Yellow Taxi Demand Prediction using Big Data

Shruti Sharma
Data Science, College of Science
San Jose State University
San Jose, USA
shruti.sharma01@sjsu.edu

***Abstract*— With continuous dependence on technology, data has been increasing exponentially. The data is not just data anymore, it has now become Big Data. To analyze this Big Data huge storage is needed. Hadoop being one such platform has made accessing and analyzing this data easy. This document talks about analyzing the big data for deep insights by making use of data visualization techniques. With increasing demand of taxis in New York city, it has become essential for the drivers to be at the right place at the right time while for customers it is essential to avail the taxis within a short distance and duration.**

Keywords—Big Data, Hadoop, MapReduce, Data Visualization, Data Analysis

I. Introduction

The taxi service in major cities is unbalanced. In certain regions, customers must wait an excessive amount of time for a cab, while in others, many taxis prowl the streets without clients. Determining where the cab should be at any given moment can benefit both the passenger and the firm. It will assist the organization in increasing profits and improving customer satisfaction. With the high demand for cab-hailing apps, the forms of public transportation have seen a drastic change over the last decade. These cab-hailing apps have created employment opportunities on a large scale. A large section of people considers this as a full-time job and it's their only primary source of income.

Such cab drivers would want a more intelligent tool to help them increase their efficiency throughout the day. They face the problem of not being present at the right location where they can have the probability of getting maximum rides. This causes wastage of fuel, time, and energy. By addressing this problem and helping cab drivers to be present at the right location at the right time, will not only help the cab drivers to gain more profit but also to reduce their costs.

Additionally, customers will have a shorter wait time and may not have to pay a surge price. Overall, it benefits the driver, customer, and the company itself. The aim is to perform a statistical analysis of the dataset and report interesting trends. This will include spatial and temporal analysis on the uber pickup dataset. Eventually, the aim is to build a recommendation engine based on different attributes to predict better pickup zones in NYC at any given time

II. Related Work

Ke et al. [1] discussed augmentation of the CNN-LSTM architecture to simulate short-term predicting demand and show the use of many convolutional-LSTM blocks that may include Spatio-temporal characteristics from various input modalities. The author tried to solve the problem of the demand-supply gap of ride-sourcing. The author divided the city into different hexagon lattices using hexagon-based convolutional neural networks and tried to incorporate the spatiotemporal properties in a hexagonal way. This approach was different from the previous approaches as earlier the city data was divided into square lattices and in this, the author divided the spatiotemporal data into hexagon-based region partition. The author used three different

approaches for hexagon-based regions i.e., Square H-CNN, Parity H-CNN, and Cube H-CNN. The parity H-CNN performs with the best accuracy out of all other machine learning algorithms for the hexagonal-based regions. It was also seen that when collaborative use of the three algorithms with the hexagon-based ensemble learning there was a significant improvement in the accuracy i.e., by 5.7%, 4.0%, and 2.3% in terms of RMSE for H-CNN, Parity H-CNN, and Cube H-CNN respectively.

This approach has various shortcomings as this technique was applied to the only type of dataset and it performed better than previous traditional machine learning algorithms but there is no theoretical proof that this approach of Hexagonal CNN LSTM architecture is better, but it performed well in their dataset. In addition, the author used three types of H-CNN considering the dimensionality and topology loss, but it was only applied to a small data set and there is no theoretical proof that this H-CNN performs better than the traditional methods.

Li et al [2] discussed various trends of spatial-temporal data of taxi scheduling based on the pickup and drop-offs. The taxi trajectory data was collected over one week. The different factors that are considered for the prediction are pick-ups, drop-offs, and the ratio of pick-ups to drop-offs, as well as the likelihood of pick-up and drop-off. The author has tried to find the spatial-temporal relationship for weekdays and weekends. The paper also has taken regional imbalances into account to scrutinize the likelihood of pick up and drop off. The Area Crossing Index (ACI) is used for calculating the probability that reflects the taxi cardinality and accessibility of a location. Point of interest(POI) and demographic data are employed as explanatory factors at the same time. The author has also considered the business hours of POIs. The author has used the hierarchical clustering technique to evaluate the similarity qualities of hourly dependent variables to investigate ridership in each hour. The authors have used stepwise linear regression to measure the collinearity and then used weighted regression on the dataset.

