Exploratory Data Analysis of New York City TLC Data

Executive summary report

Commission Prepared by **Automatidata**

Project Overview

A Comissão de Táxi e Limusine da cidade de Nova York firmou uma parceria com a Automatidata para o desenvolvimento de um modelo de regressão capaz de prever o valor das corridas de táxi. Nesta etapa do projeto, é essencial realizar uma análise aprofundada dos dados, explorá-los, tratá-los e organizá-los adequadamente antes de avançar para a fase de modelagem.

Key Insights

The Problem: Following initial exploratory data analysis (EDA) on a sample provided by the New York City Taxi and Limousine Commission (TLC), it became evident that certain entries pose challenges to accurate fare prediction. Notably, there are rides with a recorded total cost but a trip distance of “0.” These cases have been flagged as anomalies or outliers and, based on our current assessment, should either be addressed within the algorithm or removed from the dataset entirely.

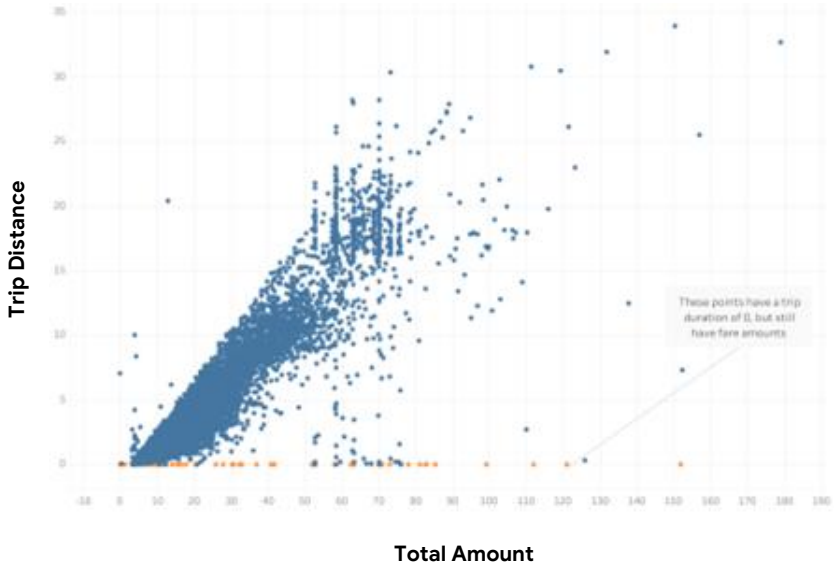
Proposed solution: Based on the conducted analysis, we recommend removing outlier entries with a recorded total distance of zero. These data points do not represent actual trips and could negatively impact the predictive accuracy of the model.

Keys to success

- **Sample validation:** Verify with TLC that the provided data accurately reflects the full dataset.
- **Handling additional outliers:** Develop a strategy for cases such as low-distance rides with high fares.

Details

Following an exploratory data analysis, the Automatidata team identified trip distance and total fare as key variables to represent a taxi ride. A scatter plot was created using Tableau to visualize the correlation between these variables, providing a more insightful and refined view of the data.



Alt Text: Graph displaying New York City TLC data plotting variables for total distance and total amount.

Next Steps

- **Outlier detection:** Identify unusual data points that may hinder fare prediction, such as locations with extended trip durations.
- **Impact variable analysis:** Determine which features most strongly influence ride fare amounts..
- **Variable selection:** Narrow down to the most relevant attributes for regression modeling, statistical analysis, and parameter tuning.