Regression Modeling for Taxi Fares

Executive Summary Report for New York City Taxi and Limousine Commission

By: Rebecca Iglesias-Flores

ISSUE / PROBLEM

One of our clients, the New York City Taxi and Limousine Commission, has asked Automitadata to build a regression model using their dataset for predicting accurate taxi fares. In this part of the project, the Automidata Data Team has completed the deliverable requested.and findings will be discussed below.

RESPONSE

The Automitadata Data Team selected a Multiple Linear Regression (MLR) as the best way to predict taxi fares based on the types, distribution, and relationships of the variables.

Our definition of model success means accurately predicting fares on the test set on several evaluation metrics. The model performed well both in the training set and the testing set, giving us confidence that it will perform will in production.

IMPACT

During exploratory analysis, many outlier and incorrect data were found. The Automidata Data Team responded by handling the outliers and incorrect data appropriately.

This resulted in improved performance in the model.

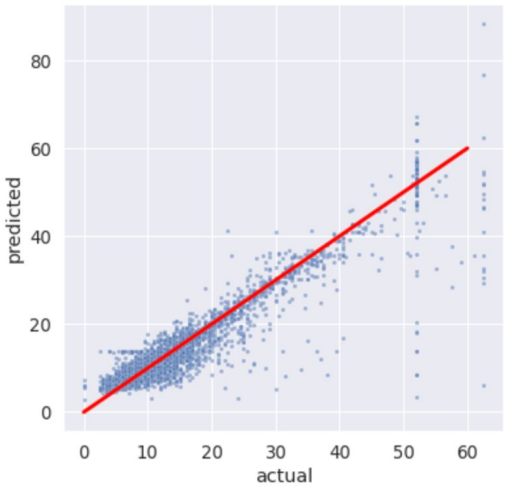
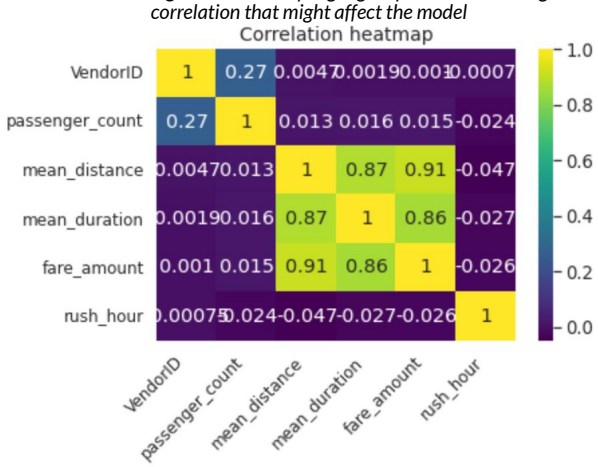


Figure 1. *Scatterplot shows predicted values vs. actual values for taxi ride fair amounts*

The linear relationship shown in the scatterplot to the left between the predicted values and the actual values around the best fit line suggests the MLR is an appropriate model for this dataset and making reliable predictions.

Figure 2. *Heatmap highlights predictors with high correlation that might affect the model*



No two predictor variables can be highly correlated with each other.

Figure 2 shows mean_distance vs. mean_duration with 87% correlation being potentially problematic.

Results. The feature with the highest weight in explaining the response is *mean_distance* with $7.13.
Interpretation. For every 3.57 miles, controlling for other variables, the fare mount increases by a mean of $7.13. For every 1 mile, controlling for other variables, the fare amount increases by a mean of $2.00

VendorID	passenger_count	mean_distance	mean_duration	rush_hour
-0.054611	0.031544	7.135758	2.811583	0.121491

KEY INSIGHTS

r-square	MAE	MSE	RMSE
0.868247	2.133658	14.327692	3.785194

- The R-square value indicates that 86% of the variance we see in taxi fare amount can be explained by the predictors we chose
- The model performance is extremely high, indicating there might be an issue with data leakage.
- More iteration and evaluation is needed on the model to ensure no data leakage is taking place. Strategies were applied in this iteration to handle outliers and missing data that might have falsely contributed to the performance gains.
- The dataset includes a VendorID = 2, which is JFK International Airport. We do not need to predict the values, they can be imputed imputed for computational efficiency and accuracy.