International Journal of Pure and Applied Mathematics
Volume 118 No. 24 2018
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

Recommendation System to improve the Availability of Cabs

Padma Lakshmi G[1], Hema N[2]
VIT, Chennai, Tamil Nadu, India
padmalakshmi.g2016@vitstudent.ac.in[1],
hema.n@vit.ac.in[2]

May 29, 2018

Abstract

One of the major difficulties faced by people commuting across the city is transportation. Few have their own transportation whereas others rely either on public transport or applications that provide transportation facilities. This paper focuses on building a system for improving the cab availability based on the New York taxi trip data set that consist of around 16 lakhs of records related to the pickup location, drop off location, fare paid and so on. This is done by clustering the coordinates according to the areas and then predicting the number of bookings that are made for any given locality. The results are classified for bookings that will be made for weekdays or weekends. The overall process is carried with python programming language. Finally the results are visualized with the help of graphs.

***Index Terms*:**ARIMA, Fore casting, Normalization, Predictive analysis Pre Processing, Time Series

1 INTROUDCTION

Commuting from one locality to another is one of the problems faced by people travelling across the city. Earlier getting cabs for short distances in a city or to go to the airport from a locality or

to travel from the airport to a locality was very inconvenient or difficult. Economic growth along with infrastructure development has made the car rental industry grow in the past years. The urban middle class population, with their increasing income (salary) has raised the need and willingness to pay for good services which includes private transportation.

Taxi services one among the transport facilities makes their journey comfortable. Even though it comes with so much ease there are some problems in existing taxi aggregators too, such as they suffer from unreasonable taxi distribution, high load rate and low dispatching efficiency. Hence People who rely on those applications for transportation facilities will face problem during the peak hours.

So to improve the Quality of service provided by cab companies forecasting is done. Forecasting the future allocation of cabs is largely and widely studied topic in many fields like trading, finance, statistics and even in computer science. Fundamental or technical analysis has been used by professional company to analyse and predict the cab allocation. Fundamental analysis is the traditional way of approaching and involving a study of the cab trip details such as trip duration, pick up location , drop off etc. on the other hand technical analysis is the study of past cab trips. Technical analyst takes the trip charts and analyses the pattern on the price paid then process to get conclusions.

Recent study on forecasting the cab allocation prediction is gaining more attention. It is because of the fact that if the movement of the cabs is predicted successfully the companys decision making will also move in the same direction. Any system which we develop constantly monitor and predict the allocation of the cabs would make the taxi aggregator company wealthier. It also predicts trends of the cab booking thus helping the company and drivers of the market.

Time series forecasting with the help of the past historical data of bookings to the company, is done by selecting a prediction variable. With the help of this variable a model is created and used to forecast the future cab availability. Time series data is a set of well-structured data items that is collected at particular points in uniform time frames. In marketing industry, forecasting is an important part. Auto regression and moving Average (ARMA) model is an extensively used method which helps in studying of time series

data. ARIMA model is generally used to predict linear time series data.

To provide the best service in the market the company has to think intelligently and provide the supply of cabs wherever it is needed at the right time. Simply allocating cabs around the city doesnt give the company the turn over expected. . The data set is taken form NYC Taxi and Limousine Commission (TLC). So to analyse the historical record data and supply cabs based on it is the need of the hour. Here we have build a system that analyses the past records and predicts where the resulting output values is derived from the pickup latitude and longitude data. The output derived is an allocation strategy that can be used to select and send cabs to a locality for a given hour depending on whether on weekday or weekend.

2 RELATED WORK

In this paper [5] Chenguang Zhu1 and Balaji address these inefficiencies and made out a model for the assigning to drivers based on a network flow. Taxi systems are considered to be the one of the best examples for supply demand system as the profit of the organization solely depends. Over here the supply side consists of cabs and the drivers whereas the demand is made by the passengers who travel by booking cabs. So such a big system is prone to have some inefficiencies like more taxis than needed, less cabs when needed, no passenger booking for long time, wrong route navigation and so on.

This novel model assigns trips to the cabs on a shift basis i.e 6 or 12 hours. Statistics are compared from taxi assignments from real world data, minimum feasible flow model (FF) and minimum cost feasible flow model (MCFF) in the experimenting phase. The model is aimed at finishing the maximum number of trips the driver can based on his location at any point of time and at the same time reducing the idle time.

The advantage of the work is that it is very scalable and is able to give detailed assignment plans of as many trips (hundreds or even thousands). On evaluating the effectiveness of the model proposed with respect to the New York city taxi trip data it was found that

only 72% of the existing yellow taxi cabs shall be required to finish off all the trips. Moreover the it was found that the average waiting time of the cabs got reduced by 32%. One of the drawbacks is that the time complexity of the models is based on the length of time bins as the number of nodes in the flow network increases sequentially with the increases in number of time bins.

