Dynamic Decision Support System for Taxi Drivers

15.093 - Optimization Methods

Pranav Girish (pranav7@mit.edu),
Krishanu Datta (krishanu@mit.edu)

Executive Summary

Problem Description: This project proposes a data-driven project focused on optimizing the decision-making process for taxi drivers by leveraging historical ride data. Currently, taxi drivers rely on manual, case-by-case decision-making when deciding whether to accept or decline a trip. This approach leaves drivers susceptible to suboptimal choices, as they lack a comprehensive understanding of the potential profitability of each trip. Further, the algorithms currently in place in ride-hailing apps only consider optimizing overall trips, without granularity to optimize the earnings of individual drivers. The key objective is thus to develop an optimization model that guides taxi drivers in determining whether to accept or decline a trip based on potential returns for the day, considering factors such as distance, time of day, prices, and historical earnings.

Data: New York City taxi trips from the first half of 2023. Narrowing the time period of the dataset down to June 2023, the dataset contains 3,307,139 rows, each a unique completed trip, and 20 features relating to ride characteristics such as location, distance, time and price. The locations are categorized into 265 zones, and the scope of this project considers trips between a subset of these zones.

Optimization: The taxi driver problem is formulated as an integer optimization model, with the objective being to maximize expected profits and the decision being whether or not to accept a trip request. The constraints are formulated using both information provided by the user (the taxi driver) regarding their preferences, and historical trip data for trip-related specifics. The developed model was seen to greatly outperform the baseline, which would be to accept every incumbent trip. The key element of the model is the calculation of expected profits, which is outlined further later in the report. The model is run every time a driver receives a trip request to generate an optimal decision - this practice is reproduced by synthetically simulating a driver's shift through weighted generation of trip requests.

Conclusion: The proposed data-driven decision-making model offers an advanced tool for taxi drivers, enhancing their efficiency and profitability. By integrating historical ride data and employing advanced optimization techniques, the optimization model provides a comprehensive approach to guide real-time decisions. Further validation and refinement of the model with diverse scenarios could enhance its applicability in practical settings.

1 Background and Motivation

The taxi industry has long operated on a traditional model where drivers make on-the-fly decisions about accepting or rejecting ride requests, often without access to comprehensive data that could inform these choices. In the face of an increasingly competitive and dynamic transportation landscape dominated by ride-hailing services, there is a growing need for a more sophisticated and data-driven approach to optimize the decision-making process for taxi drivers. The motivation behind this project lies in addressing the limitations of the current manual decision-making paradigm, which exposes taxi drivers to potential financial inefficiencies. By harnessing the power of historical ride data, this project seeks to empower taxi drivers with a predictive model that considers a multitude of factors, including distance, time of day, pricing structures, and individual driver preferences. The aim is to move beyond the one-size-fits-all algorithms used in ride-hailing apps, providing drivers with personalized insights that maximize their daily earnings while minimizing the risk of accepting rides that are unprofitable or rejecting those that are potentially lucrative in the long-term. This project's significance extends beyond the immediate financial impact on individual taxi drivers. By enhancing the efficiency and income of drivers, the overall quality of service within the taxi industry can be elevated, potentially leading to improved customer satisfaction and increased competitiveness against ride-hailing platforms. The proposed optimization model aligns with the broader trend of integrating data-driven decision-making into traditional industries, offering a transformative solution that is catered towards benefiting the taxi drivers themselves, contributing to their resilience in the industry.

2 Data Sources and Aggregation

The dataset used encompasses New York City taxi trips from the first half of 2023 [1]. By narrowing the temporal focus to June 2023, the dataset comprises 3,307,139 distinct rows, each representing a unique taxi trip, and includes 20 features capturing essential ride characteristics such as location, distance, time and price. Noteworthy among these features are pickup and drop-off times, which provide temporal information crucial for imposing time-related constraints. Additionally, the overall city is segmented into 265 zones, and the zone of the pickup and drop-off location of each trip is listed separately, facilitating the mapping of pickups and drop-offs to generalized zones, serving as valuable location-based constraints. The 'Fare amount' and 'Total amount' variables emerge as pivotal components for estimating expected profits, representing the compensation taxi drivers receive for their services.

