Regression Assumptions After Modeling

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

ISSUE / PROBLEM

The New York City Taxi & Limousine Commission contracted data team to predict taxi cab fares. In this part of the project, the data team created the deliverable for the original ask from their client: a regression model.

RESPONSE

The data team chose to create a multiple linear regression (MLR) model 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.

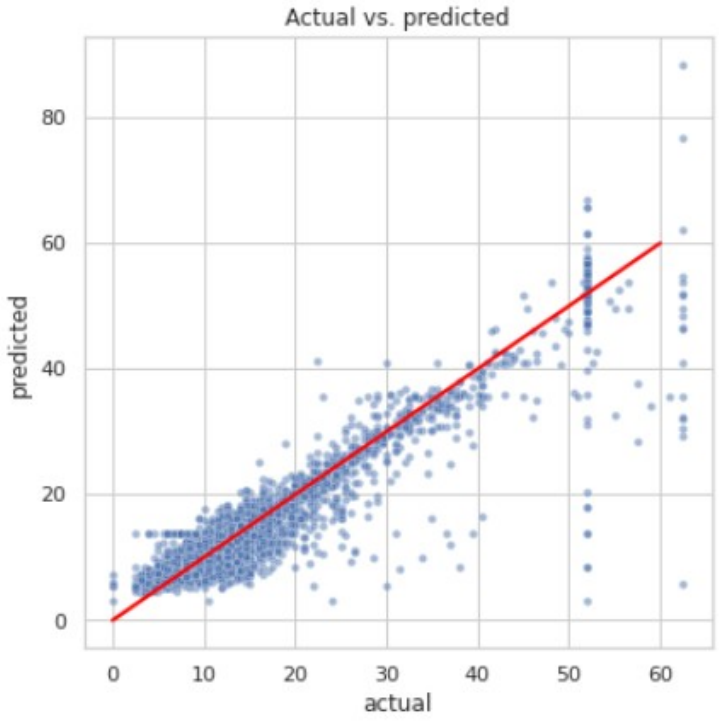
The model performance is high on the test set, suggesting that the model is not over-biased and that the model is not overfitted.

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.

In order to showcase the efficacy of the linear regression model, the data team included 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.



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 87.2% of the variance
  - ✓ is described by the model.
  - ✓ MAE 2.04
  - ✓ MSE: 13.43

KEY INSIGHTS

- The feature with the greatest effect on fare amount was ride duration, which was not unexpected.
- 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.