Ferreira et al. [6] used the New York taxi trip data as a sensor for the moments across the locations within the city . Their primary goal was on bringing up a system with complex queries so that it gives more useful visualization results than the usual analytics queries produce. Because of the size of the data, they built a system of storage, querying and results that allows interactive plotting of several visualizations. The system that they had created can produce heat maps, plots and other better visualizations than the ordinary ones. A specific example explained was a comparative heat map which shows the impacts of Hurricanes Sandy and Irene on cab ridership in the city.

Umang Patel [10] uses several big data techniques to analyze the vast New York taxi trip dataset that consists of about 180 million taxi ride details to avoid traffic routes, lower rate during less frequency in uncrowned areas and so on. The highlight of this paper is that it can effortlessly handle the huge volumes of data using big data techniques. Under the analysis of individuals the problems that are solved are: finding the driver who has travelled long distance, driver who has highest fare collected, driver with highest time travel and finding the best driver based on time taken and distance covered. All the individual analysis help in distinguishing the drivers who are able to cope up with the demand and the ones who are struggling.

Under analysis of region, map reduce programming is used to find the most pick up and drop off location. This gives us an insight that we can lower down the number of cabs in places where there is more drop off.

Next the analysis based on time and location where Hive is used to determine the average total pick up and drop off times based on location. Finally analysis based on the fare paid for the ride is doe using pig to determine the average income per hour for a driver is found. Here only few queries are only executed, using more complex queries will provide us better insight from the data.

Paper [4] proposes to predict the cab pick up time in the busy places of New York using the data available from Limousine Commission and Twitter. The advantage here is that machine learning algorithms are used to evaluate predictability so that they are accurate.

The first step is to refine the data so that it is suitable for analysis i.e data cleaning and preparation. A database of size 200 GB was generated as output. To predict the exact pick up time the data for each of the neighborhood is manipulated to single, hour long instance. The result was 60 million instances extracted for 29 neighborhoods. The final modelling process predicts the pickup count to check if the model is reliable. The features were trained with random forest regression model to find the best subset. Then a graph is found based on the mean absolute error (MAE) and neighborhood to see the prediction variation based on the prediction features. Then with the other time series graph of hourly pickups, tweets, and predicted pickups, the system is almost able to give an accurate result for Harlem Neighborhood.

To make the model much more efficient more factors such as climatic conditions, scheduled events etc are to be considered. More over only the summer months data was used here, applying the model to full calendar years data show the real accuracy of the system.

Yi-Shian Lee et [7] al utilized a numerous fields to construct models for determining time arrangement. In spite of the fact that ARIMA can be embracing to acquire an exceptionally precise straight determining show, it can't precisely gauge non straight time arrangement. Simulated neural system (ANN) can be used to fabricate more precise estimating model than ARIMA for nonlinear time arrangement, however clarification the significance of the shrouded layers of ANN is hard and, in addition, it doesn't yield a scientific condition. This amendment proposes a blend determining model for nonlinear time arrangement by join ARIMA with hereditary programming (GP) to enhance both the ANN and the ARIMA determining model. At long last, some genuine informational indexes are received to demonstrate the effectiveness of the proposed anticipating model.

3 DATA

The dataset for the study is to learn and evaluate the performance of the implemented system includes daily bookings available from the NYC Taxi and Limousine Commission (TLC) . The data is download from http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

The historical data of taxi trip consist of time series data. Anything that is related to time comes under in the category of time series data. Time series is a sequences of well-defined data points that are measured at a particular time interval over a time period. Time series analysis use statistical methods that are used to analyze a series of data and that are extracted to a meaningful statistics and characteristics about the data.

It came in the form of comma separated value (CSV) files for each month consisting of millions of records about each trip. The file consisted of the following fields namely:

- VendorID
- lpep_pickup_datetime
- Lpep_dropoff_datetime
- Store_and_fwd_flag
- RateCodeID
- Pickup_longitude
- Pickup_latitude
- Dropoff_longitude
- Dropoff_latitude
- Passenger_count
- Trip_distance
- Fare_amount
- Extra
- MTA_tax

- Tip_amount
- Tolls_amount
- improvement_surcharge
- Total_amount,
- Payment_type
- Trip_type

4 OVERALL PROCESS

This module describes how data flows from one module into another. Here we give a picture of what happens in the module. The data flows in the following order:

1. Data loading and Pre-processing
2. Data processing
3. Data Analytics
4. Visualization

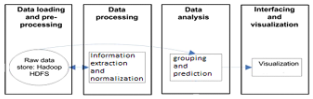

Fig .1 System Architecture

4.1 Data loading

The data needs to be uploaded in HDFS so that it can be further clustered. After that it is taken for data pre-processing which is the preliminary step in the data mining process. The process of "garbage in, garbage out" is performed here. Data-gathering methods are mostly less controlled, hence results in out-of-range values, impossible data combinations, missing values and so on. Analyzing data that is not been cleansed for problems can produce misleading results. Thus, the representation and quality of data is first and foremost step before running an analysis.

