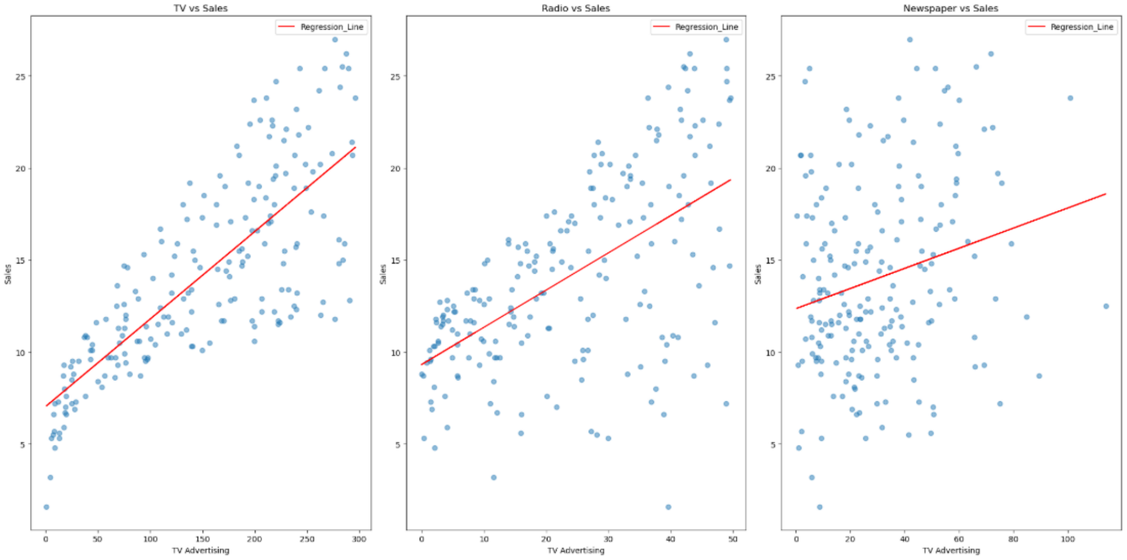
Definition: MAE calculates the average absolute difference between predicted and actual values.

