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 hired Automatidata to predict taxi fares. In this phase of the project, Automatidata's data team delivered the regression model requested by their client.

RESPONSE

The Automatidata data team chose to develop a multiple linear regression (MLR) model based on the type and distribution of the provided data. The MLR model proved successful in estimating taxi fares prior to the ride.

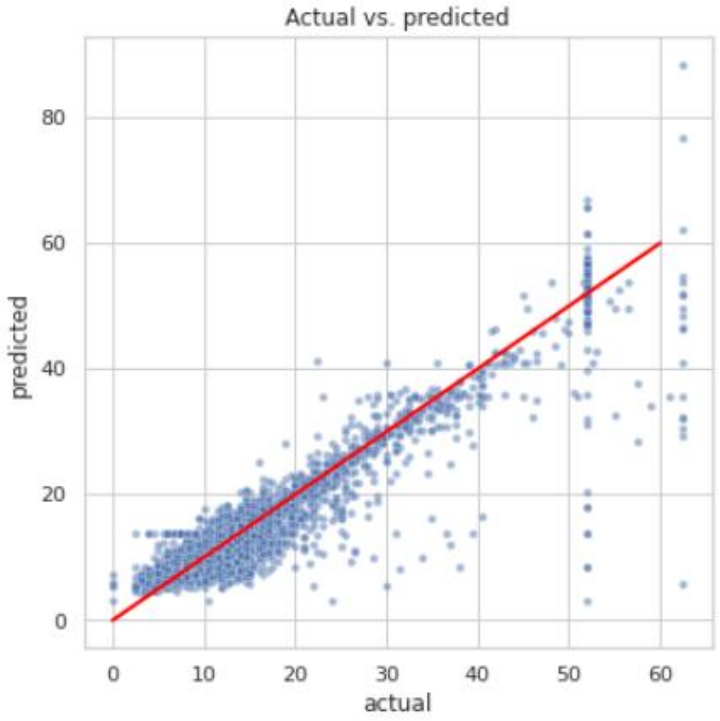
The model performed well on both the training and test sets, suggesting it is neither biased nor overfit. In fact, the model performed better on the test data.

IMPACT

Imputing outliers optimized the model, particularly regarding the variables fare amount and duration.

The linear regression model provides a robust framework for predicting the estimated fare amount for taxi rides.

To showcase the effectiveness of the linear regression model, the Automatidata data team included a scatter plot comparing the predicted and actual fare amounts. This model can be used to predict taxi fare amounts with reasonable confidence. The provided notebook offers further analysis of the model residuals.



Alt-text: The scatter plot shows a linear regression model plot illustrating predicted and actual fare amount for taxi cab rides.

Model metrics:

- Net model tuning resulted in:
  - ✓ R^2 0.87, meaning that 86.8% of the variance is described by the model.
  - ✓ MAE 2.1
  - ✓ MSE: 14.36
  - ✓ RMSE 3.8

KEY INSIGHTS

- The feature with the greatest impact on fare amount was ride duration, which was expected. The model showed an average increase of $7 for each additional minute; however, this figure is not a reliable benchmark due to high correlation between some features.
- Request additional data for under-represented itineraries.
- The New York City Taxi and Limousine Commission can leverage these findings to develop an app that allows users (TLC riders) to view the estimated fare before their ride begins.
- The model delivers a generally strong and reliable fare prediction that can be used in subsequent modeling efforts.