IV. METHODS

1. Linear Regression(LR)

Linear regression [1] is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

**Types of Linear Regression** [1]

There are two main types of linear regression:

* Simple Linear Regression

This is the simplest form of linear regression, because it has only one independent variable and one dependent variable. The formula for simple linear regression is:

y = β0 +β1X

*Where:*

*Y is the dependent variable*

*X is the independent variable*

*β0 is the intercept*

*β1 is the slope*

* Multiple Linear Regression

It used to model relationships between a dependent variable and two or more independent variables. It extends simple linear regression by considering multiple predicted factors to estimate the outcome. Its formula is:

y = β0 + β1X + β2X + … + βnXn

*Where:*

*Y is the dependent variable*

*X1, X2, …, Xn are the independent variables*

*β0 is the intercept*

*β1, β2, …, βn are the slopes*

However, we simply introduce Simple Linear Regression within the scopes of this study.

**Check Assumptions:**

**A diagram of a graph

Description automatically generated**

* 1. **Linearity:**

**Assessment Using Residual Plots:**

* If the residuals are randomly scattered around the horizontal line (zero line), this indicates that the linearity assumption is likely met.
* If there are patterns in the residuals, such as systematic curves or trends, it suggests a violation of linearity.

**Analysis of the Residual Plots:**

In the three residual plots (validation, test, and training sets), there appear to be patterns where residuals systematically deviate from the zero line, particularly in regions of higher and lower predicted values. This suggests that the relationship between the predictors and the response variable may not be perfectly linear, indicating a potential violation of the linearity assumption.

* 1. **Variance Uniformity**

**Variance of Residuals:**

In each of the residual plots (validation, test, and training sets), the spread of residuals does not seem uniform. For instance:

In the validation set, the variance appears to increase as the predicted values increase.

In the test set, there are sections where the spread of residuals widens and narrows inconsistently.

In the training set, the residuals also exhibit non-uniform variance, with higher spread in certain regions of predicted values.

**Possible Trends:**

Patterns of increasing or decreasing variance in residuals across the predicted values indicate heteroscedasticity, suggesting that the assumption of variance uniformity is violated.

**A graph of a graph

Description automatically generated with medium confidence**

* 1. **Normality**

**Analysis of Q-Q plots for each dataset:**

Training set: The residuals closely follow the theoretical line, indicating that the residuals approximately follow a normal distribution. However, there are slight deviations at the extreme ends (lower and upper quantiles), suggesting potential outliers.

Test set: Similar to the training set, the residuals mostly align with the theoretical line, with minor deviations at the tails. This indicates the residuals are approximately normally distributed.

Validation set: The Q-Q plot shows significant deviations from the theoretical line, particularly at the lower and upper ends. This indicates that the residuals in this dataset deviate from a normal distribution, especially at the extremes.

**Conclusion on normality assumption:**

Training and test sets: The residuals generally satisfy the normality assumption, with only minor deviations at the tails.

Validation set: The residuals violate the normality assumption, especially in the extreme quantiles, suggesting that the residuals in this set are not normally distributed.

* 1. **Independence**

A black screen with white text

Description automatically generated

**Understanding Durbin-Watson Statistic:**

* The **Durbin-Watson statistic** ranges from 0 to 4:
  + A value of **2** indicates no autocorrelation (perfect independence).
  + A value **less than 2** suggests **positive autocorrelation**.
  + A value **greater than 2** suggests **negative autocorrelation**.
  + Values close to **0** indicate **strong positive autocorrelation**, which violates the independence assumption.

**Analysis of the Values:**

1. **Training Set**:
   * Durbin-Watson statistic: 0.01750.0175
   * This value is extremely close to 00, indicating **strong positive autocorrelation** in the residuals of the training set. Independence is clearly violated.
2. **Test Set**:
   * Durbin-Watson statistic: 0.02770.0277
   * Similar to the training set, this value is still very close to 00, suggesting **strong positive autocorrelation** and a violation of independence in the test set.
3. **Validation Set**:
   * Durbin-Watson statistic: 0.01620.0162
   * This is also very close to 00, indicating **strong positive autocorrelation** and a clear violation of independence in the validation set.

**VI. Result and Recommadation:**

**Overall Assessment:**

The model assumptions are significantly violated in multiple areas, particularly in linearity, variance uniformity, and independence. While the normality assumption holds reasonably well for the training and test sets, the validation set shows deviations. These issues highlight the need for adjustments in the model, including potential non-linear transformations, addressing autocorrelation, and revising the error structure to improve overall model validity.

**Recommended Next Steps:**

Reassess the model's appropriateness, particularly for addressing non-linearity and heteroscedasticity.

Consider alternative models that can better handle autocorrelation, such as time-series models for temporal data or generalized least squares (GLS).

Apply transformations (e.g., logarithmic or polynomial) to stabilize variance and better capture the relationship between predictors and the response variable.

Validate any updated models against new datasets to ensure that assumptions are satisfied.