Automatidata

The New York Taxi and Limousine Commission

Project Goal

The New York Taxi and Limousine Commission is looking for a way to utilize the data collected from the New York City area to predict fare amounts for cab rides.

Disclaimer: This is a demo project for a fictional company. Completed for educational purposes.

All stakeholders are fictional, and the data has been partially fabricated.

This projected was originally a part of the Coursera, Google, Advanced Data Analytics Certificate

Preliminary Data Summary

Overview

The NYC Taxi & Limousine Commission has contracted with Automatidata to build a regression model that predicts taxi cab fares. In this part of the project, the Automatidata data team performed a preliminary inspection of the data supplied by the NYC Taxi and Limousine Commission in order to inform the team of key data variable descriptions, and ensure the information provided is suitable for generating clear and meaningful insights.

Project Status

- Explored dataset to find any unusual values.
- Considered which variables are most useful to build predictive models (in this case: total amount and trip distance, which work together to depict a taxi cab ride).
- Considered potential interactions between the two chosen variables.
- Examined which components of the provided data will provide relevant insights.
- Built the groundwork for future exploratory data analysis, visualizations, and models.

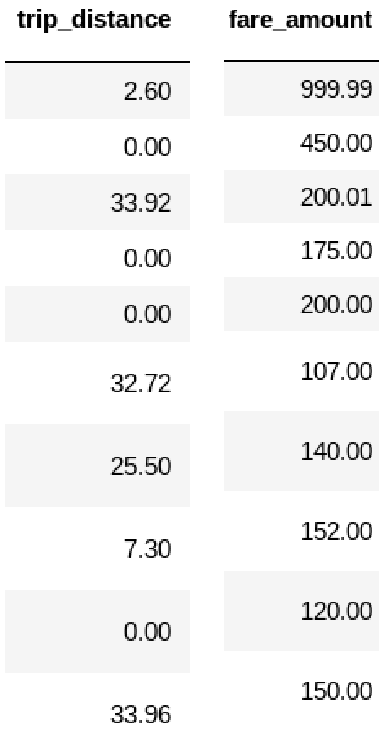
Key Insights

- This dataset includes variables that should be helpful for building prediction model(s) on taxi cab ride fares.
- The identified unusual values are trips that are a short distance but have high charges associated with them, as shown in the total amount variable.

Next Steps

1. Conduct a complete exploratory data analysis.
2. Perform any data cleaning and data analysis steps to understand unusual variables (e.g., outliers).
3. Use descriptive statistics to learn more about the data.
4. Create and run a regression model.

Total_amount variables



trip_distance	fare_amount
2.60	999.99
0.00	450.00
33.92	200.01
0.00	175.00
0.00	200.00
32.72	107.00
25.50	140.00
7.30	152.00
0.00	120.00
33.96	150.00

[Alt-text] The total_amount variable indicates the necessity of further analyzing outlier variables.

Exploratory Data Analysis

Overview

The NYC Taxi & Limousine Commission has contracted with Automatidata to build a regression model that predicts taxi cab ride fares. In this part of the project, the data needs to be analyzed, explored, cleaned and structured before modeling.

Key Insights

The Problem: After running initial exploratory data analysis (EDA) on a sample of the data provided by New York City TLC, it is clear that some of the data will prove an obstacle for accurate ride fare prediction. Namely, trips that have a total cost entered, but a total distance of “0.” At this point, our analysis indicates these to be anomalies or outliers that need to be factored into the algorithm or removed completely.

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

Keys to Success

Ensuring with New York City TLC that the sample provided is an accurate reflection of their data as a whole.

Plan for handling other outliers, such as low trip distance paired with high costs.

Next Steps

Determine any unusual data points that could pose a problem for future analysis in predicting trip fares.

