Medical Insurance Cost Prediction Using Linear Regression

Ayan Maity

December 18, 2024

Abstract

This study aims to predict medical insurance costs using linear regression based on key attributes such as age, gender, BMI, number of children, smoking habits, and region. With an R^2 score of 0.7174 and RMSE of 7205.8864, the model demonstrates decent predictive performance. Smoking status emerged as the most significant factor affecting charges. Future improvements include exploring nonlinear models to address residual non-normality.

Introduction

Accurate prediction of medical insurance costs is crucial for insurance companies to manage risks effectively and ensure fair pricing. This study uses a dataset comprising 1337 records with attributes such as age, gender, BMI, number of children, smoking habits, and region to build a linear regression model. The main objective is to identify key factors influencing costs and develop a reliable predictive model. Insights from this analysis can assist in improving pricing strategies and risk assessment.

Dataset

The dataset contains 1337 records with no missing values, covering the following attributes:

- **age:** Age of the primary beneficiary.
- **sex:** Insurance contractor's gender (Female or Male).
- **bmi:** Body mass index, providing an understanding of body weights that are relatively high or low relative to height. The index is an objective measure of body weight (kg/m^2) using the ratio of height to weight, ideally between 18.5 and 24.9.
- **children:** Number of children covered by health insurance / Number of dependents.
- **smoker:** Whether the person smokes or not.
- **region:** The beneficiary's residential area in the US, categorized into northeast, southeast, southwest, and northwest.

- **charges:** Individual medical costs billed by health insurance.

Multiple Linear Regression (MLR)

MLR is a powerful statistical tool for modeling the relationship between a dependent variable (Y) and multiple independent variables ($X_1, X_2, \ldots, X_k$). It extends simple linear regression by incorporating multiple predictors to explain variations in Y. The mathematical form of the MLR equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon, \tag{1}$$

where:

- Y: Dependent variable (response).
- β_0: Intercept, representing the expected value of Y when all predictors are zero.
- $\beta_1, \beta_2, \ldots, \beta_k$: Coefficients of independent variables, representing their contribution to Y.
- ϵ: Random error term accounting for variability not explained by predictors.

The goal of MLR is to find the optimal values of $\beta_0, \beta_1, \ldots, \beta_k$ using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals:

$$RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y_i} \right)^2, \tag{2}$$

where $\hat{Y_i}$ is the predicted value of Y_i.

Assumptions of the Gauss-Markov Theorem

The Gauss-Markov Theorem guarantees that the OLS estimators are the Best Linear Unbiased Estimators (BLUE) if the following assumptions hold:

Linearity of the Model

The relationship between the dependent variable Y and independent variables $X_1, X_2, \ldots, X_k$ is linear in the parameters. This implies that the expected value of Y can be expressed as:

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k. \tag{3}$$

Linearity can be visually inspected by plotting the residuals against the predicted values. Non-linear patterns suggest a violation.

No Perfect Multicollinearity

Independent variables should not be perfectly correlated. Perfect multicollinearity makes it impossible to estimate coefficients uniquely. Detecting multicollinearity involves calculating the Variance Inflation Factor (VIF):

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \tag{4}$$

where R_j^2 is the coefficient of determination when X_j is regressed on all other predictors. A VIF > 10 indicates problematic multicollinearity.

Homoskedasticity

The variance of the residuals (ϵ) should be constant across all levels of the predictors:

$$\text{Var}(\epsilon | X) = \sigma^2. \tag{5}$$

When this condition is violated (heteroskedasticity), it leads to inefficient estimates. Diagnostic tools like residual plots and tests like the Goldfeld-Quandt test can detect heteroskedasticity.

Independence of Errors

The residuals should be independent, with no autocorrelation:

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \text{for } i \neq j. \tag{6}$$

This assumption is crucial for time-series data, where autocorrelation is common. The Durbin-Watson test (discussed below) is used to assess this assumption.

Zero Mean of Residuals

The average of residuals should be zero:

$$\mathbb{E}(\epsilon) = 0. \tag{7}$$

This ensures unbiased coefficient estimates.

Exogeneity

The independent variables should not be correlated with the error term (ϵ). This condition is expressed as:

$$\text{Cov}(X_i, \epsilon) = 0 \quad \forall i. \tag{8}$$

Violations occur due to omitted variables, measurement errors, or simultaneity, leading to biased and inconsistent estimates.

