Executive Summary of Exploratory Data Analysis for New York TLC Data

Commission Prepared by Automatidata

Project Overview

The NYC Taxi & Limousine Commission has consulted with Automatidata to build a regression model to predict taxi cab fares. The Automatidata team performed exploratory data analysis to reveal data trends and perform data wrangling to prepare for model development.

Details

Automatitidata has decided that trip distance and total amount are the features most significant to predicting the target. The following Tableau Public visualization depicts the relationship between these variables.

Key Insights

Data Issues:
- There are values for features that directly impact the target that are inaccurate
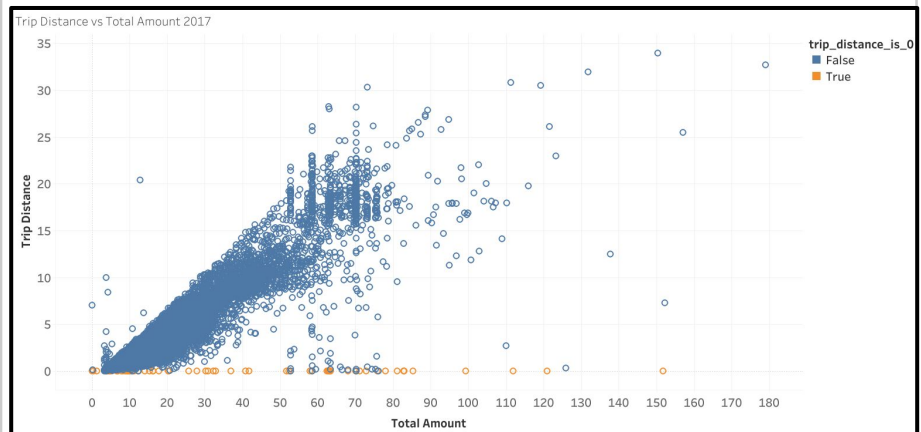- There are trips with a duration of 0 miles but with a positive total amount paid

Solutions:
- Remove duration outliers from the dataset prior to training

Future Steps:
- Remove outliers from other significant features
- Robust data cleaning for the whole dataset

Next Steps

- Remove other outliers from significant features that impact the target
- Find features that have the largest importance or correlation with the target
- Remove insignificant features from the training set