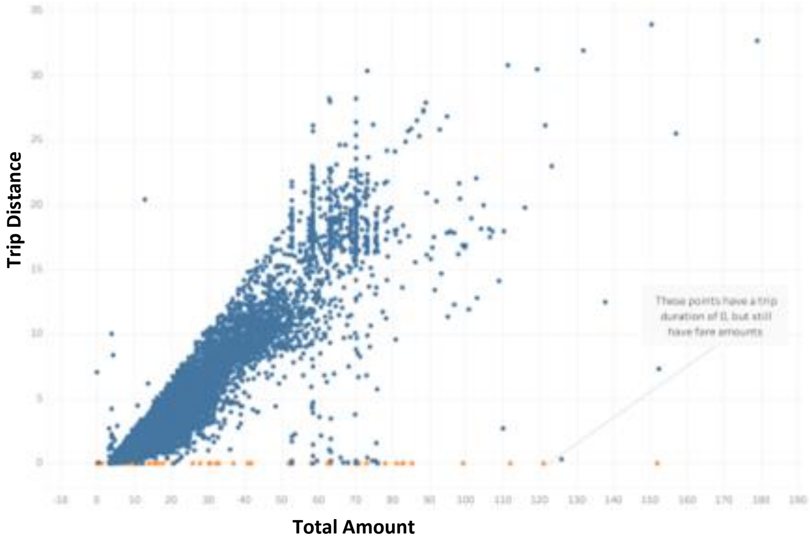
For example, locations that have longer durations.

Determine the variables that have the largest impact on trip fares.

Filter down to consider the most relevant variables for running regression, statistical analysis, and parameter tuning.

Total_amount variables

As a result of the conducted exploratory data analysis, the Automatidata data team considered trip distance and total amount as key variables to depict a taxi cab ride. The provided scatter plot shows the relationship between the two variables. This scatter plot was created in Tableau to enhance the provided visualization.



Statistical Review & A/B Testing

Overview

The purpose of this project is to predict taxi cab fares before each ride. At this point, this project's focus is to find ways to generate more revenue for New York City taxi cab drivers. This part of the project examines the relationship between total fare amount and payment type.

Problem

Taxi cab drivers receive varying amount of tips. While examining the relationship between total fare amount and payment type, this project seeks to discover if customers who pay in credit card tend to pay a larger total fare amount than customers who pay in cash.

Solution

The Automatidata team ran an A/B test to analyze the relationship between credit card payment and total fare amount. The key business insight is that encouraging customers to pay with credit cards may generate more revenue for taxi drivers. However, further analysis is needed.

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 Results

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 Steps

The Automatidata data team recommends that the New York City TLC encourages customers to pay with credit cards, and create strategies to promote credit card payments. For example, the New York City TLC can install signs that read "Credit card payments are preferred" in their cabs, and implement a protocol that requires cab drivers to verbally inform customers that credit card payments are preferred.

Regression Assumptions After Modaling

Issue / Problem

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.

Details

Imputing outliers optimized the model, specifically in regards to the variables of: fare amount and duration.

The linear regression model provides a sound framework for predicting the estimated fare amount for taxi rides.

Key Insights

The feature with the greatest effect on fare amount was ride duration, which was not unexpected. The model revealed a mean increase of $7 for each additional minute, however, this is not a reliable benchmark due to high correlation between some features.

Request additional data from under-represented itineraries.

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.

The model provides a generally strong and reliable fare prediction that can be used in downstream modeling efforts.

Model metrics:

Net model tuning resulted in:

R^2 0.87, meaning that 86.8% of the variance is described by the model.