Normality of Errors

For hypothesis testing and confidence intervals to be valid, residuals should follow a normal distribution. This is especially important for small sample sizes.

Key Statistical Tests

Jarque-Bera (JB) Test

The JB test checks whether residuals are normally distributed. It combines skewness (S) and kurtosis (K) into a single statistic:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right), \tag{9}$$

where n is the sample size.

- Null Hypothesis (H_0): Residuals are normally distributed ($S = 0$ and $K = 3$).
- Alternative Hypothesis (H_1): Residuals are not normally distributed.

A high JB statistic with a low p-value (< 0.05) leads to rejecting H_0, indicating non-normality. Residual normality is assessed via histograms or Q-Q plots alongside the JB test.

Durbin-Watson (DW) Test

The DW test evaluates autocorrelation in residuals. The test statistic is:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}, \tag{10}$$

where e_t are residuals.

- Null Hypothesis (H_0): No autocorrelation ($\rho = 0$).
- Alternative Hypothesis (H_1): Positive or negative autocorrelation ($\rho \neq 0$).

The DW statistic ranges from 0 to 4:

- $DW \approx 2$: No autocorrelation.
- $DW < 2$: Positive autocorrelation.
- $DW > 2$: Negative autocorrelation.

Critical values depend on the number of observations (n) and predictors (k). Values near 0 or 4 indicate strong autocorrelation, violating the independence assumption.

Implications of Assumption Violations

- **Linearity Violation:** The model may fail to capture relationships, leading to poor predictions. Non-linear transformations or advanced models (e.g., polynomial regression) may be needed.
- **Heteroskedasticity:** Coefficient estimates remain unbiased but are inefficient, with unreliable standard errors. Remedies include weighted least squares or robust standard errors.

- **Autocorrelation:** Coefficients are unbiased but inefficient. Time-series models like ARIMA or adding lagged variables can address this.
- **Non-Normality of Residuals:** It impacts hypothesis testing and confidence intervals. Large datasets (via the Central Limit Theorem) mitigate this issue.
- **Endogeneity:** It results in biased estimates, invalidating inferences. Instrumental variable regression or two-stage least squares (2SLS) can address endogeneity.

Exploratory Data Analysis (EDA)

Data Overview

Basic Statistics:

- Mean Age: 39.2 years
- Mean BMI: 30.66
- Average Charges: 13,279.12

Data Types: Numerical (age, bmi, children, charges), Categorical (gender, smoker, region).

Missing Values

Finding: No missing values were found in any columns.

Feature Relationships

Correlation Heatmap:

