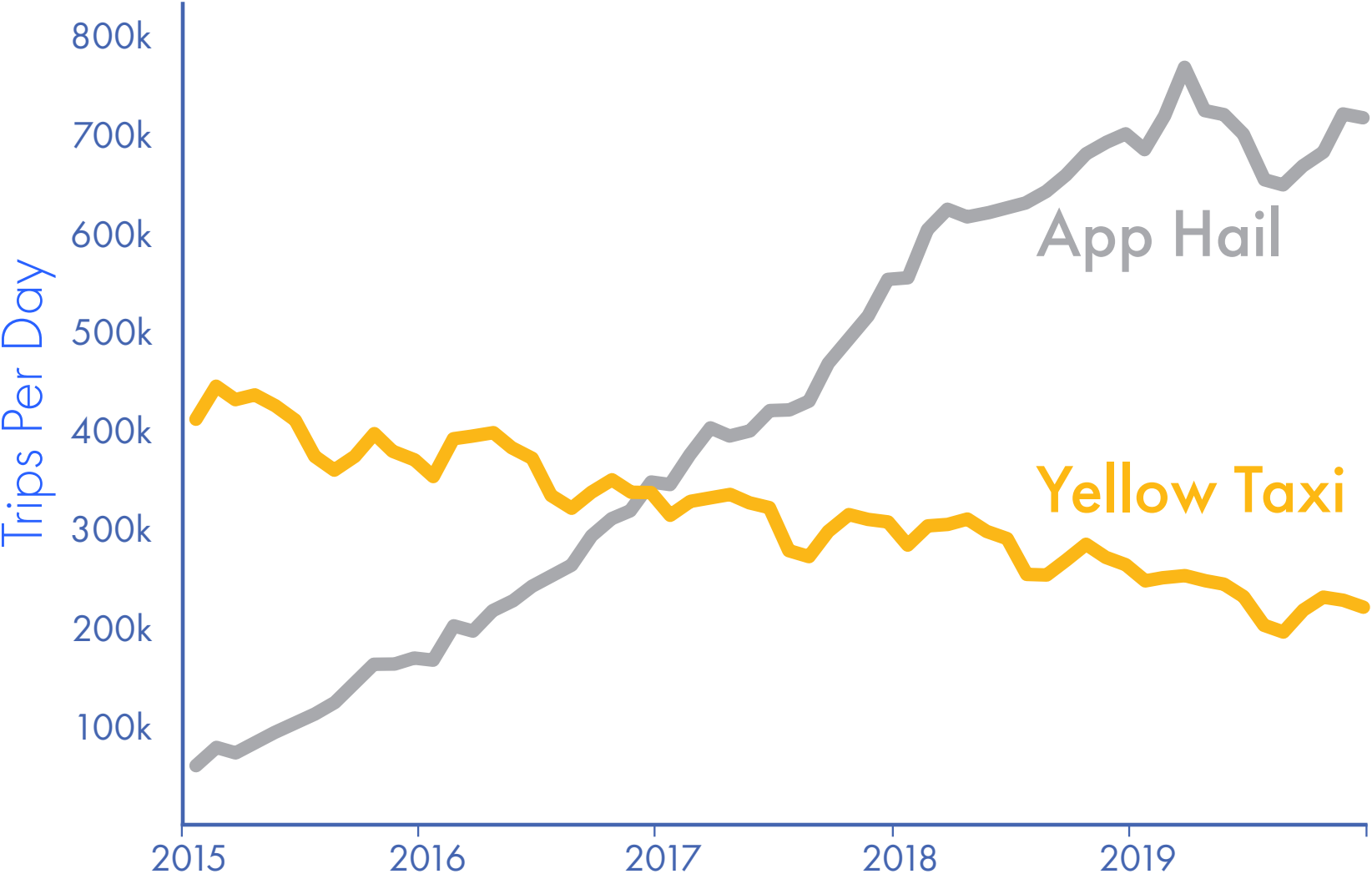


NYC Taxi: PageRank Pickup Zone Recommendation

Team 146 | Laurent Aeschbach | Karn Agarwal | William Coningsby | David Kaczmarkiewicz | John Torossian

Taxi drivers can **make more** per shift by using our **PageRank** pickup zone recommendations.

