NEW YORK CITY TLC PROJECT

EXECUTIVE SUMMARY REPORT

COMMISSION PREPARED BY AUTOMATIDATA

Preliminary Data Summary

Executive summary report

Commission Prepared by Automatidata

ISSUE/ PROBLEM

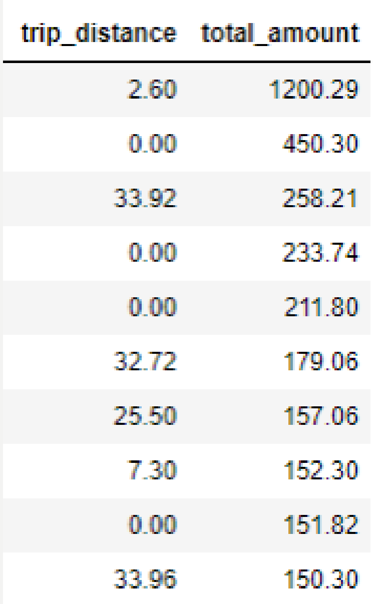
- The Automatidata data team conducted an initial examination of the dataset provided by the NYC Taxi & Limousine Commission to understand key variable descriptions and assess the data's suitability for generating insightful predictions in the regression model for taxi cab fares.

RESPONSE

- Explored dataset for anomalies
- Prioritized total_amount and trip_distance for predictive modeling, analyzed their interactions
- Identified relevant data components
- Laid the foundation for future exploratory analysis, visualizations, and modeling.

KEY INSIGHTS

- The dataset contains variables conducive to building predictive models for taxi cab ride fares.
- Unusual values include short-distance trips with disproportionately high charges, evident in the total_amount variable. Reference screenshots for illustration.



trip_distance	total_amount
2.60	1200.29
0.00	450.30
33.92	258.21
0.00	233.74
0.00	211.80
32.72	179.06
25.50	157.06
7.30	152.30
0.00	151.82
33.96	150.30

NEXT STPES

- Execute a thorough exploratory data analysis.
- Implement data cleaning and analysis procedures to identify unusual variables like outliers.
- Utilize descriptive statistics to gain insights into the data.
- Develop and execute a regression model.

Exploratory Data Analysis

Executive summary report

Commission Prepared by Automatidata

OVERVIEW

Initial exploratory data analysis (EDA) on a sample of the New York City TLC data reveals obstacles to accurate ride fare prediction, specifically trips with a total cost entered but a total distance of "0." These anomalies or outliers need to be addressed in the algorithm or considered for removal.

PROJECT STATUS

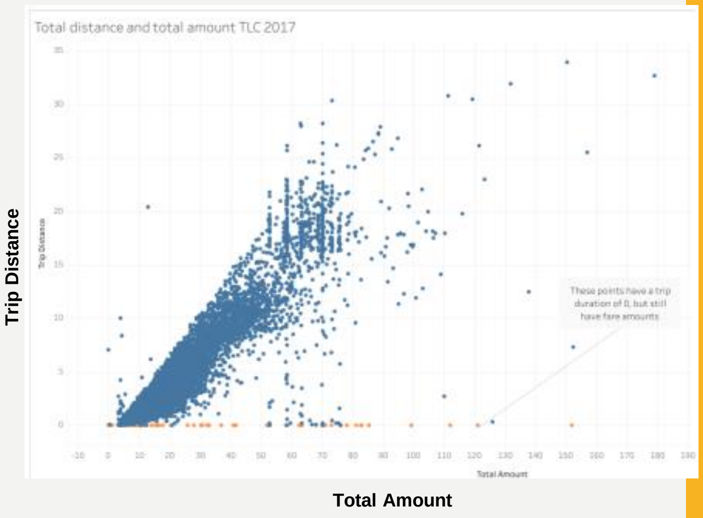
After analysis, we recommend removing outliers with a total distanced recorded of 0.

NEXT STEPS

- Identify atypical data points that may hinder accurate prediction of trip fares, such as locations with unusually long durations.
- Identify the variables with the greatest influence on trip fares.
- Narrow down the dataset to focus on the most significant variables for regression, statistical analysis, and parameter tuning.

KEY INSIGHTS

Based on the exploratory data analysis, the Automatidata data team identified trip distance and total amount as crucial variables depicting a taxi cab ride. The provided scatter plot illustrates the relationship between these variables, enhancing visualization through Tableau.



Statistical Review and A/B Testing

Executive summary report

Commission Prepared by Automatidata

OVERVIEW

The objective of this project is to predict taxi cab fares preemptively. Currently, the focus is on devising methods to increase revenue for New York City taxi cab drivers. This phase of the project investigates the correlation between total fare amount and payment type.

PROBLEM

In exploring the relationship between total fare amount and payment type, this project aims to ascertain whether customers paying by credit card tend to contribute larger total fare amounts compared to those paying in cash, considering the variable nature of tips received by taxi cab drivers.

SOLUTION

Automatidata conducted an A/B test to assess the connection between credit card payments and total fare amounts. The primary finding suggests that incentivizing customers to use credit cards could boost revenue for taxi drivers.

DETAILS

Steps conducted in the A/B test

- Collected sample data from an experiment in which customers are randomly selected and divided into two groups:
  - Customers who are required to pay with credit card.
  - Customers who are required to pay with cash. This enables us to draw causal conclusions about how payment method affects fare amount.
- Computed descriptive statistics to better understand the average total fare amount for each payment method available to the customer.
- Conducted a two-sample t-test to determine if there is a statistically significant difference in average total fare between customers who use credit cards and customers who use cash.