The findings suggest that there is a high likelihood of pickups and drops off on weekends compared to weekdays. On weekends, travellers find it difficult to get taxis near parks or residential areas, while it is considerably simpler in high-density areas such as restaurants and internal stations. As a result, they suggest that cabs should be routed to areas with a larger density of residential areas. The Probability of pick up and drop off is higher near restaurants, especially during dinner times. The researchers suggest that pick-ups should increase linearly with the increase in drop-offs at these locations, yet based on the ratio of pick-ups and drop-off, restaurants appear to be a source location. As a result, areas with a high concentration of restaurants require better cab management. This limitation of this study is that the data used for analyzing the trend was of only one week. Moreover, when the only points of interest and demographic data are analyzed, the suggested important factors may lack a scientific nature.

In [3] Divine Carson et al. have suggested appropriate strategies to predict potential pick-up and drop-off locations by addressing the two problems of Braess Paradox and Congestion & Vehicle clustering by resolving it through collective intelligence (COIN). The discussion on mobility optimization states most of the human interaction between home and work. To recognize the mobility patterns Variable-Order Markov Model and Partial Matching for prediction of the next location with accuracy ranging from 60%-81% is mentioned. The author uses different machine learning techniques to predict the best pick up location preference of the user

Chao Wang et al. [4] used Convolution Neural Network based deep learning techniques for accurate prediction of the ride-hailing trip. The CNN model accurately predicts taxi prediction in about 1 km of the zone every 10 minutes. The

model performed 30% better compared to LSTM. Along with the trip location and time data, the author also considered the weather data that was taken from World Weather Online that includes temperature, humidity, and weather conditions. The author used 6 frame kernels for the convolution layer. Two convolution layers followed by the max-pooling for each layer. The increase the nonlinear properties the author further uses Rectified Linear Unit (RELU) function which further goes into the flattening layer. The author implies 3 models for comparison instanton model, LSTM and CNN. It was observed that LSTM and CNN provide better predictions than the instanton model. The Mean Absolute Error ranged the lowest for CNN with 3.0616, LSTM with 3.1078 and Instanton with 4.6783. The training computation is then further examined with CNN showing 30% faster than LSTM despite having multiple parameters. CNN showed training time as 11.73s while LSTM showed as 17.59s. Hence the CNN model turned out to be a more promising model for 100 zones in 10 minutes.

In [1] Divine Carson et al. have suggested appropriate strategies to predict potential pick-up and drop-off locations by addressing the two problems of Braess Paradox and Congestion & Vehicle clustering by resolving it through collective intelligence (COIN). The discussion on mobility optimization states the majority of human interaction between home and work. To recognize the mobility patterns Variable-Order Markov Model and Partial Matching for prediction of next location with accuracy ranging from 60%-81% is mentioned. The author uses different machine learning techniques to predict the best pick up location preference of the user.

The author uses PCA analysis on the features that results in certain attributes being able to explain 55.9% record in variance while 5 attributes explained 73.7% of variance that contributed to selecting the most optimal features. The author considers three scenarios for model calibration. Predict drop-off area with only pick-up data- Neural network showed the least mean squared error. Also, it results out linear regression as a potential algorithm in predicting the drop-off areas within 13 blocks of radius

Predict drop-off area with partial information- Ensemble models like Random Forest and Adaboost predicted drop-off location more accurately within 1 block radius.

