NYC Taxi Hotspot Prediction and Analytics

Rena Agrawal
Georgia Institute of Technology
Atlanta, Georgia
ragrawal66@gatech.edu

Gautam Awalegaonkar
Georgia Institute of Technology
Atlanta, Georgia
gawalegaonkar3@gatech.edu

Narendra Bhattacharya
Georgia Institute of Technology
Atlanta, Georgia
nbhattacharya3@gatech.edu

Bhageerath Bogi
Georgia Institute of Technology
Atlanta, Georgia
bbogi3@gatech.edu

Sumana Haldar
Georgia Institute of Technology
Atlanta, Georgia
shaldar7@gatech.edu

1. PROBLEM DEFINITION

Cost of a NYC taxi medallion has dropped from a peak $1.3m in 2014 to $150k in 2017, driven by the entry of ride-sharing apps such as Uber/Lyft. While there are several livery and ride-sharing apps that allow drivers to maximize revenues by identifying demand and directing drivers to high revenue locations, there is little support for the yellow taxis as this is mostly a ride-hail model. Given a day, time and geographical location, our project seeks to identify zones with the highest likelihood of rides for taxi drivers with a user engagement driven visual/analytical solution. The solution will visualize NYC taxi data on a GEO map and quantify it based on driver engagement index.

2. HEILMEIER QUESTIONS

(1) What are you trying to do? Articulate your objectives using absolutely no jargon.

Based on the driver's reference location, weather condition, day and time predict the geographical locations/zone that has higher relative probability of finding a ride request also called Hotspot using regression modelling. Create an interactive app to enable cab drivers identify high potential & profitable pick up points in and around their vicinity.

(2) How is it done today; what are the limits of current practice?

Currently, yellow taxis do not have a user interface for a driver to pick a location of interest to get quick ride requests, while this technology exists with ride-sharing companies.

Limitation: Hotspot selection is solely based on driver's limited experience and success rates. No intelligent algorithm / visual solution that makes use of historical data to route drivers to prime hotspots.

(3) What's new in your approach? Why will it be successful?

We propose an approach to use the historical taxi trip data and come up with hotspot indicator based on driver engagement index to color the map. This approach will be successful as routing the driver to areas of low engagement will help the driver to get quick ride requests at the same time reducing the customer wait time.

This is a novel approach looking at the driver's perspective of the yellow taxi business. A regression-based model that considers historical pick up date, day, time, location, weather to predict the anticipated ride requests. We plan to use randomly generated driver movement data to show interactively the change in driver engagement index as the drivers engage with the application.

(4) Who cares?

Individual / independently owned taxi drivers.

(5) If you're successful, what difference and impact will it make, and how do you measure them

(e.g., via user studies, experiments, ground truth data, etc.)?

Measure:

- Ratio of current customer wait time to previous customer wait times
- Ratio of current driver idle time to previous idle time

Average revenue per day for per driver

Impact:

- Increase customer/driver satisfaction via reduced wait times
- Reduce the wait time of a driver to get a ride request
- Maximize the revenue for yellow taxis
- Increased competitiveness for yellow taxis against ride sharing companies

(6) What are the risks and payoffs?

Risks:

Can route more drivers than anticipated. Needs real time location feedback from current drivers to avoid overcrowding. Real time driver GPS data needs to be updated to ensure recent trends are being refreshed.

Payoffs:

Same as 'impact' section above

(7) How much will it cost?

No cost since we use open-source data set, software applications and surveys.

(8) How long will it take?

45 days (20+25)

(9) What are the midterm and final exams to check for success? How will progress be measured?

- Midterm
  Development of a proto-type, including Data Cleaning, Feature Extraction, Preliminary data analysis + Visualization with sample data
- Final
  Deliver final solution including Scale Up + Fine-tuning data analysis + Visualizing

3. INNOVATION

This is a novel approach looking at the driver's perspective of the ride sharing business.

4. PLAN OF ACTIVITIES

- Literature Survey: Understanding NYC Taxi Data
- Product Design: Identifying KPI's, Determine driver engagement index
- Data Analysis – Sanitize the data for product consumption
- POC - Build D3 Visualization
- Scale Up - Monitor/Control

We intend to finish the following activities:

- By 15 March:
  Product Design: Come up with a roadmap/schematic/wireframe for the product
  Identify Key Performance Indicators and methodology to derive these metrics from the available data
- By 1 April:
  Data Cleanup / Sanitization:
  Put together a sample POC viz template
- By 20 April:
  Scale up/ Monitor/Control the environment
- By end of semester:
  Report writing/ Poster Prep

All team members have contributed similar amount of effort and expected to participate equally for the duration of the project.