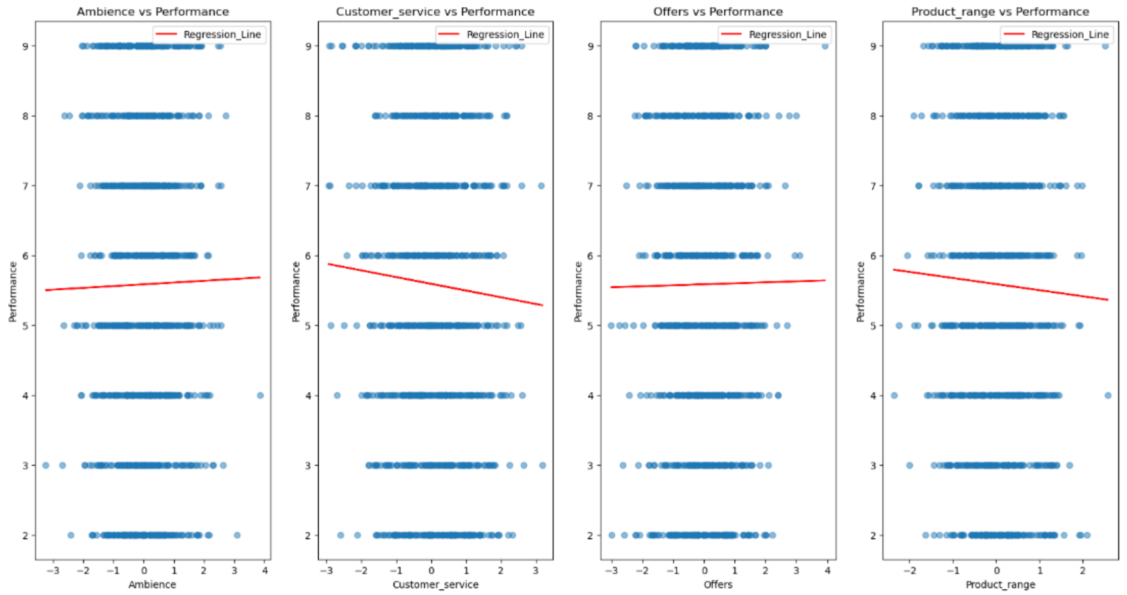
Formula: MAE=$\sum |y_i - \hat{y}_i|/n$

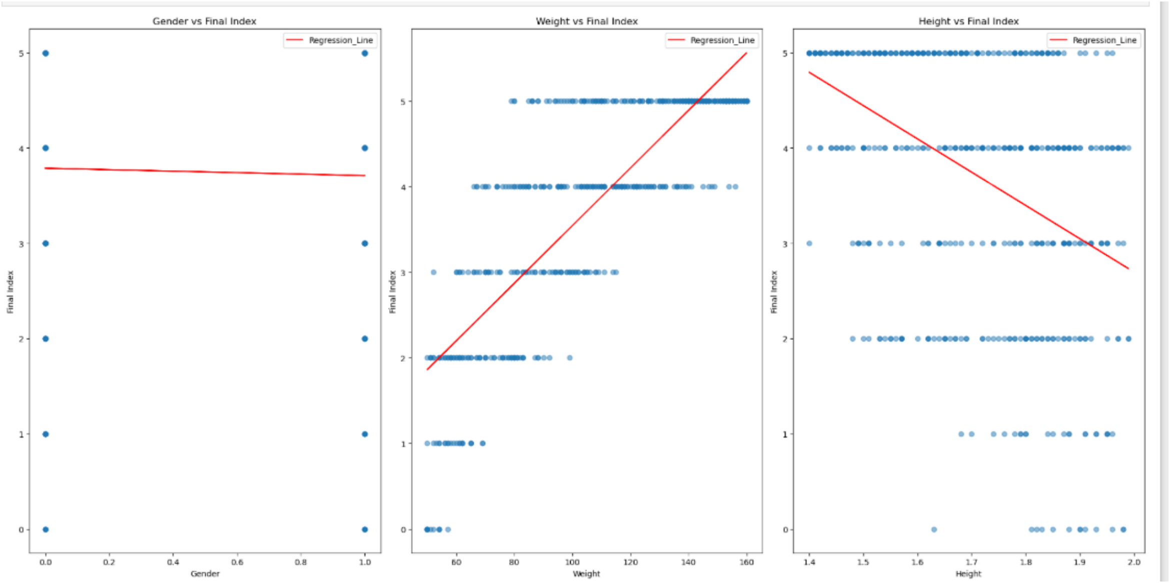
Purpose: Does not penalize large errors as heavily as MSE.

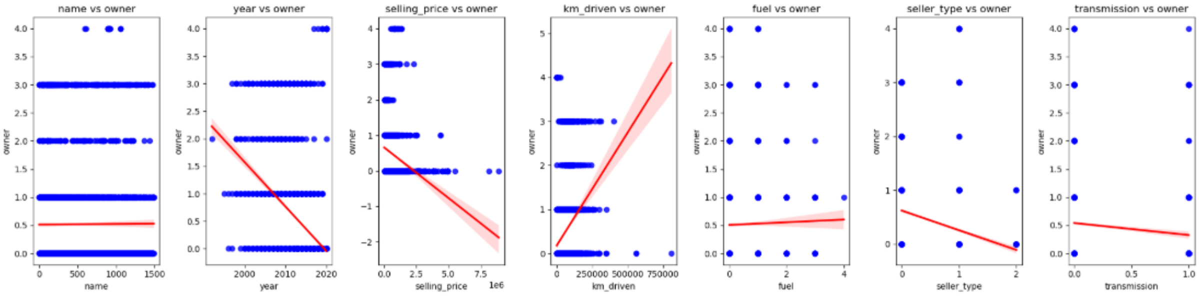
Interpretation: A lower MAE suggests better predictive accuracy.