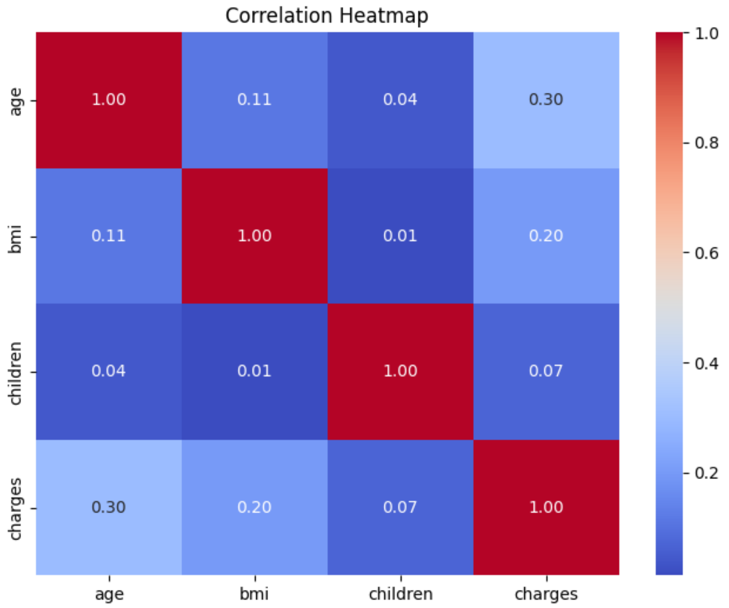
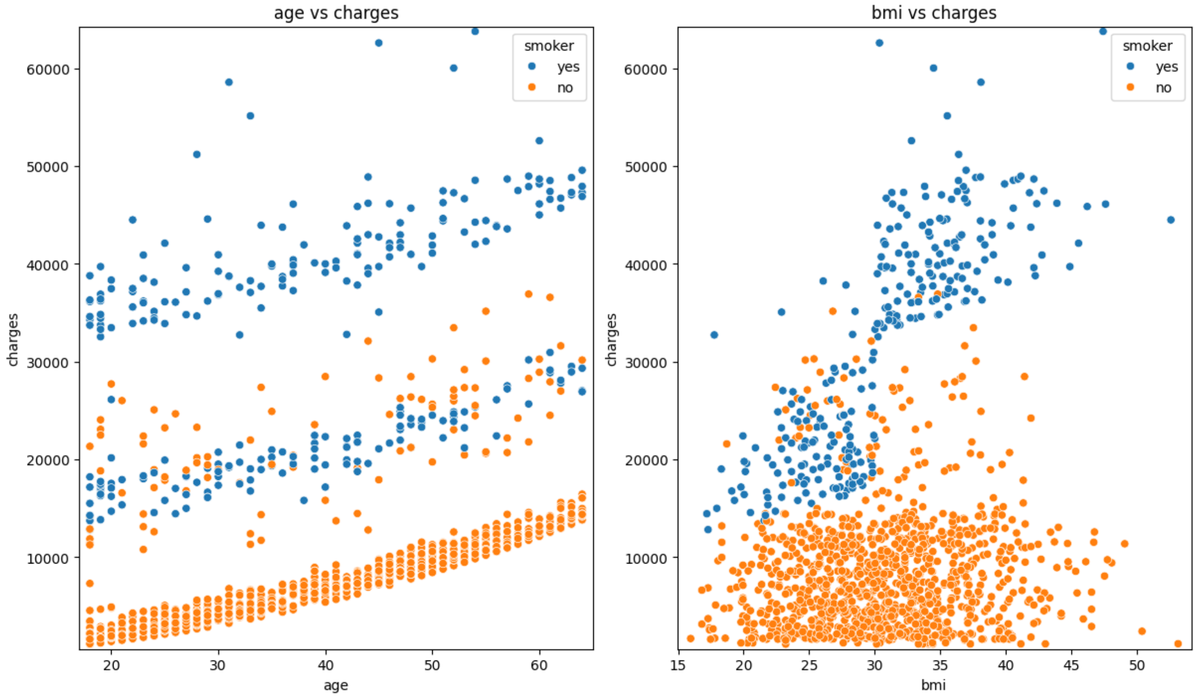


Figure 1: Correlation Heatmap

Age vs charges & BMI vs Charges:



Outlier Detection:

- Boxplots revealed outliers in bmi and charges.

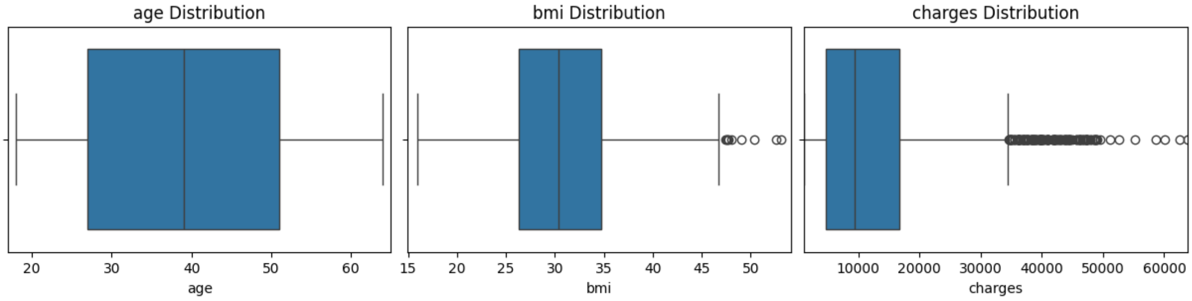


Figure 2: Boxplots

Distributions of Features

- **Charges**:
  - Original Skewness: 1.5137 (positively skewed).
  - Log-Transformation reduced skewness to -0.0895.