5. LITERATURE SURVEY

Taxis remain a key asset for urban mobility despite the tremendous growth of modern mobility-on-demand service providers such as Uber and Lyft. A fundamental understanding of the factors that affect the taxi demand is essential for planning an effective multi-modal transportation system and can shed lights on new on-demand services. This study addressed a gap in literature by investigating the historical taxi pick up data and routing the drivers accordingly to reduce taxi wait times hence increase driver engagement.

Qiuyuan Yang1 et al. [2] addresses the social attributes of functional regions upon big traffic data in Beijing and apply the knowledge to maximize drivers' profit. Time-Location-Sociality model is introduced in order to identify three-dimensional properties of city dynamics, which can effectively predict the distribution of passengers for different social functional regions.

In Han-wen Chang et al. [1], a four-step approach is proposed to solve the taxi demand analysis problem. Considering the context, taxi request records are filtered. These records are clustered according to the spatial distance. For each cluster identified, corresponding roads are found, and the cluster is associated to the semantic meaning of the representative roads. Hotness index is calculated based on the property of the clusters and the distance from the taxi driver to the cluster.

There are researches focusing on finding significant locations from GPS traces. Ashbrook and Starner (2003) [12] use k-means-like iterative approach to cluster places into locations. Palma et al. (2008) [9] propose a clustering method based on speed measurement to distinguish stops and moves in a single trajectory

Samuel Palmisano, the IBM CEO, proposed the concept of smart city in his speech [14]. In recent years, researchers start to keep a watchful eye on big data generated by widely deployed sensors in smart city.

Walravens et al. [15] provide insight of the current state of the mobile services, and regard the mobile sensors as the primary interface to modern cities. With the development of information and communication technology, the project of smart city has been constructed in New York and Singapore [4].

A number of methods have been proposed to maximize the profits of taxi drivers by using historical GPS trajectories of taxis, and enhance the opportunity of finding vacant taxis for passengers simultaneously [6]–[8].

Our strategy goes one-step further uses a decision tree based regression approach to evaluate the number of anticipated ride requests based on the historical data. We also plan to include a dummy driver movement data (in reality it should be derived from real time driver gps data) to visualize the change in driver engagement index based on driver movement to different location id's. We are also probing to include other attributes such as population growth, weather data into the regression models to improve the accuracy of our prediction, which is lacking in current literature/research.

6. REFERENCES

[1]   Han-wen Chang, Yu-chin Tai and Jane Yung-jen Hsu*, "Context-aware taxi demand hotspots prediction"

[2]   Qiuyuan Yang1 ,Zhiqiang Gao1 ,Xiangjie Kong1,2 ,Azizur Rahim1 ,Jinzhong Wang1 ,Feng Xia1, "Taxi Operation Optimization Based on Big Traffic Data"

[3]   Eric J. Gonzales, Ph.D., Ci (Jesse) Yang, M.S., Ender Faruk Morgul, M.S., and Kaan Ozbay, Ph.D. , "Modeling Taxi Demand with GPS Data from Taxis and Transit"

[4]   K. Su, J. Li, and H. Fu, "Smart city and the applications," .

[5]   Umang Patel, "NYC Taxi Trip and Fare Data Analytics using BigData"

[6]   D. Zhang, T. He, S. Lin, S. Munir, and J. Stankovic, "pcruise: Online cruising mile reduction for large-scale taxicab networks," .

[7]   N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," .

[8]   D. Zhang, L. Sun, B. Li, C. Chen, G. Pan, S. Li, and Z. Wu, "Understanding taxi service strategies from taxi gps traces,".

[9]   Palma, A.T., Bogorny, V., Kuijpers, B. and Alvares, L.O. (2008) 'A clustering-based approach for discovering interesting places in trajectories'.

[10]   Yong Gao, Dan Jiang & Yan Xu (2018) Optimize taxi driving strategies based on reinforcement learning, International Journal of Geographical Information Science

[11]   Joya A. Deri, Franz Franchetti, and Jose M.F. Moura, 2016, "Big Data Computation of Taxi Movement in New York City"

[12]   Yeqian Lin, Wenquan Li, Feng Qiu, and He Xu, "Research on Optimization of Vehicle Routing Problem for Ride-sharing Taxi," .

[13]   Ashbrook, D. and Starner, T. (2003) 'Using GPS to learn significant locations and predict movement across multiple users'.

[14]   Soderstr ¨ om, T. Paasche, and F. Klauser, "Smart cities as corporate ¨ storytelling".

[15]   N. Walravens, "Mobile city applications for brussels citizens: Smart city trends, challenges and a reality check"

[16]   Todd W. Schneider "Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengean