Multiple Linear Regression Variable Results for TLC Data

Executive Summary from Automatidata

Overview

The operations manager with New York City TLC is seeking more insight through regression modeling. The team's next milestone is to run a regression model for taxi fares based on variables in the dataset.

Objective

- Determine the correct modeling approach
- Build a regression model
- Finish checking model assumptions
- Evaluate the model

Results

- The mean_distance and mean_duration are the independent variables. The fare_amount is the dependent variable.
- Training set = 80% of total samples, Test set = 20% of total samples.
- The training data metrics are as follows:
  - Residual Sum of Squares (RSS): 325133.3316093109
  - Explained Variance Score (R^2): 0.8397085382230706
  - Mean Absolute Error (MAE): 2.1912213115414287
  - Mean Squared Error (MSE): 17.904803767239983
  - Root Mean Squared Error (RMSE): 4.231406830740809
- The test data metrics are as follows:
  - Residual Sum of Squares (RSS): 65183.08950669534
  - Explained Variance Score (R^2): 0.8679750234271767
  - Mean Absolute Error (MAE): 2.137045209651203
  - Mean Squared Error (MSE): 14.357508701915272
  - Root Mean Squared Error (RMSE): 3.7891303358310693

Next Steps

- Consider other variables that correlates with fare_amount for prediction.
- Implement any new variables into the multiple linear regression model.
- Adjust the training/test set proportions if necessary.