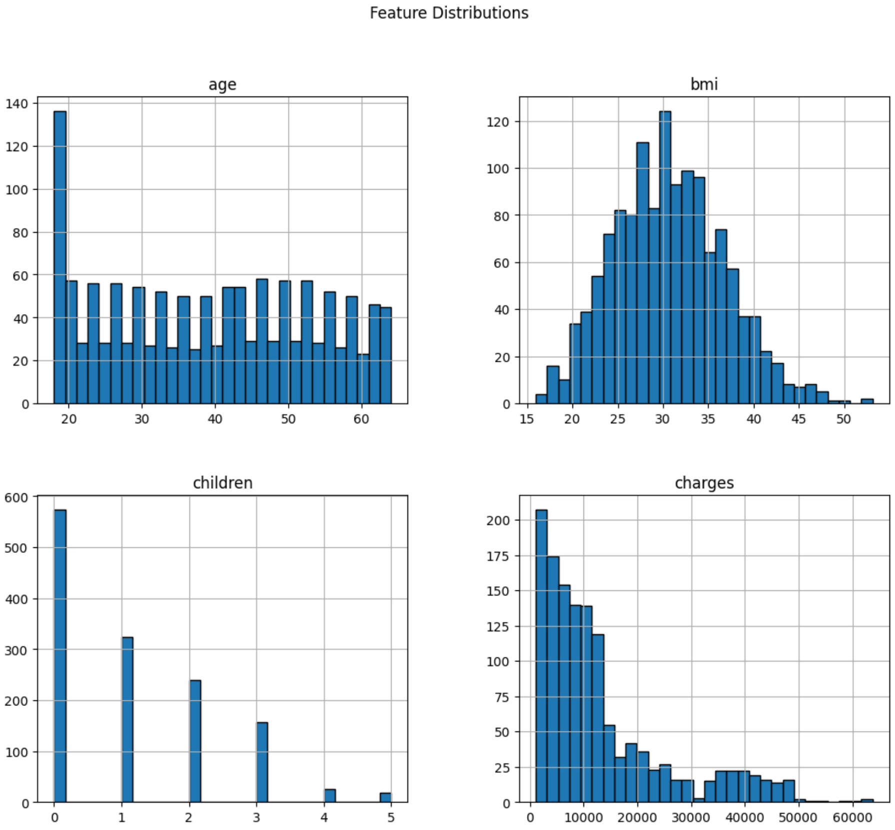


Figure 3: feature distribution

Data Preprocessing

Encoding Categorical Variables:

- gender: Female = 0, Male = 1.
- smoker: No = 0, Yes = 1.
- region: Ordinal encoding.

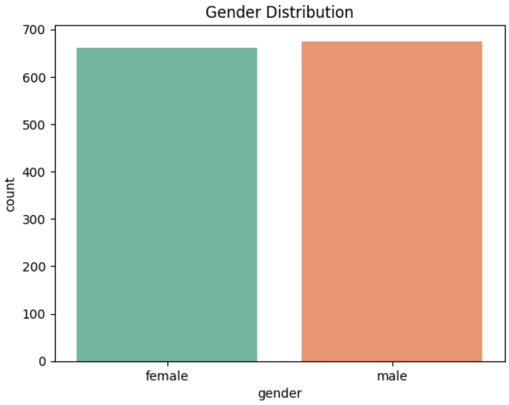


Figure 4: Gender distribution

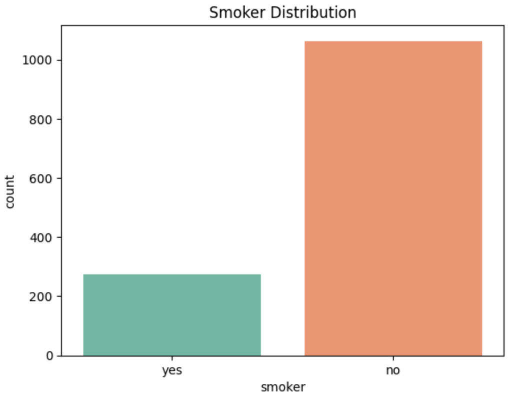


Figure 5: Smoker distribution

Log Transformation: Log-transformed charges to address skewness.

