Multiple Linear Regression Model

For New York TLC dataset

Prepared by **Automatidata**

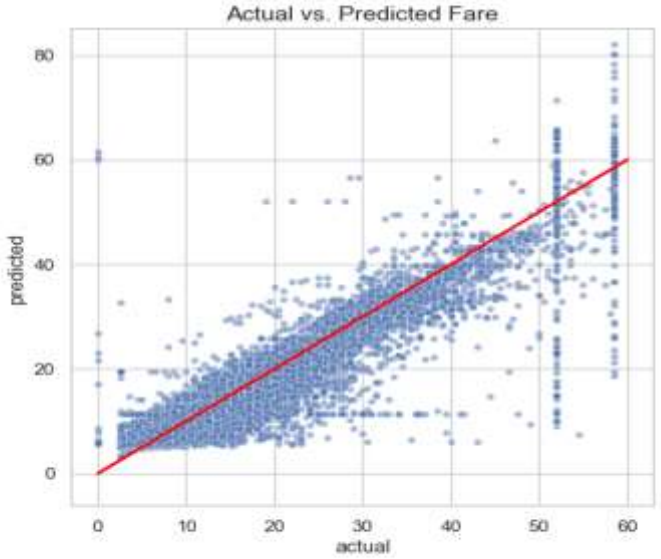
The New York City Taxi & Limousine Commission contracted Automatidata to predict taxi cab fares.The project proposal was approved, data was acquired, cleaned, explored and hypotheses were formed and tested. Now it's time for the ultimate project objective: To build a machine learning model to predict taxi fares before a ride is started.

Key Insights

- The fare amount is highly dependent on mean distance and mean duration of a particular source and destination pair.
- For every mile travelled, the fare amount increases by a mean of $2.
- This however is not a reliable benchmark due to high correlation between some features. .
- The errors are on the lower side which is a good scenario.
- We got R^2 score of 0.869 meaning that 86.9% of the variance in the Fare Amount variable is described by the model.
- The model can be improved by adding additional features. Like this data was just of January month. So other months' data is needed.
- The *rush_hour* feature too can be improved by splitting in intervals
- Historical traffic data and holiday data too can significantly improve the model.

Details

The scatter plot shows the actual vs predicted fare from the linear regression model illustrating the success of the model.



Model Performance

R2 Score:	0.89
Mean Absolute Error:	1.94
Mean Squared Error:	10.87
Root Mean Squared Error:	3.30

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 as mentioned in insights.
- Additional predictions can be made from the same data like total trip amount, tax, tips from rider, expected time of the trip, etc.