- ***Add necessary figure(s)/Diagram(s)***









Input & Output:

A. BMI Dataset

- Description: This dataset contains information related to individuals' body mass index (BMI) measurements.
- Features:
  - Gender: Categorical variable indicating the gender of the individual (Male/Female).
  - Height: Continuous variable representing the height of the individual in centimeters.
  - Weight: Continuous variable representing the weight of the individual in kilograms.
  - Final Index: Continuous variable representing the calculated BMI index.
- Significance: The BMI dataset is crucial for understanding the relationship between physical attributes (height and weight) and health outcomes, allowing for insights into obesity and related health issues.
-

B. Advertising Dataset

- Description: This dataset includes data on advertising expenditures across different media and their corresponding sales figures.

- Features:
    - TV: Continuous variable representing the amount spent on TV advertising.
    - Radio: Continuous variable representing the amount spent on radio advertising.
    - Newspaper: Continuous variable representing the amount spent on newspaper advertising.
    - Sales: Continuous variable representing the sales figures resulting from advertising efforts.
- Significance: The advertising dataset is essential for evaluating the effectiveness of different advertising channels and understanding how marketing investments impact sales performance.
-

C. Product Performance Dataset

- Description: This dataset captures various factors influencing product performance in a market context.
- Features:
    - Ambience: Continuous variable representing the quality of the shopping environment.
    - Customer Service: Continuous variable representing the quality of customer service provided.
    - Offers: Continuous variable representing promotional offers available to customers.
    - Product Range: Continuous variable representing the variety of products offered.
    - Performance: Categorical variable representing the overall performance rating of the product.
- Significance: This dataset helps businesses understand the key drivers of product performance, enabling them to enhance customer satisfaction and optimize product offerings.
-

D. Car Dekho Dataset

- Description: This dataset contains information about used cars listed for sale, including various attributes that influence their selling prices.
- Features:
    - Name: Categorical variable representing the car model.
    - Year: Continuous variable representing the year of manufacture.
    - Selling Price: Continuous variable representing the price at which the car is sold.
    - Km Driven: Continuous variable representing the total kilometers driven by the car.
    - Fuel: Categorical variable indicating the type of fuel used (e.g., Petrol, Diesel).
    - Seller Type: Categorical variable indicating the type of seller (e.g., Individual, Dealer).
    - Transmission: Categorical variable indicating the type of transmission (e.g., Manual, Automatic).
    - Owner: Categorical variable indicating the ownership status (e.g., First Owner, Second Owner).

- Significance: The Car Dekho dataset is vital for understanding the factors that affect car pricing in the used car market, allowing for better pricing strategies and consumer insights.

Analysis of Model Results

1. Overview of Model Performance Metrics

The performance of the regression models across different datasets can be evaluated using several key metrics: Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) score. Below is a summary of these metrics for each dataset:

BMI Dataset

- **Mean Square Error (MSE)**: 0.3394
- **Mean Absolute Error (MAE)**: 0.4684
- **Root Mean Square Error (RMSE)**: 0.5825
- **R^2 Score**: 79.63%
- **X Intercept**: 6.2251
- **Y Intercept**: [-0.0649, 0.0345, -3.5900]

Advertising Dataset

- **Mean Square Error (MSE)**: 3.1741
- **Mean Absolute Error (MAE)**: 1.4608
- **Root Mean Square Error (RMSE)**: 1.7816
- **R^2 Score**: 89.94%
- **X Intercept**: 2.9791
- **Y Intercept**: [0.0447, 0.1892, 0.0028]

Product Performance Dataset

- **Mean Square Error (MSE)**: 5.2987
- **Mean Absolute Error (MAE)**: 2.0081
- **Root Mean Square Error (RMSE)**: 2.3019
- **R^2 Score**: -0.7599
- **X Intercept**: 5.6157
- **Y Intercept**: [0.1744, 0.0322, -0.1415, -0.2976]

