Predicting Travel Behavior in NYC

Gabriel Eze

Introduction

The NYC Taxi & Limousine Commission dataset provides a rich environment for exploring rider behavior, traffic flow, and travel patterns across the city.

This project focuses on whether simple statistical models can reliably infer:

- Trip duration
- Total fare
- Likelihood of multiple passengers

Our goal is to understand how well interpretable models perform on a noisy, real-world database, and whether they can support decision-making about comfort, cost, and traffic predictability in NYC taxis.

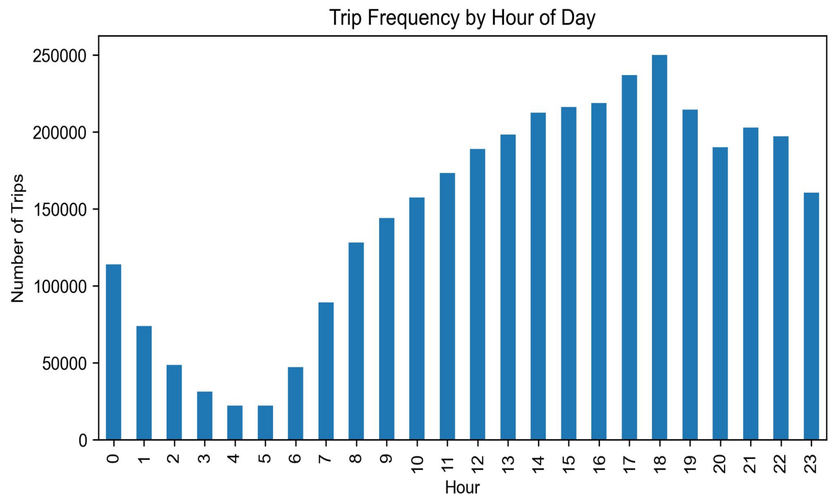

Figure 1. Rush Hour Pattern

Data & Feature Engineering

The project uses millions of NYC Yellow Taxi trip records with the following engineered features:

- hour
- weekday
- Pickup_Borough / Dropoff_Borough
- trip_time_min
- passenger_count
- total_amount
- trip_distance

Trip filtering removed unreliable trips:

- trip_distance: 0.1–25 miles
- trip_time_min: 1–100 minutes
- total_amount: $5.99–$250

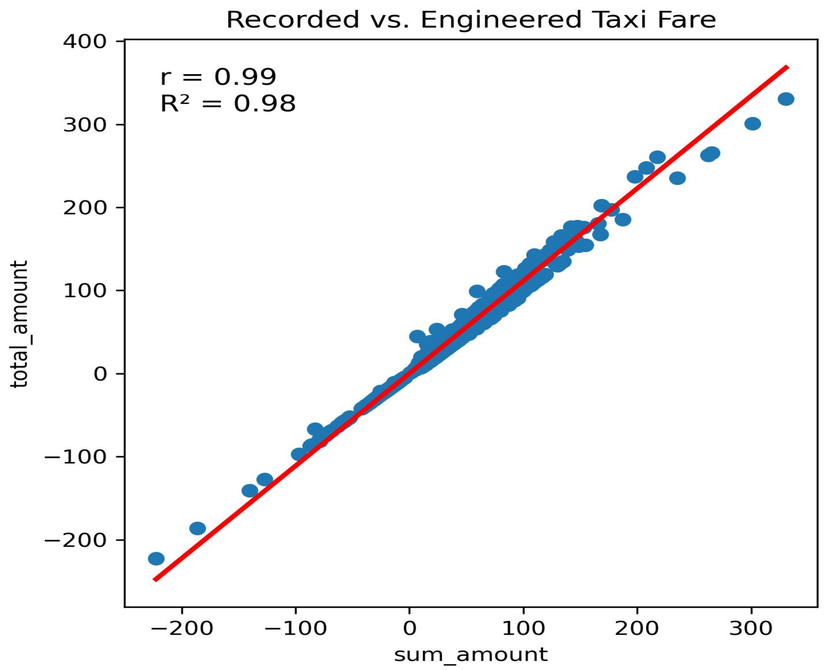

Figure 2. Taxi Fare Anomaly

Methods

We trained three independent models:

1. Trip Duration
Linear Regression
Target: trip_time_min

2. Total Fare
Linear Regression
Target: total_amount

3. Multi-Passenger Likelihood
Logistic Regression
Target: multi_passenger

Table 1. Model Performance Summary

Target Variable	Method	Score
Trip Duration (min)	Linear Regression	**MAE = 4.28** RMSE = 6.75
Total Fare ($)	Linear Regression	**MAE = 3.43** RMSE = 6.28
>1 Passenger Likelihood	Logistic Regression	**AUC = 0.59**

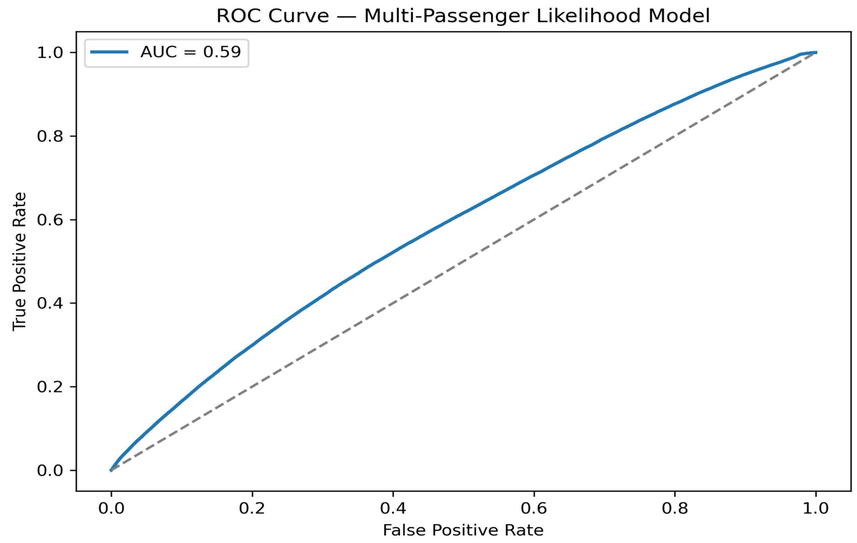

Figure 3. ROC Curve

Discussion

- **Trip duration** and **total fare** is strongly influenced by trip distance, hour, and borough pair, capturing traffic and routing patterns at a coarse level.
- **Passenger likelihood** is harder to model because it depends on unobserved factors (trip purpose, party size, social context) that are not captured in TLC records.

Conclusion

This project shows that:

- **Simple linear regression models** can effectively forecast **trip duration** and **total fare** for NYC taxi trips using only basic features.
- A **logistic regression model** for multi-passenger detection yields **AUC ≈ 0.59**, establishing a realistic baseline for behavior-focused modeling.

Future work could incorporate:

- Weather, events, and real-time traffic data
- Route-level features (bridges, tunnels, corridors)
- Nonlinear or interaction-focused models (e.g., GAMs, tree ensembles) for targeted improvements

Acknowledgements

Thanks to:

- **DSC 500: Data Science Capstone** at Centre College
- Faculty and peers who helped provide feedback