Linear Regression Analysis for New York City TLC

Executive Summary Report

ISSUE / PROBLEM

The New York City Taxi & Limousine Commission (TLC) contracted data team to predict taxi cab fares.

RESPONSE

The team chose to create a multiple linear regression (MLR) model based on the type and distribution of data. The MLR model had successfully estimates taxi cab fares prior to ride.

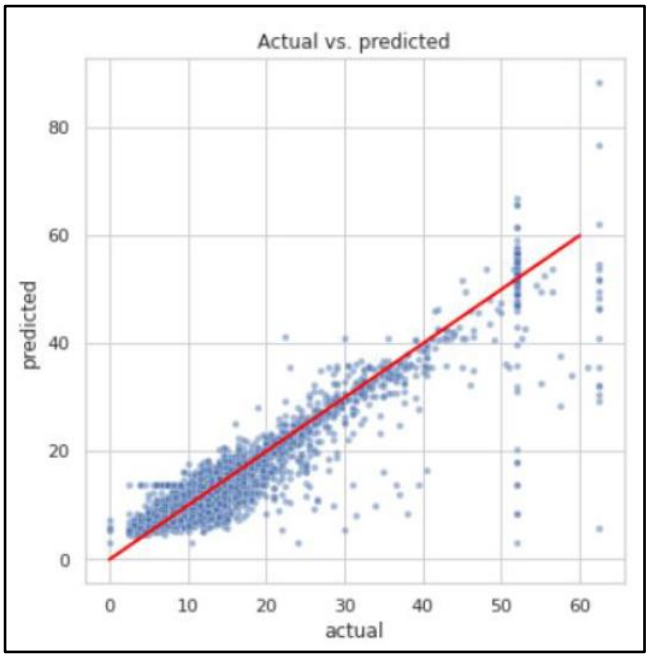
The model performance is high on both training and test sets, suggesting that the model is not over biased and that the model is not overfit. The model performed better on the test data.

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 scatterplot shows a linear regression model plot illustrating predicted vs 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.
- The model provides a generally strong and reliable fare prediction that can be used in downstream modeling efforts.