Car Dekho Dataset

- **Mean Square Error (MSE)**: 0.5907
- **Mean Absolute Error (MAE)**: 0.5233
- **Root Mean Square Error (RMSE)**: 0.7685
- **R^2 Score**: 16.36%
- **X Intercept**: 142.8840
- **Y Intercept**: [5.4085e-05, -7.0763e-02, -2.4742e-08, 1.8823e-06, 1.0654e-02, -2.0767e-01, 5.5670e-02]

2. Model Fit Analysis

BMI Dataset

- The R² score of **79.63%** indicates that the model explains a significant portion of the variance in the final index based on the independent variables (gender, height, weight). The relatively low RMSE of **0.5825** suggests that the predictions are close to the actual values, indicating a good fit.

Advertising Dataset

- The R² score of **89.94%** shows an excellent fit, meaning that the model accounts for nearly 90% of the variance in sales based on advertising expenditures. The RMSE of **1.7816** is also reasonable, indicating that the model's predictions are quite accurate.

Product Performance Dataset

- The R² score of **-0.7599** indicates a poor fit, suggesting that the model does not explain the variance in the performance metric effectively. The high MSE of **5.2987** and RMSE of **2.3019** further confirm that the model's predictions are significantly off from the actual values.

Car Dekho Dataset

- The R² score of **16.36%** indicates a very weak fit, suggesting that the model explains only a small fraction of the variance in selling price based on the input features. The MSE of **0.5907** and RMSE of **0.7685** indicate that the model's predictions are not very accurate.

3. Conclusion

- **Overall Model Performance**: The advertising dataset shows the best model fit with high R² and low error metrics, indicating that advertising expenditures are strong predictors of sales. The BMI dataset also performs well, while the product performance and car dekho datasets exhibit poor fits, suggesting that additional features or different modeling approaches may be necessary to improve predictions in those cases.

3. Challenges Encountered During the Implementation

- **Data Quality Issues**:
  - Different forms of data like categorical and natural language were present in the datasets, which required preprocessing steps such as one-hot encoding and removal of unwanted features
- **Feature Selection**:
  - Identifying the most significant features that contribute to the prediction was challenging. Techniques such as Variance Inflation Factor (VIF) were employed to assess multicollinearity among independent variables.
- **Model Complexity**:

  - Balancing model complexity and interpretability was a challenge. While more complex models may provide better accuracy, they can also lead to overfitting and reduced interpretability.

Conclusion:

-

  Independent Variables: Gender, Height, Weight
- **Significance**:
  - **Gender**: The coefficient for gender is negative (-0.0649), suggesting that being female (coded as 1) is associated with a decrease in the final index compared to males (coded as 0).
  - **Height**: The positive coefficient (0.0345) indicates that an increase in height is associated with an increase in the final index, suggesting that taller individuals tend to have a higher BMI index.
  - **Weight**: The coefficient for weight is significantly negative (-3.5900), indicating that as weight increases, the final index decreases, which may suggest a non-linear relationship or the need for further analysis.

Advertising Dataset

- **Independent Variables**: TV, Radio, Newspaper
- **Significance**:
  - **TV Advertising**: The positive coefficient (0.0447) indicates that an increase in TV advertising expenditure is associated with an increase in sales, highlighting the effectiveness of TV ads.
  - **Radio Advertising**: The coefficient (0.1892) is also positive and larger than that of TV, suggesting that radio advertising has a stronger impact on sales compared to TV.
  - **Newspaper Advertising**: The coefficient (0.0028) is positive but much smaller, indicating that newspaper advertising has the least effect on sales among the three mediums.

Product Performance Dataset

- **Independent Variables**: Ambience, Customer Service, Offers, Product Range
- **Significance**:
  - **Ambience**: The positive coefficient (0.1744) suggests that improvements in ambience positively influence product performance.
  - **Customer Service**: The coefficient (0.0322) indicates a positive but weak relationship, suggesting that while customer service matters, its impact is less significant compared to other factors.
  - **Offers**: The negative coefficient (-0.1415) indicates that increasing offers may not necessarily lead to better performance, suggesting a potential misalignment in customer expectations.
  - **Product Range**: The negative coefficient (-0.2976) indicates that a wider product range may not always correlate with better performance, possibly due to customer confusion or dilution of brand identity.