4.2 Data Processing and Normalization

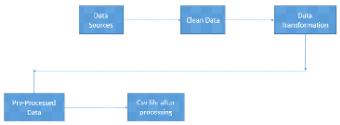

Fig 2 Navigation From Source To Destination

The primary goals of the data cleaning and preparation process were as follows:

- Clean the data by e.g. remove rows with missing values
- Create a structured data table to allow us to complete the analysis
- Extract data from the database and populate table

After the cleansing of the data it is taken for further processing. In order to find the locality from which the height number of booking occur at any given point of time in the day we have normalize some of the fields. We find that the column lpep_pickup_datetime in the dataset has the time and date of booking the cab clubbed together. To carry out our analysis we need all of them separated and have to determine whether its a weekend or weekend as the trend of booking changes accordingly.

The new features added by extracting from the existing column and for further processing are:

- Week

As we are going to take into account the week number which the bookings are high we add up here a column which states the week number it is in the month.

- Week_day

This field help us in identifying whether it is a weekday or weekend. We number days like: 1 for Monday, 2 for Tuesday and son on.

- Month_day

This column represents the date in the month. For eg: 1 30 or 31 depending on the month considered.

- Hour

The hour in which the cab got booked is extracted and stored in a separate column.

• Day type

In this column we classify if it is a weekday or weekend day and then mark it is a 1 for weekday and 2 for weekend

Some of the attributes that are not needed for the processing like 'Store_and_fwd_flag', 'RateCodeID', 'Passenger_count', 'Tolls_amount', 'Tip_amount', 'Ehail_fee' and 'improvement_surcharge' were removed so that the data is completely ready for further analysis.

4.3 Clustering Process

For a clustering data, the process is subjective it can be achieved by plenty of ways. Each of the clustering algorithm or methodology follows some similarity measure to cluster the data. There are many famous iterative clustering algorithms which determine the cluster based on their distance or say closeness to the centroid. These models help can even group the data and give the clusters based on the number of required.

But in this system, we have to cluster based on the booking pick up latitude and longitude. Because bookings that come from the same area or location needs to be grouped together. There cannot be only one latitude and longitude values for bookings that fall under a particular location. So to group them together one of the approaches can be made:

1) We need to find the list of latitude and longitude combinations which fall under the same locality. Then group them to form a cluster

2) Built a polygon with the latitude and longitude values for that area if the given latitude and longitude value fall under that polygon then add them into the cluster.

Here the second method is adopted and a polygon is built for the localities where we find bookings occur. After forming the polygons for each locality we need to check record by record for the dataset whether the (pickup latitude, pickup longitude) fall under the area of the polygon of any the locality.

Here is the graph for the localities we have considered which represents the booked that occur for the localities at a given point of time in a day.

