Regression Assumptions After Modeling

Executive summary report for the New York City Taxi and Limousine Commission

ISSUE / PROBLEM

The New York City and Limousine Commission needs to predict taxi cab fares. In this part of the project, the deliverable, a regression model, was created based on the original ask of the client.
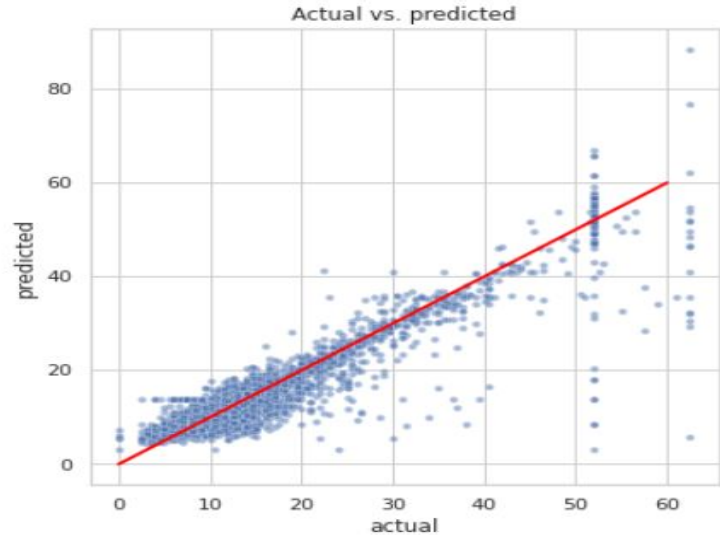
RESPONSE

A multiple linear regression model was created based on the type and distribution of data provided. The MLR model showed a successful model that estimates taxi cab fares prior to the ride.

IMPACT

Imputing outliers optimized the model, specifically in regards to the variables of: fare amount and duration.

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

To show efficacy of the linear regression model, below is a scatter plot comparing the predicted and actual fare amount. This model can be used to predict the fare amount of taxi cab rides with reasonable confidence. The provided notebook exhibits further analysis on 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 in the outcome variable is described by the model
  - MAE 2.1
  - MSE 14.46
  - RMSE 3.8

KEY INSIGHTS

- The feature with the greatest effect on fare amount was ride distance, which was not unexpected. The model revealed a mean increase of $7 for each 3.57 mile traveled; 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.
- The model provides a generally strong and reliable fare prediction that can be used in downstream modeling efforts.