Pick-up point prediction after drop-off – Here Random Forest showed the highest R-Squared value of 0.934

In general, neural networks performed best for overall scenarios of pick-up and drop-off locations when provided with only pick up points. Ensemble models perform well in being able to predict both pick up and drop off points. The following results benefitted in various aspects that includes, Vehicle utilization is improved, and time becomes more efficient, reduces idle time once a trip ends and improves customer and driver experience. reduced transport costs less emissions.The paper proposed by the author gives a commendable result but lacks to widen the horizon on more about the implementation of neural networks and if more deep learning models like RNN and CNN could provide us with better results. The model proposed above fails to predict the accurate results for autonomous vehicles as with the rising demand for these vehicles the industry most probably see a boom in switching from manual cars to autonomous cars. The model could be further extended to be deployed on large number of areas with fairly accurate results. Also, it can further be extended to come up with a solid solution to the Braess paradox problem and consider mobility pattern making use of Markov models

Chao Wang et al. [2] used Convolution Neural Network base deep learning techniques for

accurate prediction of ride hailing trip. The data used is of China, Chengdu given by DiDi Chuxing. The CNN model accurately predicts taxi prediction in about 1 km of zone every 10 minutes. The model performed 30% better compared to LSTM. Along with the trip location and time data the author also considered the weather data that was taken from World Weather Online that includes temperature, humidity, and weather conditions.

The author used 6 frame kernels for the convolution layer. Two convolution layers followed by the max-pooling for each layer. To increase the nonlinear properties the author further uses Rectified Linear Unit (RELU) function which further goes into flattening layers. The author implies 3 models for comparison instanton model, LSTM and CNN. It was observed that LSTM and CNN provide better predictions than the instanton model. The Mean Absolute Error ranged the lowest for CNN with 3.0616, LSTM with 3.1078 and Instanton with 4.6783. The training computation is then further examined with CNN showing 30% faster than LSTM despite having multiple parameters. CNN showed training time as 11.73s while LSTM showed as 17.59s. Hence the CNN model turned out to be a more promising model for 100 zones in 10 minutes. The model could benefit in overall operation efficiency, fuel consumption, customer satisfaction, time management.

The overall performance of the model is appreciable but different features that affect taxi predictions like the calendar week, festive occasions, national holidays etc could be considered too for the more accurate results on different days of the year.

The model could further be extended to multi-step predictions that could be applied on autonomous vehicles. The model prediction was still satisfied when predicting the ride 60 minutes ahead. Furthermore, the fleet dispatching system can be developed based on the demand prediction using this model. Fleet management will save a lot of idle time and travel distance between rides.

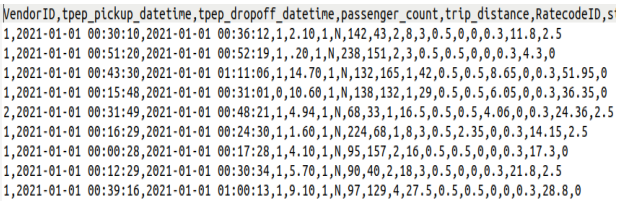
III. Experimental Evaluation

With a vision to make the life of taxi drivers and riders easy and efficient this project is designed to dig into deep analysis. of taxi demand considering time and location of the day. The evaluation is done on the basis of clustering the data and visualizing it into different regions.

A. Data Sources

The data collected for the evaluation is a seven month data set for the year 2021 that has been taken from yellow taxi data in New York city. The dataset consists of 16M data points. The feature dataset consists of Vendor Id,location id , pickup date and drop off date, pickup location, drop off location ,total fare, number of passengers, trip rating, etc. Out of these we take the pickup time, dropoff time, pickup location, dropoff location, and total fare has been taken into account to analyse important trends