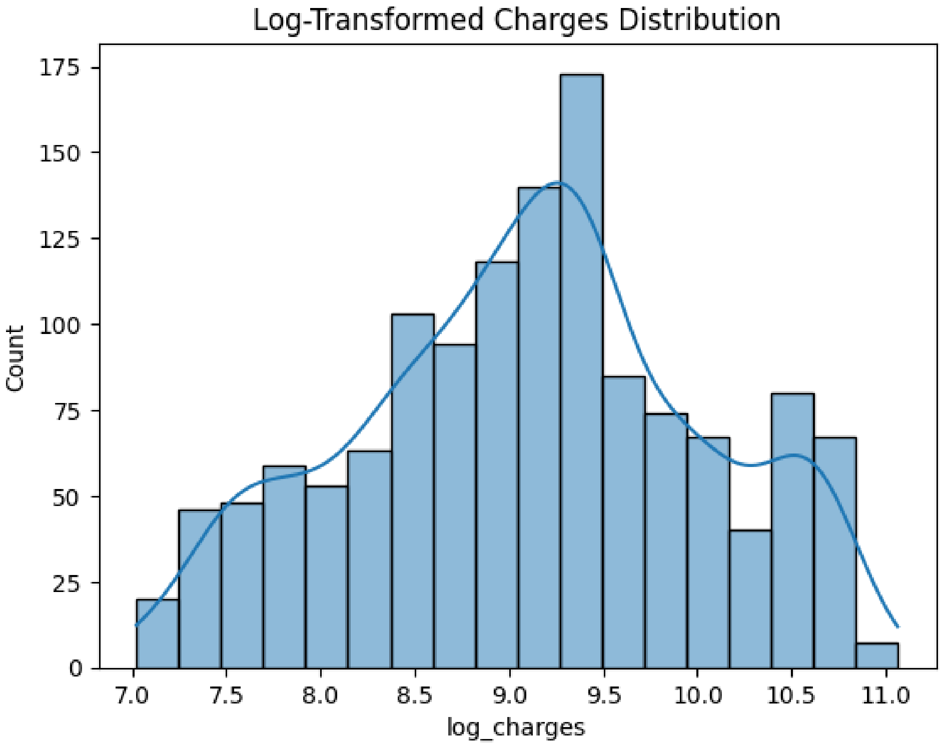


Figure 6: After log transformation the distribution of target variable

Model Development

Train-Test Split

Data split into 80% train (1069 records) and 20% test (268 records).

OLS Regression Model

- **Features Used**: age, gender, bmi, children, smoker, region.

- **Target**: Log-transformed charges.

Model Performance

Evaluation Metrics:

- RMSE on Test Set: 7205.8864
- R^2 Score on Test Set: 0.7174

Actual vs Predicted Charges:

- Scatter plot shows good alignment along the line $y = x$, indicating decent predictions.

