New York City TLC Project – Regression Analysis

Executive Summary Report IV

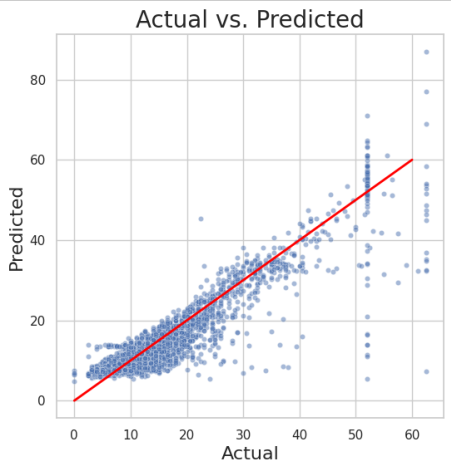
Commission Prepared by **Automatidata**

Project Overview

The New York City Taxi & Limousine Commission enlisted Automatidata to develop a model that can predict taxi cab fares. As part of this project phase, the team delivered the requested regression model. A multiple linear regression (MLR) approach was selected based on the nature and distribution of the provided data.

Key Insights

- The resulting MLR model accurately estimates taxi fares before the journey begins with strong performance on both training and testing datasets.
- Ride duration is heavily correlated with ride distance which is expected.
- Ride distance emerged as the most influential factor in determining fare, which aligns with expectations. The analysis suggests an average fare increase of about $9.5 for each additional mile of ride distance.
- When we isolate the trip duration from the distance, we observe an average fare increase of about $9 for each additional minute of ride time.
- Approximately 85.1% of the variation in fare amounts is explained by the model.



Details

To highlight the effectiveness of the linear regression model, a scatter plot is included showing the comparison between predicted and actual fares. The model is capable of predicting taxi fare amounts with a reasonable degree of confidence. Model metrics include:

- R^2 : 0.85
- MAE (Mean Absolute Error): 2.45
- MSE (Mean Squared Error): 16.15
- RMSE (Root Mean Squared Error): 4.02

The model is balanced (neither overly biased nor overfit) with superior performance on the testing data. The model's optimization involved handling outliers, especially those related to fare amount and ride duration, which improved its accuracy. The linear regression framework serves as a reliable method for estimating taxi fare amounts.

Next Steps

1. Develop an application that allows riders to view an estimated fare before starting their journey.
2. Collect additional data from routes that are currently underrepresented to further refine the model.