With the rise of app-hail services like Uber, yellow taxis have seen declines in both trips per day and market share.



Our model demonstrates a 5% increase in the average earnings of a driver in a typical 8 hour shift.

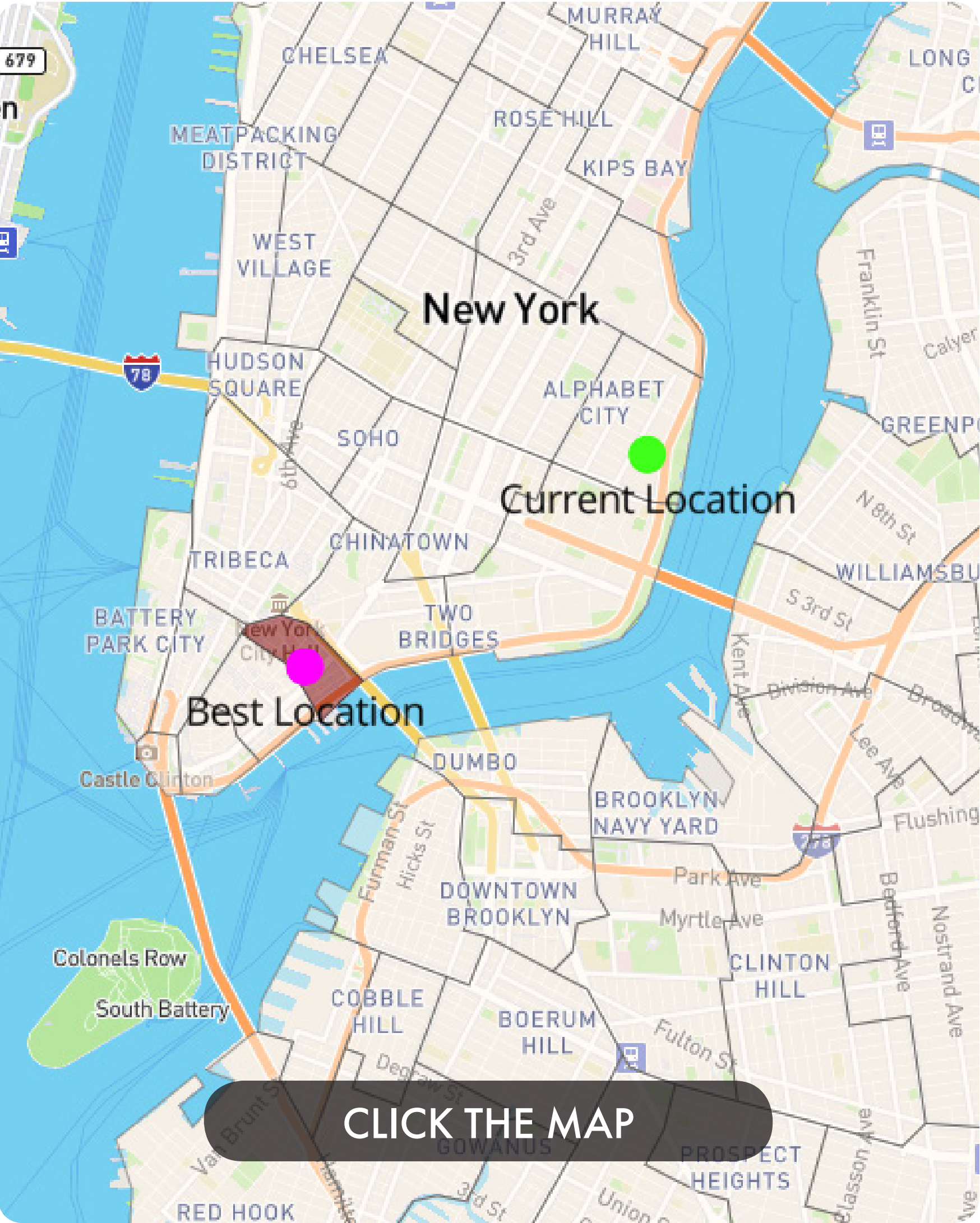
Category	Earnings
Model	\$411
Baseline	\$390