In the data aggregation phase, the June 2023 trips dataset is processed to organize the information into 30-minute intervals for each pair of locations (pickup location and drop-off location). This granular temporal segmentation allowed for the calculation of average fares, time taken, and the total number of trips within each designated time slot. Furthermore, a strategic selection of connected subset locations was undertaken, taking into account the spatial distribution of taxi services to simulate a scenario where a driver operates within a specific region. This aggregation strategy lays the groundwork for the development of an optimization model that incorporates both temporal and spatial constraints for informed decision-making by taxi drivers.

3 Optimization Methods

3.1 Overview

The implementation aims to enhance the decision-making process for taxi drivers by simulating their choices in accepting or rejecting ride requests. Leveraging historical ride data, the model utilizes an expected return calculation to guide decisions, optimizing for overall profit while considering constraints such as time intervals, trip availability, and destination probabilities.

3.2 Problem Formulation

Considering a single trip request, system can be formulated as an integer (binary) optimization problem, with the decision being whether or not to accept a ride request.

3.2.1 Input Parameters

The input parameters for the proposed optimization model are twofold; the information and preferences of the driver, and the historical data of the region.

The parameters entered by a driver are:

- Starting Location: This is the location from which a driver starts their shift.
- Shift Duration: This is the period of time for which a driver would like to be online
- Number of Trips: This is the number of trips the driver is ready to embark on over the course of their shift.
- Current Ride Request: This is the details of the ride request the driver has received, i.e. its pick-up location, drop-off location, and time of receiving the request.

The parameters obtained using historical trip data and obtained through proper data preprocessing are:

- Pickup Locations: This is the set of locations from which trips have been started. There are a total of 265 such zoned locations in the available dataset.
- Drop-off Locations: This is the set of locations at which trips have been ended. There are a total of 265 such zoned locations in the available dataset.
- Time interval: This is the set of intervals splitting a day into 48 buckets (each bucket represents a 30 minute period).
- Total Number of Trips: This is the aggregated (sum of) number of trips, indexed based on pickup location, drop-off location, and time interval of the trip.
- Fare Amount: This is the aggregated (averaged) fare for trips, indexed based on pickup location, drop-off location, and time interval of the trip.

3.2.2 Decision Variable

The decision variable for the proposed optimization model is a choice z ($z \in \{0, 1\}$), which represents the binary decision of whether or not to accept a ride request.

3.2.3 Objective Function

The main objective for the proposed optimization model is to maximize the driver's expected profit. It is formulated as a maximization problem, comparing the expected return when the driver accepts the incumbent trip request against when they reject it.

3.2.4 Expected Returns Calculation

To improve computational time, the calculation of expected returns is done through an auxilary function rather than within the optimization formulation. The developed algorithm is designed to compute the expected return for a taxi driver over a series of potential future trips, given a current trip request. This computation takes into account both the immediate fare earned from a trip and the recursive exploration of subsequent trips with their associated probabilities.

The implementation considers a trip request from a given pickup location to a given drop location at a given time. If the driver chooses to embark on this trip, they are expected to earn an amount equal to the sum of the fare of the current trip and the expected profits from future trips. The former is known, and the latter is computed by finding the probability of receiving a trip request from the current trip's destination to each subsequent location, which recurses through all possible future steps in the driver's shift.

The algorithm described below highlights the main elements of the calculation. Here, D represents all the possible further destination locations from the destination location of the current trip. The algorithm thus finds the expected returns by iterating through each following step over the course of the driver's shift.

Algorithm 1 Expected Returns Calculation

Data: Trip Details, Remaining trips n, Current Time t

if *number of remaining trips* $= 0$ *or shift end time* $\leq t$ **then**
  return 0
else
  Expected Future Returns $= \sum_D$ Expected Returns(Next_Trip, $n - 1$, $t - t_{trip}$)
  Expected Returns (Trip, n, t) = Current Trip Fare + Expected Future Returns
  return Expected Returns (Trip, n, t)

Result: Expected Returns of the concerned trip

3.2.5 Mathematical Formulation

The overall integer optimization problem is thus formulated as:

Objective:

$$\text{maximize} \quad EP$$

Subject to:

$$\begin{aligned} EP &= z \cdot ER(T, n-1, t) + (1-z) \cdot ER(NT, n, t) \\ ER(T, n, t) &= \text{Fare}(T, t) + \gamma \sum_{\text{Possible Trips}} P(\text{Next Trip}) \cdot ER(\text{Next Trip}, n-1, t + t_T) \\ n - z &\geq 0 \\ t + t_T &\leq \text{end\_time} \\ z &\in \{0, 1\} \end{aligned}$$

