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 TLC collaborated with Automatidata to develop a model which could predict the fares before rides begin.

RESPONSE

The Automatidata team chose to develop the Multiple Linear Regression Model based on several variables. The model has been developed with great performance.

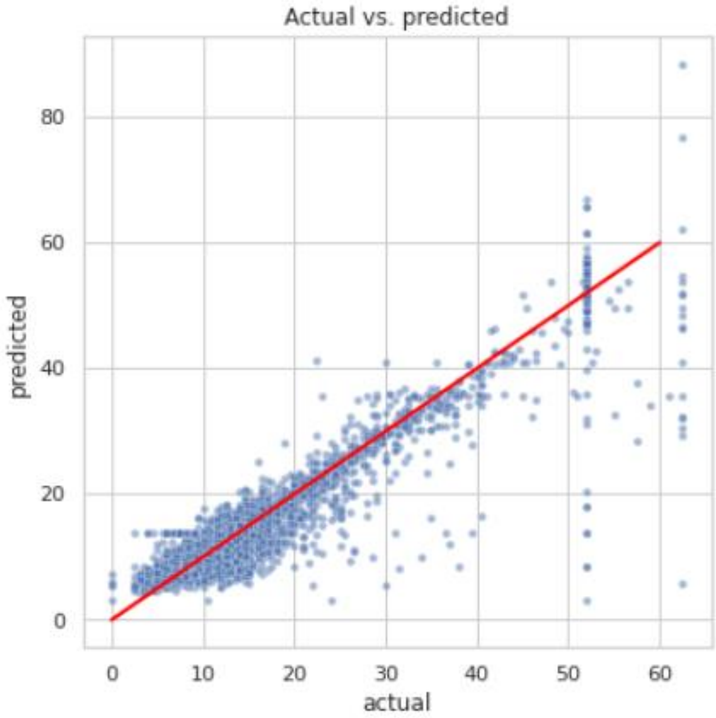
The model has not seen underfitting bias nor overfitting variance, this ensures that the model built is generalized and has higher performance on test data.

IMPACT

Imputed the outliers in variety of variables, especially fare amount and duration.

The model has also incorporated the prediction of rides passing by John F. Kennedy Airport.

To demonstrate the model's efficacy, Automatidata included a scatter plot comparing predicted and actual fare amounts. This model predicts taxi fare with confidence, and the provided notebook delves into residual analysis.



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 effect on fare amount was ride duration, which was not unexpected. The model revealed a mean increase of $7 for each additional minute, however, this is not a reliable benchmark due to high correlation between some features.
- Request additional data from under-represented itineraries.
- The New York City Taxi and Limousine commission can use these findings to create an app that allows users (TLC riders) to see the estimated fare before their ride begins.
- There were most of rides passing by JFK airport having almost same fare of $52. The model can also explain the prediction of rides passing by JFK airport.