A/B test result

There is a statistically significant difference in the average total fare between customers who use credit cards and customers who use cash. Customers who used credit cards showed a higher total amount compared to cash.

NEXT STPES

The Automatidata data team suggests that the New York City TLC promotes credit card payments and devises strategies to encourage their use. For instance, installing signs stating "Credit card payments are preferred" in cabs and implementing a protocol where cab drivers verbally inform customers about the preference for credit card payments could be effective measures.

Regression Analysis

Executive summary report

Commission Prepared by Automatidata

OVERVIEW

The New York City Taxi & Limousine Commission contracted Automatidata to predict taxi cab fares. In this part of the project, the Automatidata data team created the deliverable for the original ask from their client: a regression model.

PROJECT STATUS

The Automatidata data team opted for developing a multiple linear regression (MLR) model, leveraging the characteristics and dispersion of the available data. This MLR model effectively predicts taxi cab fares before the commencement of a ride.

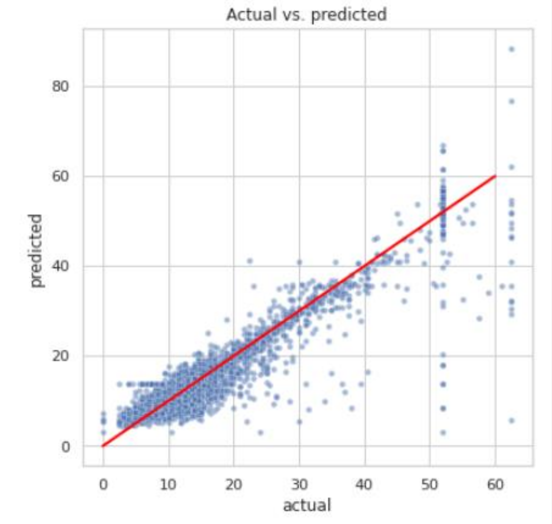
Demonstrating robustness, the model exhibits notable performance across both training and test datasets, indicating absence of over-bias and overfitting. Particularly noteworthy is its superior performance on the test data subset.

DETAILS

- Applying outlier imputation techniques led to optimization of the model, particularly enhancing its performance concerning the variables of fare amount and duration.
- Utilizing a linear regression framework, the model offers a reliable means of predicting the estimated fare amount for taxi journeys.

Model metrics:
Net model tuning resulted in:

- R^2 0.87, meaning that 86.8% of the variance is described by the model.
- MAE 2.1 MSE: 14.36 RMSE 3.8



NEXT STEPS

The key factor influencing fare amount was found to be the duration of the ride, a result that was anticipated. According to the model, there's an average rise of $7 for every extra minute; however, this figure may not be entirely dependable due to the significant correlation among certain variables. Gathering more data from less represented routes is recommended.

These insights can be leveraged by the New York City Taxi and Limousine Commission to develop an application enabling users (TLC riders) to preview the estimated fare prior to commencing their journey. The model's fare predictions exhibit a generally robust and trustworthy performance, suitable for further modeling endeavors.

Machine Learning Model

Executive summary report

Commission Prepared by Automatidata

OVERVIEW

The New York City Taxi & Limousine Commission has enlisted the Automatidata data team to construct a machine learning model tasked with forecasting whether passengers in NYC TLC taxi cabs will be generous tippers.

PROJECT STATUS

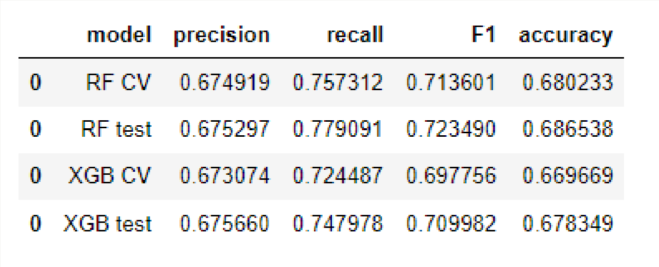
Following ethical considerations, the initial objective of predicting non-tippers was dismissed. Instead, the focus shifted to predicting "generous" tippers, defined as those who tip 20% or more. This adjustment aimed to strike a balance between the interests of taxi drivers and prospective passengers, which can sometimes conflict.

SOLUTION

The data team employed two distinct modeling architectures and conducted a comparative analysis of their outcomes. While both models demonstrated satisfactory performance, the random forest architecture delivered slightly superior predictions. Consequently, the team suggests initiating beta testing with taxi drivers to gather additional feedback before proceeding further.

DETAILS

- The assumption made by the data team was that factors such as a trip's route, predicted fare amount, and time of day might possess a sufficiently strong correlation with tip amounts, enabling accurate prediction of generous tipping behavior.
- Both models are acceptable, but the random forest model is the champion. It is clear that these factors do indeed help predict tipping. The model's F1 score was 0.7235.



	model	precision	recall	F1	accuracy
0	RF CV	0.674919	0.757312	0.713601	0.680233
0	RF test	0.675297	0.779091	0.723490	0.686538
0	XGB CV	0.673074	0.724487	0.697756	0.669669
0	XGB test	0.675660	0.747978	0.709982	0.678349

Future model suggestions

- Collect/add more granular driver and user-level data, including past tipping behavior.
- Cluster with K-means and analyze the clusters to derive insights from the data

NEXT STEPS

Moving forward, the Automatidata data team can proceed by consulting the New York City Taxi and Limousine Commission to present the model results and suggest its potential utility as an indicator of tip amounts. Nonetheless, it should be noted that substantial enhancements to the model would require the acquisition of additional data.