VISUALIZATION

Our application providers drivers with a zone recommendation to increase expected total shift earnings. The driver selects:

- current zone
- month, weekday/weekend, day/night
- hours remaining in shift
- minutes willing to travel to more profitable zone

The application then identifies the best nearby zones to which the driver should move in order to maximize expected earnings.



INTRODUCTION

Problem: Can a New York City taxi driver increase their overall earnings by moving to the most profitable near by pick up zones?

Importance: New York City taxi drivers recently engaged in a hunger strike to bring light to financial struggles brought on by the emergence of app based ride hailing services (Uber and Lyft) and the COVID-19 pandemic. Taxi drivers will need to leverage technology to improve their competitiveness with Uber and Lyft in order to have a sustainable business.

APPROACH

Our approach solves the problem of increasing the profitability of a driver’s day by maximizing the total fares earned within the time remaining in the driver’s shift.

Innovation: We have demonstrated a novel way of determining the best pickup zones using PageRank. To evaluate the algorithm we developed a random walk simulation to estimate the total amount a driver can make after a specified number of trips. Evaluation shows that it is better than random selection of near by pickup zones. Our application allows drivers to identify the best nearby zone to obtain a fare by selecting the relevant inputs and their current zone via an interactive map.

1 Create Graph

Create a weighted directed graph where each taxi zone is a node and the weight of an edge is the number of trips between them in a given time frame.

2 Drop Off Zone Score

Use PageRank to find the steady state probability distribution of ending up in any given zone after a sufficient number of steps.

3 Pick Up Zone Value

Calculate a weighted profit for each pickup zone as a measure of profitability of that zone. This weight factor is determined by the steady state probability from the previous step.

4 Recommendation Engine

For each potential trip use a simulation to find the best pick up zone based on the transit time, time left in a driver’s shift, and other relevant factors.

EXPERIMENTS & RESULTS

Evaluation: To evaluate the PageRank algorithm we set up a simulation where we compared the average amount a driver can make starting his day in one of the top 10 pickup zones as recommended by our algorithm vs. a random starting zone. We ran 1,000 experiments for each where each experiment consists of an 8 hour random walk. We compared this for the historical dataset from 2016 to 2019.

Results: As shown in the figure to the right, on average the expected amount that can be made from a pickup zone recommended by our algorithm is more than a random selection. The functionality of finding the best neighboring pickup zones given a current zone is best demonstrated through our [visualization](#).

The following table shows the average earnings per 8 hour shift based on 2019 data by day of week and time of day.

	Baseline	Model	Increase
Weekday Day	\$360.80	\$378.84	5.00%
Weekday Night	\$406.05	\$426.33	4.99%
Weekend Day	\$393.62	\$419.64	6.61%
Weekend Night	\$402.14	\$419.52	4.32%