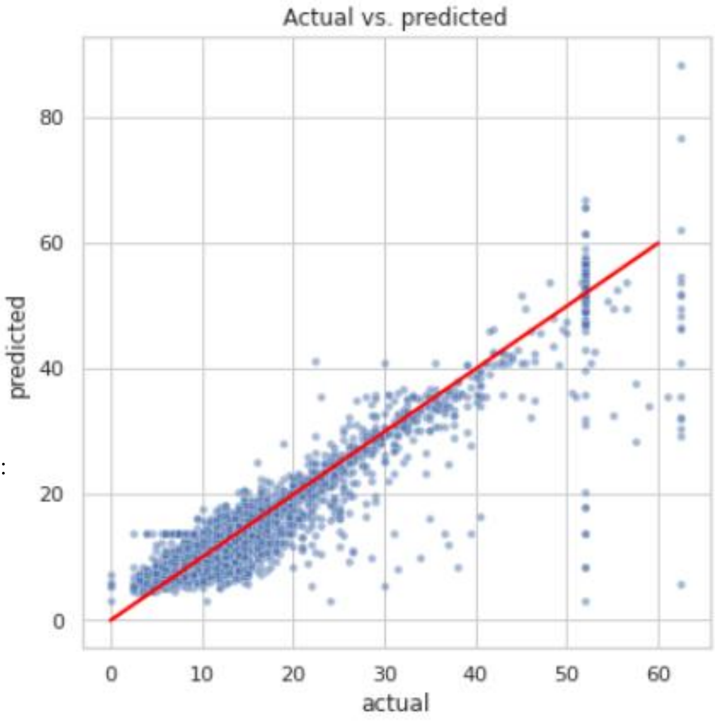
MAE 2.1

MSE: 14.36

Response

The Automatidata data team chose to create a multiple linear regression (MLR) model based on the type and distribution of data provided. The MLR model showed a successful model that estimates taxi cab fares prior to the ride.

The model performance is high on both training and test sets, suggesting that the model is not over-biased and that the model is not overfit. The model performed better on the test data.



In order to showcase the efficacy of the linear regression model, the Automatidata data team included a scatter plot comparing the predicted and actual fare amount. This model can be used to predict the fare amount of taxi cab rides with reasonable confidence. The provided notebook exhibits further analysis on the model residuals.

Alt-text: The scatter plot shows a linear regression model plot illustrating predicted and actual fare amount for taxi cab rides.

Machine Learning Model Outcomes

Overview

New York City Taxi & Limousine Commission has contracted the Automatidata data team to build a machine learning model to predict whether a NYC TLC taxi cab rider will be a generous tipper.

Problem

After rejecting the initial modeling objective (predicting non-tippers) out of ethical concern, it was decided to predict "generous" tippers—those who tip ≥ 20%. This decision was made to balance the sometimes competing interests of taxi drivers and potential passengers.

Solution

The data team used two different modeling architectures and compared their results. Both models performed acceptably, with a random forest architecture yielding slightly better predictions. As a result, the team would recommend beta testing with taxi drivers to gain further feedback.

Details

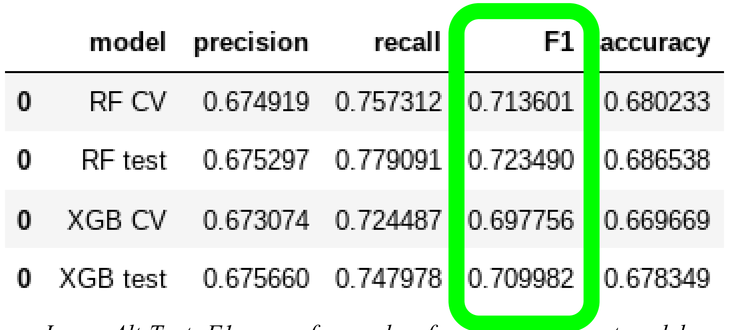
Behind the Data

The data team's assumption was that a trip's itinerary, predicted fare amount, and time of day may have a strong enough relationship with tip amount that we could accurately predict generous tipping.

After the data team built the identified models and performed the testing, it is clear that these factors do indeed help predict tipping. The model's F_1 score was 0.7235.

Results Summary

The resulting algorithm is usable to predict riders who might be generous tippers, with reasonably strong precision, recall, F_1, and overall accuracy scores. Refer to the "next steps" section for suggestions.



	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

Image Alt-Text: F1 scores for random forest and XGboost models

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

As a next step, the Automatidata data team can consult the New York City Taxi and Limousine commission to share the model results and recommend that the model could be used as an indicator of tip amount. However, additional data would be needed to realize significant improvement to the model.