In the above formulation, the objective function maximizes overall expected profits. The first constraint finds the expected profit in terms of the decision variable z (whether to accept or reject a request). In order to calculate the expected returns of rejecting a ride request, the trip NT has the same drop-off location and pickup location. The second constraint computes the expected returns as a function of future steps, with future returns discounted by a factor γ - here the sum to find expected returns from subsequent trips is taken over "Possible Trips", which is the set of future trips from the destination of the current trip T (t_T is the time taken by the current trip). The third and fourth constraints ensure that the conditions on number of trips in a shift and time of a shift are met. Finally, last constraint ensures that the decision is binary.

3.2.6 Simulating a Real-World Scenario

The optimization model formulated in the previous section only considers the decision for a single trip. In reality, a driver would have to make a sequence of decisions throughout their shift to optimize their overall profits. In order to emulate this, a simulation of a driver's shift is used. Here, given a starting location and driver preferences, the simulation runs through the driver's shift by synthetically generating a trip request through random sampling of historical trips. At each step in the simulation a trip is generated, the optimization model is executed, and the location of the driver and time are updated with the driver making a decision in accordance to the optimal solution of the model.

While we ran simulations across an assortment of differing driver preferences including various starting locations, start times, end times and maximum number of trips desired, we opted to display our results from a subset of these tests. In Table 1 below, we showcase a taxi driver working the graveyard shift (from midnight to 8AM) and their expected profits varying by their start location, at any one of the six most common taxi zones in New York. For each of these simulations, we compare the expected profits of a series of trips guided by optimal decision making to the baseline strategy of accepting every trip the driver receives. In each case, the driver is better off following the optimal decision method as increases in profits range as high as 68.5% depending on starting time and location.

Table 1: Profit Comparison Under Different Methods (No. Trips = 4, EndTime = 8AM)

Starting Zone	Driver Start Time	Profit (Opt)	Profit (Accepted Every Ride)	Improvement (%)
Upper East Side South	1:00 AM	204.04	156.99	29.97
Upper East Side South	4:00 AM	281.5	218.2	29.01
Midtown Center	1:00 AM	246.1	160.71	53.133
Midtown Center	4:00 AM	254.26	231.14	10.003
Upper East Side North	1:00 AM	231.71	147.9	56.667
Upper East Side North	4:00 AM	368.69	277.17	33.019
JFK Airport	1:00 AM	193.24	114.66	68.533
JFK Airport	4:00 AM	125.24	124.18	0.854
Midtown East	1:00 AM	280.27	272.01	3.037
Midtown East	4:00 AM	373.18	282.57	32.066
Times Sq/Theatre District	1:00 AM	329.22	297.19	10.778
Times Sq/Theatre District	4:00 AM	198.12	167.01	18.628

4 Insights and Discussion

Optimal Decision-Making: The optimization model presented in this project significantly enhances the decision-making process for taxi drivers. Through rigorous simulation and expected return calculations, the model guides drivers to make informed choices regarding trip acceptance or rejection. Key insights reveal instances where the model successfully directed drivers to accept trips that are profitable in the long run over those that are only lucrative in the short term, ultimately leading to improved profitability over manual decision-making.

Real-World Applicability: In evaluating the model's practicality, it becomes evident that the data-driven approach aligns well with the challenges faced by taxi drivers. The integration of historical ride data provides a robust foundation for decision-making in a dynamic and competitive landscape. Discussions around real-world applicability consider the model's adaptability to diverse taxi service operations and potential benefits for drivers navigating complex urban environments.

Limitations: For real-world application, decisions should be quick. In the current implementation however, the speed of execution depends heavily on the number of locations used. An approach that serves to solve this limitation but is beyond the scope of this project is to cluster the available regions based on geographical proximity, and make decisions for trips between these formed clusters. This approach is reasonable as a trip between locations within a single cluster would be less time-consuming and hence is a comparatively straightforward decision to accept, as opposed to longer trips. Such an implementation can aid drivers to make informed decisions on taking up longer trips, where they may not have existing knowledge of the destination region.

References

[1] NYCTaxis. Historical trip data, 2023. Available at kaggle.com.