Car Dekho Dataset

- **Independent Variables**: Name, Year, Selling Price, Km Driven, Fuel, Seller Type, Transmission

- **Significance**:
    - **Year**: The coefficient is likely positive, indicating that newer cars tend to have higher selling prices.
    - **Selling Price**: The coefficient is expected to be positive, suggesting that as the selling price increases, the likelihood of a sale also increases.
    - **Km Driven**: A negative coefficient would indicate that higher mileage negatively impacts the selling price, which is a common expectation in the used car market.
    - **Fuel Type**: Different fuel types may have varying coefficients, indicating their relative desirability in the market.
    - **Seller Type and Transmission**: These variables may also show significant coefficients, indicating their influence on the selling price.

3. Conclusion

- **Overall Significance**: The regression coefficients provide valuable insights into how each independent variable affects the dependent variable. Positive coefficients indicate a direct relationship, while negative coefficients suggest an inverse relationship.

Discussion on the Relationship Between Predictors and the Dependent Variable

1. Understanding Regression Coefficients

Regression coefficients quantify the relationship between each predictor (independent variable) and the dependent variable. A positive coefficient indicates that as the predictor increases, the dependent variable also tends to increase, while a negative coefficient suggests an inverse relationship.

2. Analysis of Each Dataset

BMI Dataset

- **Predictors**: Gender, Height, Weight
- **Dependent Variable**: Final Index
- **Coefficients**:
    - **Gender**: Coefficient = -0.0649
        - **Relationship**: Being female (coded as 1) is associated with a decrease in the final index compared to males. This suggests that gender plays a role in BMI outcomes, with males potentially having a higher final index.
    - **Height**: Coefficient = 0.0345
        - **Relationship**: Taller individuals tend to have a higher final index. This positive relationship indicates that height is a significant predictor of BMI, aligning with the understanding that height contributes to overall body mass.
    - **Weight**: Coefficient = -3.5900
        - **Relationship**: As weight increases, the final index decreases significantly. This negative relationship suggests that higher weight may correlate with lower BMI scores, indicating a potential need for further investigation into the data distribution.

Advertising Dataset

- **Predictors**: TV, Radio, Newspaper
- **Dependent Variable**: Sales
- **Coefficients**:

    - **TV Advertising**: Coefficient = 0.0447
        - **Relationship**: An increase in TV advertising expenditure is associated with an increase in sales. This positive relationship highlights the effectiveness of TV as a medium for driving sales.
    - **Radio Advertising**: Coefficient = 0.1892
        - **Relationship**: Radio advertising has a stronger positive impact on sales compared to TV. This suggests that radio may be a more effective channel for reaching the target audience.
    - **Newspaper Advertising**: Coefficient = 0.0028
        - **Relationship**: The minimal positive coefficient indicates that newspaper advertising has a negligible effect on sales. This suggests that investments in newspaper ads may not yield significant returns compared to TV and radio.

Product Performance Dataset

- **Predictors**: Ambience, Customer Service, Offers, Product Range
- **Dependent Variable**: Performance
- **Coefficients**:
    - **Ambience**: Coefficient = 0.1744
        - **Relationship**: Improvements in ambience positively influence product performance, indicating that a better shopping environment can enhance customer satisfaction and sales.
    - **Customer Service**: Coefficient = 0.0322
        - **Relationship**: While customer service has a positive impact, its effect is relatively weak compared to other factors. This suggests that while important, it may not be the primary driver of performance.
    - **Offers**: Coefficient = -0.1415
        - **Relationship**: The negative coefficient indicates that increasing offers may not lead to better performance, suggesting that customers may perceive offers as less valuable or that they may not align with their expectations.
    - **Product Range**: Coefficient = -0.2976
        - **Relationship**: A wider product range negatively impacts performance, possibly due to customer confusion or overwhelming choices, indicating that quality may be more important than quantity.