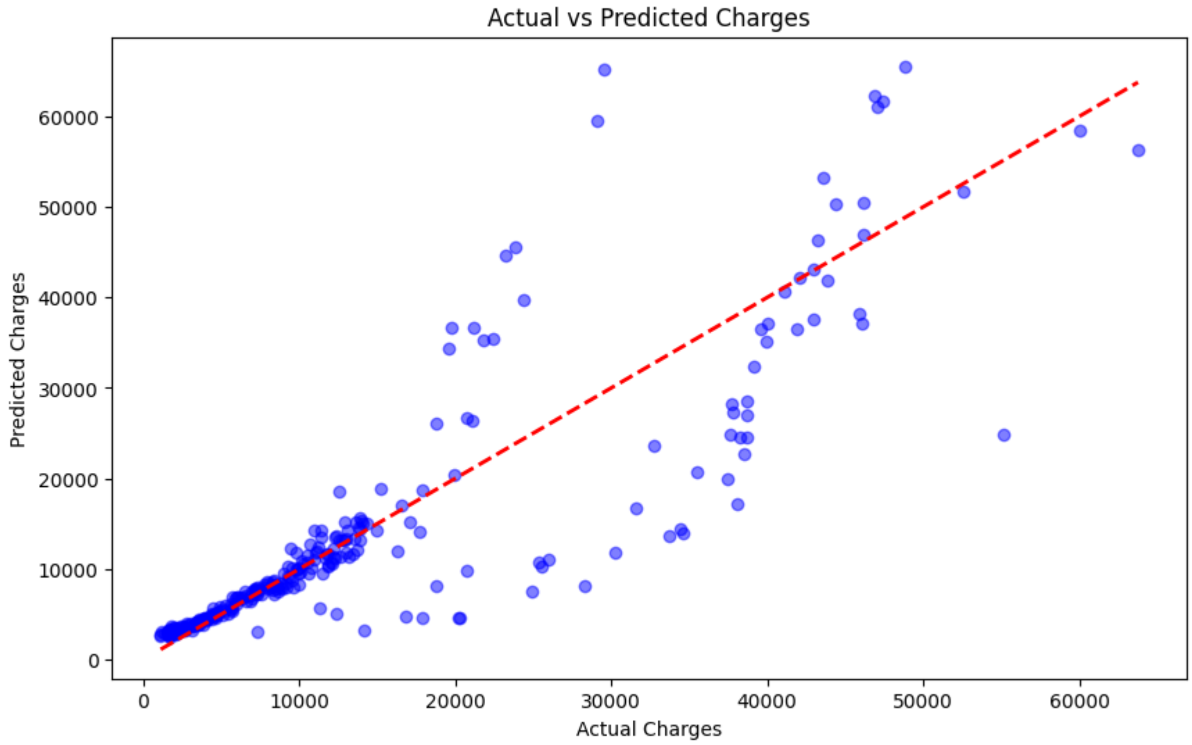


Figure 7: fitted line on Test data

Gauss-Markov Theorem

Mean of Residuals

- **Value**: $-1.168424052033944 \times 10^{-14}$
- **Interpretation**: The mean of the residuals is close to zero, which is a good sign. Ideally, the residuals should have a mean of zero, indicating no bias in the model's predictions.

Goldfeld-Quandt Test

- **Statistic**: 0.8935
- **P-value**: 0.9021
- **Null Hypothesis**: The data is homoscedastic (constant variance of the error terms).
- **Alternative Hypothesis**: The data is heteroscedastic (non-constant variance of the error terms).
- **Interpretation**: Since the p-value is very high (0.9021), we fail to reject the null hypothesis. This suggests that there is no evidence of heteroscedasticity in the model residuals, meaning the assumption of homoscedasticity holds.

Durbin-Watson Statistic

- **Statistic**: 1.8715
- **Interpretation**: The Durbin-Watson statistic tests for autocorrelation in the residuals. Values near 2 suggest no autocorrelation, while values closer to 0 or 4 suggest positive or negative autocorrelation, respectively. Since the statistic is around 1.87, it indicates a very slight presence of positive autocorrelation, but it's generally considered acceptable and does not indicate major issues.

Correlation Between Residuals and Independent Variables

- **Values**:
  - age: $-4.384779 \times 10^{-15}$
  - gender: 7.577059×10^{-16}
  - bmi: 1.233938×10^{-15}
  - children: 2.592880×10^{-15}
  - smoker: 1.237009×10^{-15}
  - region: 3.175586×10^{-15}
- **Interpretation**: The correlation values between the residuals and the independent variables are all extremely close to zero, suggesting there is no significant linear relationship between the residuals and the predictors. This is a good sign as it implies that the model has adequately accounted for the relationships between the predictors and the target variable.

Jarque-Bera Test

- **Statistic**: 1291.9008
- **P-value**: 2.93×10^{-281}
- **Null Hypothesis**: The residuals are normally distributed (skewness = 0 and kurtosis = 3).

- **Alternative Hypothesis**: The residuals are not normally distributed.
- **Interpretation**: The p-value is extremely small, indicating that we reject the null hypothesis. This suggests that the residuals are not normally distributed, and there might be skewness or excess kurtosis in the error terms. This could be a potential issue for certain statistical tests, although it does not necessarily invalidate the model.

Conclusion

The linear regression model demonstrated a strong ability to predict medical insurance charges, achieving an R^2 score of 0.7174. Among the factors considered, smoking status had the most significant impact on the charges, highlighting its importance in the prediction model. Although the model performed reasonably well, there is potential for improvement by exploring nonlinear models such as Random Forest or Gradient Boosting, which may better capture complex relationships. Additionally, the non-normality of the residuals could be addressed using advanced techniques, further enhancing the model's performance and reliability.