Regression Assumptions After Modeling

Executive summary report for the New York City Taxi and Limousine Commission
Prepared by **Automatidata**

ISSUE / PROBLEM

The New York City Taxi & Limousine Commission (TLC) partnered with Automatidata to create a regression model to predict taxi fares before a trip. At the client's request, the Automatidata data team developed and tested a regression model based on provided trip data.

RESPONSE

The Automatidata team chose to create a multiple linear regression (MLR) model based on the type and distribution of the data. The MLR model successfully predicted taxi cab fares before trips.

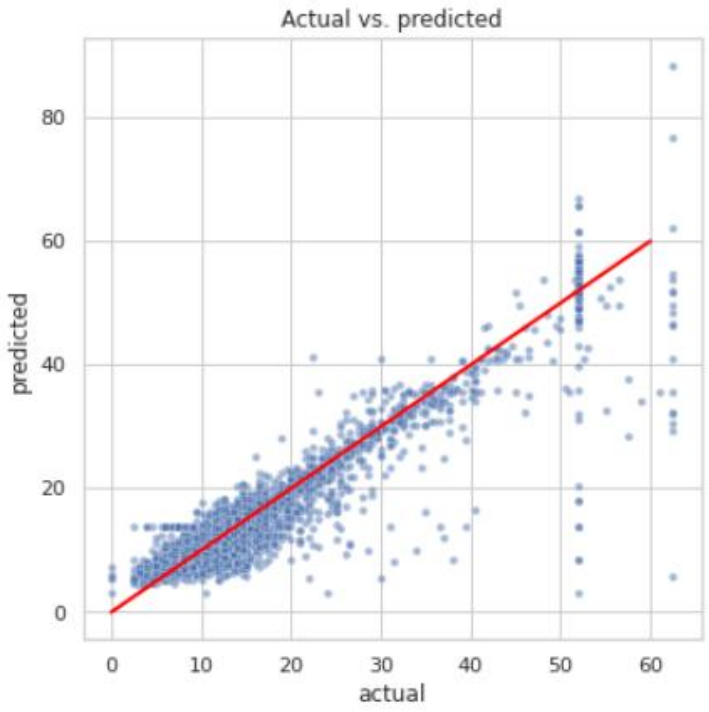
The model's performance was consistently high across both training and test sets, indicating that the model was neither over-biased nor overfit. It performed best on the test data.

IMPACT

Imputing outliers improved the model's performance, particularly for the variables fare amount and trip duration.

The linear regression model provides a reliable framework for predicting estimated taxi fares.

To demonstrate the model's effectiveness, the Automatidata team exported a scatter plot comparing predicted fare amounts to actual fares. This model can be used to predict taxi fares with reasonable confidence. Additional residual analysis is included in the provided notebook.



Alt-text: The scatter plot shows predicted versus actual taxi fare amounts based on the linear regression model.

Model Metrics:

- ✓ R^2 = 0.87 (86.8% of variance explained by the model)
- ✓ MAE = 2.1
- ✓ MSE = 14.36
- ✓ RMSE = 3.8

KEY INSIGHTS

- Trip duration had the greatest effect on fare amount, which is expected. The model revealed a mean increase of $7 for each additional minute. However, this should not be treated as a fixed benchmark because of strong correlations between certain features.
- Additional data should be requested for under-represented routes to further improve model accuracy.
- TLC can use these findings to create an app that allows riders to view estimated fares before their trips.
- The model provides a strong and reliable prediction for fare amounts and can support future modeling efforts.