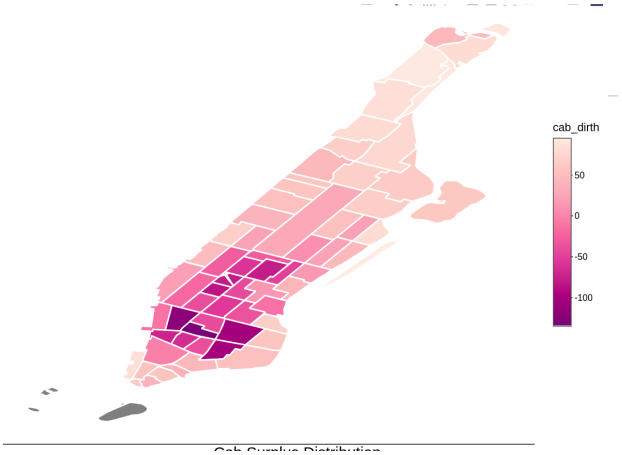
Fig. 1. Image of the Dataset

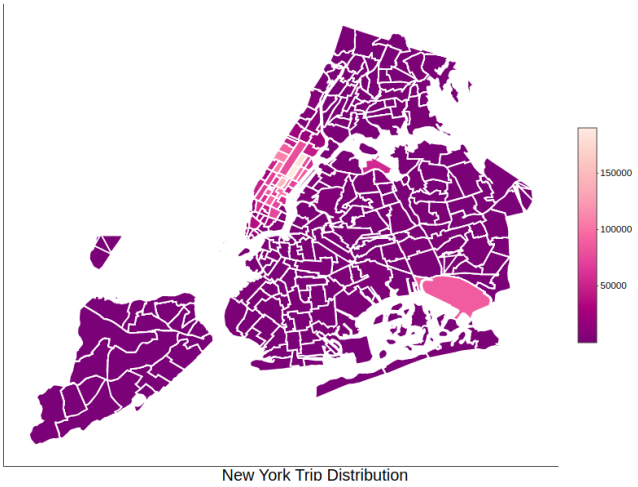


```
VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeID,s
1,2021-01-01 00:30:10,2021-01-01 00:36:12,1,2.10,1,N,142,43,2,8,3,0.5,0,0,0.3,11.8,2.5
1,2021-01-01 00:51:20,2021-01-01 00:52:19,1,.20,1,N,238,151,2,3,0.5,0.5,0,0,0.3,4.3,0
1,2021-01-01 00:43:30,2021-01-01 01:11:06,1,14.70,1,N,132,165,1,42,0.5,0.5,8.65,0,0.3,51.95,0
1,2021-01-01 00:15:48,2021-01-01 00:31:01,0,10.60,1,N,138,132,1,29,0.5,0.5,6.05,0,0.3,36.35,0
2,2021-01-01 00:31:49,2021-01-01 00:48:21,1,4.94,1,N,68,33,1,16.5,0.5,0.5,4.06,0,0.3,24.36,2.5
1,2021-01-01 00:16:29,2021-01-01 00:24:30,1,1.60,1,N,224,68,1,8,3,0.5,2.35,0,0.3,14.15,2.5
1,2021-01-01 00:00:28,2021-01-01 00:17:28,1,4.10,1,N,95,157,2,16,0.5,0.5,0,0,0.3,17.3,0
1,2021-01-01 00:12:29,2021-01-01 00:30:34,1,5.70,1,N,90,40,2,18,3,0.5,0,0,0.3,21.8,2.5
1,2021-01-01 00:39:16,2021-01-01 01:00:13,1,9.10,1,N,97,129,4,27.5,0.5,0.5,0,0,0.3,28.8,0
```

B. Big Data Processing

Big data processing is a set of techniques or programming paradigms for accessing data at scale in order to extract useful information to support and make decisions. Platforms like Hadoop, Spark. Hive is the most commonly used technology to achieve this. Hadoop is used to reduce the dataset via the MapReduce methodology to extract useful aggregated information. The Map and Reduce functions are user programmed to handle large data distributed over a heterogeneous number of nodes.





New York Trip Distribution

To allow for easier processing of data, we convert couple of the attributes in the dataset:

- Pickup Time and Location Time is divided into eight slots of 3 hours each in a 24 hour day. Each slot is represented as an integer 0-7. The numbers represent 12am to 3am, 3am to 6am, 6am to 9am, 9am to 12pm etc. slots respectively (clockwise).

- The city of New York is divided into 265 different zones, Each zone is uniquely represented by an integer between 1-265.

We use hadoop to perform the following aggregation :-

- Pickup Aggregation: The pickup aggregated data aggregates the number of trips for a unique Pickup location ID and time slot tuple.

- Dropoff Aggregation: The pickup aggregated data aggregates the number of trips for a unique Drop Off Location ID and time slot tuple.

- Pickup vs Drop Off: This aggregates all trips for a unique permutation of Pickup Location ID, Drop Off Location ID and Pickup Time Slot.