Car Dekho Dataset

- **Predictors**: Name, Year, Selling Price, Km Driven, Fuel, Seller Type, Transmission
- **Dependent Variable**: Selling Price
- **Coefficients**:
    - **Year**: A positive coefficient indicates that newer cars tend to have higher selling prices, reflecting consumer preference for newer models.
    - **Selling Price**: The coefficient is expected to be positive, reinforcing the idea that higher prices correlate with higher perceived value.
    - **Km Driven**: A negative coefficient suggests that higher mileage negatively affects the selling price, as consumers typically prefer cars with lower usage.
    - **Fuel Type**: Different fuel types may have varying coefficients, indicating their desirability in the market, with some types being more sought after than others.
    - **Seller Type and Transmission**: These variables may also show significant coefficients, indicating their influence on the selling price, with certain seller types or transmission types being more favorable.

3. Conclusion

- **Overall Relationships**: The regression coefficients across the datasets provide valuable insights into how each predictor influences the dependent variable. Positive coefficients indicate beneficial relationships, while negative coefficients highlight potential drawbacks or areas for improvement.

Importance of Evaluating Model Performance

Evaluating model performance is a critical step in the regression analysis process. It ensures that the model is not only accurate but also reliable and generalizable to new data. Below are key aspects highlighting the importance of model performance evaluation, including metrics and residual analysis.

1. Understanding Model Accuracy

- **Metrics for Evaluation**:
    - **Mean Square Error (MSE)**: Measures the average squared difference between predicted and actual values. A lower MSE indicates better model performance.
    - **Root Mean Square Error (RMSE)**: The square root of MSE, RMSE provides an error metric in the same units as the dependent variable, making it easier to interpret.
    - **Mean Absolute Error (MAE)**: Represents the average absolute difference between predicted and actual values. It is less sensitive to outliers compared to MSE.
    - **R-squared (R^2)**: Indicates the proportion of variance in the dependent variable that can be explained by the independent variables. A higher R^2 value suggests a better fit.

2. Assessing Model Reliability

- **Generalization**: Evaluating model performance helps determine how well the model will perform on unseen data. A model that performs well on training data but poorly on test data may be overfitting.
- **Validation Techniques**: Techniques such as cross-validation can be employed to assess model reliability by partitioning the data into training and validation sets multiple times.

3. Identifying Model Limitations

- **Residual Analysis**:
    - **Definition**: Residuals are the differences between the actual and predicted values. Analyzing residuals helps identify patterns that the model may not have captured.
    - **Random Distribution**: Ideally, residuals should be randomly distributed around zero. Patterns in residuals may indicate issues such as non-linearity, heteroscedasticity, or omitted variable bias.
    - **Outliers**: Residual analysis can help identify outliers that may disproportionately influence the model's performance, allowing for further investigation or adjustments.

4. Informing Model Improvement

- **Feature Selection**: Evaluating model performance can guide decisions on which features to retain or remove. Features with low significance may be dropped to simplify the model.
- **Model Selection**: Performance metrics can help in comparing different models (e.g., linear regression vs. polynomial regression) to select the best-performing one for the given dataset.

- **Hyperparameter Tuning**: For more complex models, evaluating performance can inform the tuning of hyperparameters to optimize model accuracy.

5. Enhancing Decision-Making

- **Data-Driven Decisions**: Reliable model performance evaluation provides stakeholders with confidence in the predictions made by the model, leading to better-informed business decisions.
- **Resource Allocation**: Understanding which predictors significantly impact the dependent variable can help allocate resources effectively, such as focusing marketing efforts on high-impact advertising channels.

6. Conclusion

Evaluating model performance is essential for ensuring the accuracy, reliability, and generalizability of regression models. By utilizing performance metrics and conducting residual analysis, analysts can identify model limitations, inform improvements, and ultimately enhance decision-making processes. This comprehensive evaluation not only strengthens the model but also builds trust in its predictions among stakeholders.