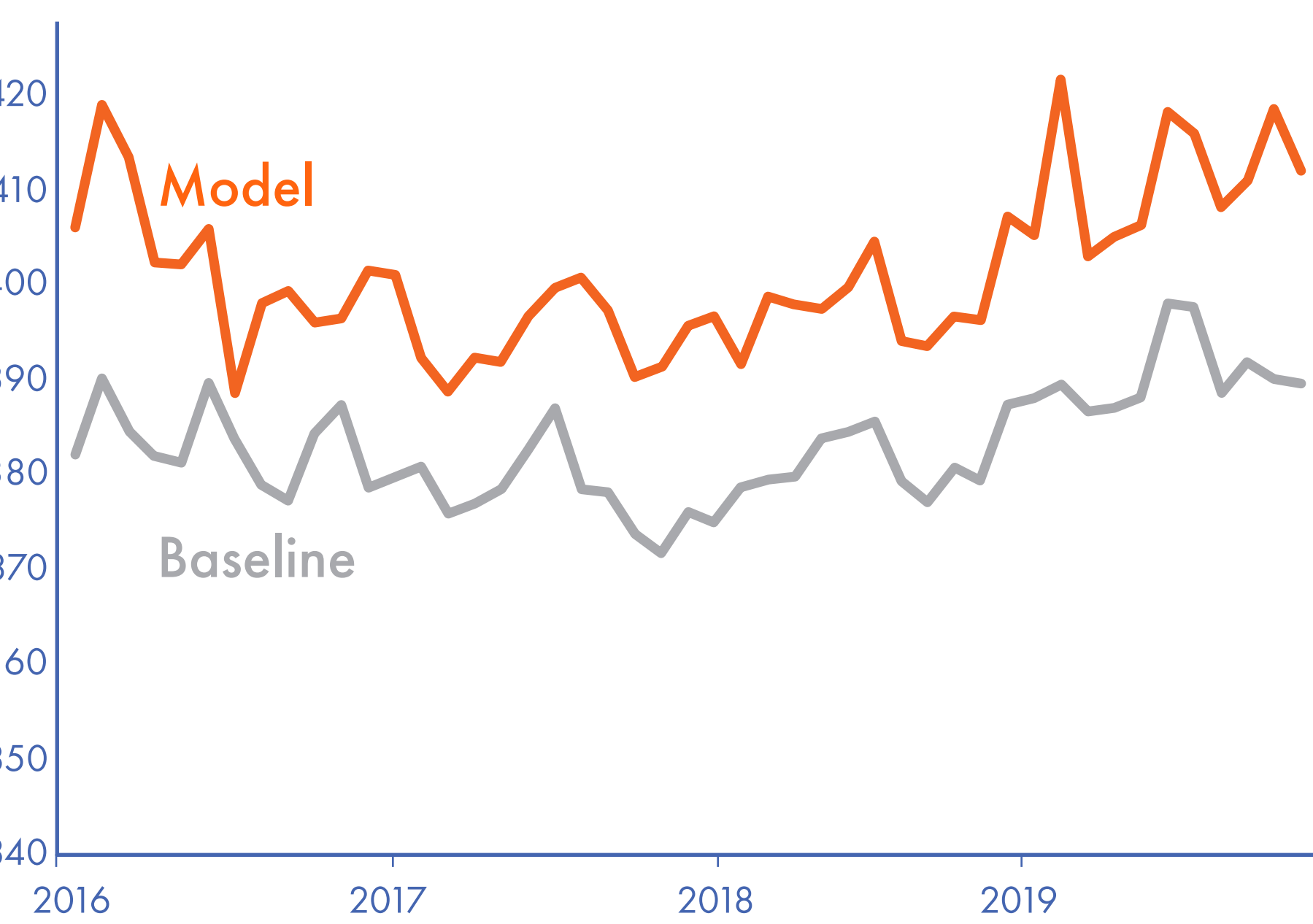
Comparison: We explored other models including a two-dimensional embedding approach in order to identify the best zones, but this approach did not result in an improvement in a driver’s earnings.

DATA

Source: The raw data is published by the New York City Taxi and Limousine Commission. We obtained lightly processed data in the form of a PostgreSQL database from a public facing [github](#). Our project uses data from calendar years 2016-2019.

Characteristics: The subset of the data we are using has 478,840,071 rows, where each row represents a trip. The size of the subset is over 90GB.

Model vs Baseline (random) starting nodes from the previous month for 8 hour shift. Results averaged across day, night, weekdays, and weekends.



Model vs Baseline (random) starting nodes for a typical 8 hour shift. Distribution show from weekday days July 2017, but all months show a similar distribution.