C. Data Visualization

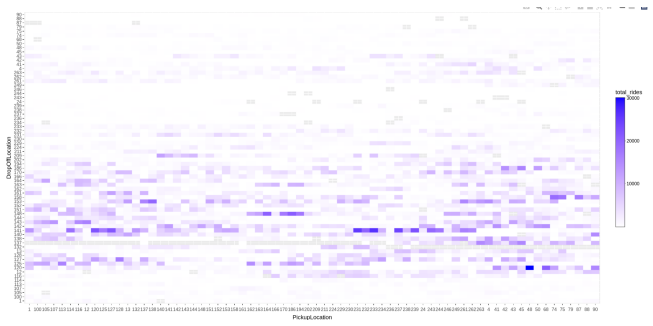
Data Visualization was implemented in R as it provides a versatile set of tools for creating plots and charts. Initially we visualized the distribution of trips over different Pickup and Dropoff locations. More than 92% data was found to be concentrated in the Manhattan region along with the airports surrounding that region. As a result, we focus our analysis for zones in this region (~95 zones).

IV. Data Analysis

We compared the dropoff and the pickup for each zone in our region of interest. Assuming a very low number of empty cabs, we assume cab dearth

in a zone if it has a higher number of drop offs as compared to the number of pickups. Conversely, we assume a cab surplus if the number of pickups is greater than the number of pickups. Based on this we found some interesting trends.

- East Village had the maximum number of pickups/dropoffs and cab dearth between



- 9pm and 3 am. This is owing to the fact that East village is a central hub of happening nightlife in Manhattan.
- Between 6am to 9am Penn Station has a maximum number of pickups and faces a large dearth of cabs. It houses one of the busiest rail stations and connects Midtown Manhattan to the surrounding regions. Consequently, there is a large amount of traffic during office hours.
- There are a large number of taxi pickups in Upper East side north but pickup requests get catered fairly during the same time i.e. there is almost no cab dearth.
- Despite Penn Station and Upper East Village receiving a similar number of pickup requests, there were lesser chances of competition (and more chance of pickup) in Penn Station due to higher cab dearth.
- Midtown Centre being the tourist attraction was a hub for all drop offs
- During 9am to 12pm, despite upper east side north and south having maximum pickups, it seemed to be doing fine because of the large influx of cabs in the area as people travel between the two counties
- There was a spike in pickup requests near Wall street around 12pm-3pm (due to it being lunch time).
- At 3pm to 6pm a lot of pickup is observed at Midtown as people go to their homes from their workplaces.
- During 6pm to 9pm there is spike in number of drops at East and West Village due to its social nightlife while maintaining a dearth in Midtown
- A huge spike of pickups is observed in East village from 9pm to 12am as people tend to home after social gatherings and nightlife

Following this we performed some preliminary analysis on the data and found a large number of pickups are in between 12pm-6pm. We also found the rides paying the highest amount usually emanated from the airports.

Following this we performed a heatmap plot between different pickup and dropoff locations. We were able to verify the above observations using the heatmap too.

Acknowledgment

The authors are thankful to Professor Mike Wu for providing the necessary guidance and support.

References

[1] Ke, Jintao, Hai Yang, Hongyu Zheng, Xiqun Chen, Yitian Jia, Pinghua Gong, and Jieping Ye. "Hexagon-based convolutional neural network for supply-demand forecasting of ride-sourcing services." IEEE Transactions on Intelligent Transportation Systems 20, no. 11 (2018): 4160-4173.

[2] Li, Bozhao, Zhongliang Cai, Lili Jiang, Shiliang Su, and Xinran Huang. "Exploring urban taxi ridership and local associated factors using GPS data and geographically weighted regression." Cities 87 (2019): 68-86.

[3] Carson-Bell, Divine, Mawutor Adadevoh-Beckley, and Kendra Kaitoo. "Demand Prediction of Ride-Hailing Pick-Up Location Using Ensemble Learning Methods." Journal of Transportation Technologies 11, no. 2 (2021): 250-264.

[4] Wang, Chao, Yi Hou, and Matthew Barth. "Data-driven multi-step demand prediction for ride-hailing services using convolutional neural networks." In Science and Information Conference, pp. 11-22. Springer, Cham, 2019.