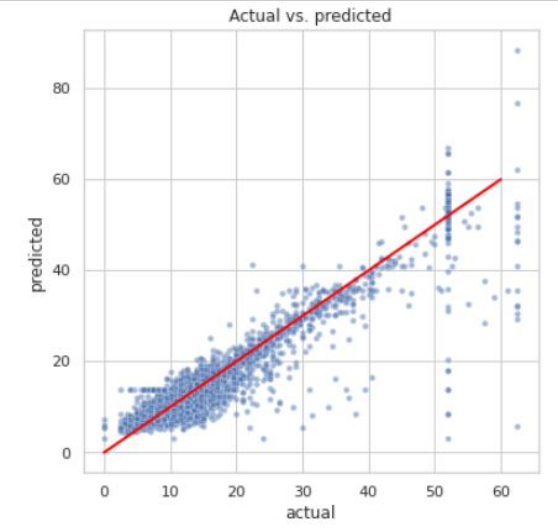
Regression Assumptions After Modeling

Executive Summary Report

Project Overview

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

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.
- The model provides a generally strong and reliable fare prediction that can be used in downstream modeling efforts.

Details

The Automatidata 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 both training and test sets, suggesting that the model is not over-biased and that the model is not overfit. The model performed better on the test data.
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.

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

Next Steps

- 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.
- Request additional data from under-represented itineraries.