Title New York City TLC Linear Regression Model Project

Project Overview

This project's purpose is to develop a multiple linear regression model, through the request of New York City TLC to predict the taxi fare rides in advance. By taking some variables into account, we can determine the important effect each has to the fare rides, and evaluate which one(s) will be more effective, Along with that, we'll consider the model's results and performance to determine whether it is a good candidate for deploying.

Key Insights

- The multiple regression model yields a pretty well results. Specifically, here are the scores of essential metrics:

*** R-squared: 86.8%, which shows the proportion of variance in fare amount can be explained by the model (or predictor variables)**

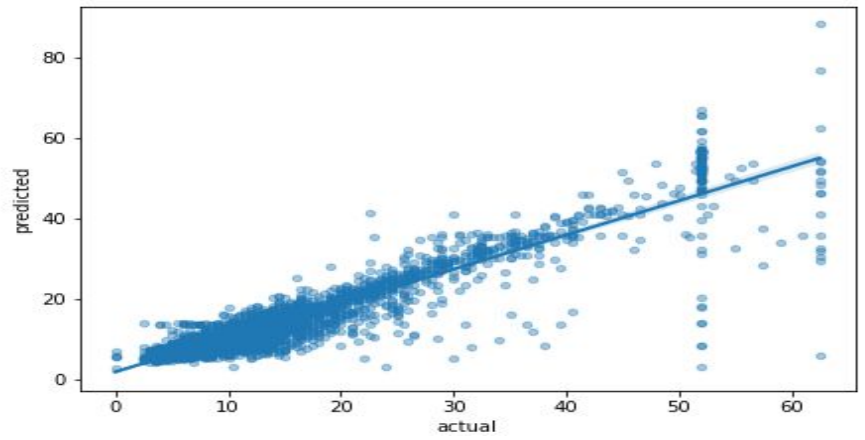
*** Mean squared error: 14.33**

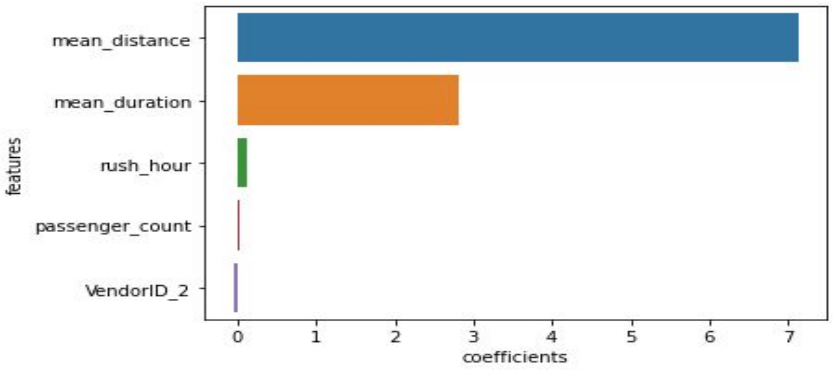
*** Root mean squared error: 3.78**

*** Mean absolute error: 2.13**

- Looking at the plot of actual values of fare amount vs. predict values from the model, we can easily see mainly it predicts pretty well and close to the actual data.
- Mean distance shows as a most influential variable to the fare amount of taxi rides. Particularly, if we increase one standard deviation of distance traveled (=3.57 miles), we expect the fare amount to increase on average of $7.13, while holding other variables constant.
- Besides, this model appears some concern to address, specifically about data leakage problem when developing the model.

Details





Next Steps

- Since this model achieves pretty well results, the New York City TLC can refer to it and decide to develop an app for predicting taxi fare rides.
- However, our Automatidata team would recommend to take into consideration for developing some advanced models. By considering that, we want to determine if other models can yield higher results and performance in prediction, along with comparing this multiple regression model.