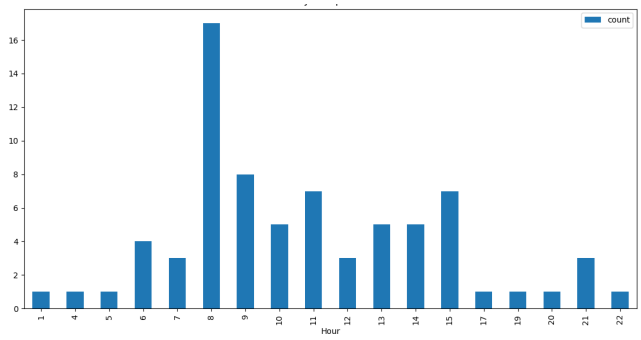

Fig 3 trip count for JFK airport

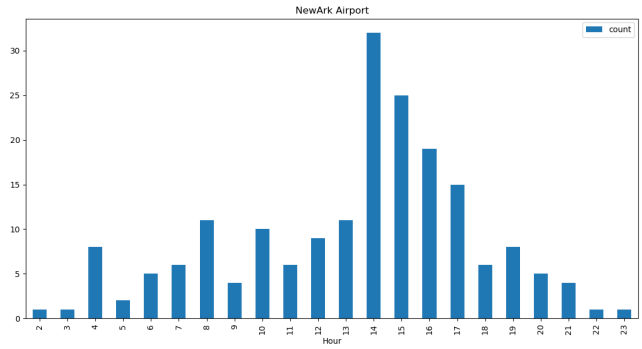

Fig 4 trip count for Newark Airport

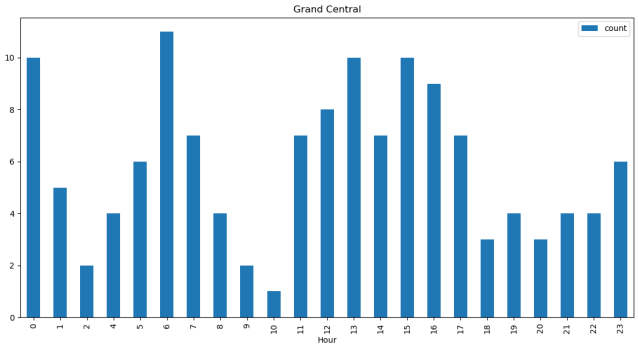

Fig 5 Trip count for grand central

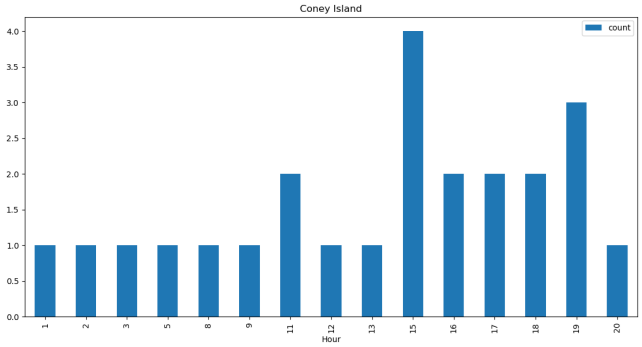

Fig 6 Trip count for Cooney Island

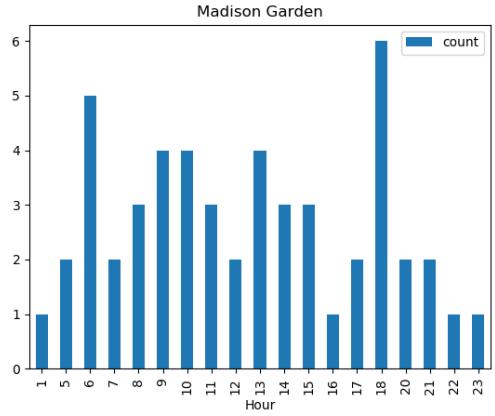

Fig 7 Trip count for Madison Garden

4.4 Prediction Process

After grouping the cab booking based on the areas they are booked we move to do the final process of prediction. This process consists of many statistical techniques where it analyses the current and historical information to make prediction regarding the future or reveal out some interesting or unknown events. We have omitted JFK airport locality to be included in the prediction as the trip count for it various tremendously.

Here we have used SARIMAX algorithm to then forecast the cab allocation prediction. An SARIMAX model is a linear regression model that uses a SARIMA type process. This model is useful in

cases we suspect that residuals may exhibit a pattern or seasonal trend.

$$w_t = y_t - \rho_1 x_{1,t} - \rho_2 x_{2,t} - \cdots - \rho_b x_{b,t}$$

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - \sum_{j=1}^{p} \phi_j L^{j \times s})(1 - L^s)^D w_t - \eta$$

$$= (1 + \sum_{i=1}^{q} \theta_i L^i)(1 + \sum_{j=1}^{Q} \theta_j L^{j \times s}) a_t$$

$$a_t \sim 1.1.d \sim \Phi(0, \sigma^2)$$

- L is the lagi (aka back-shift) operator.
- y_t is the observed output at time t.
- $x_{k,t}$ is the k-th exogenous input variable at time t.
- is the coefficient value for the k-th exogenous (explanatory) input variable.
- β_k is the number of exogenous input variables.
- b is the number of exogenous input variables.
- w_t is the auto-correlated regression residuals.
- p is the order of the non-seasonal AR component.
- P is the order of the seasonal AR component.
- q is the order of the non-seasonal MA component.
- Q is the order of the seasonal MA component.
- s is the seasonal length.
- D is the seasonal integration order of the time series.
- η is a constant in the SARIMA model
- a_t is the innovation, shock or error term at time t.

- $\{a_t\}$ time series observations are independent and identically distributed (i.e. i.i.d) and follow a Gaussian distribution (i.e.$(0, \sigma^2$)

Re-ordering the terms in the equation above and assuming the differenced (both seasonal and non-seasonal) results in a stationary time series (z_t) yields the following:

$$z_t = (1 - L)^d (1 - L^s)^D w_t$$

$$\mu = E[z_t] = \frac{\eta}{(1 - \phi_1 - \phi_2 - \cdots - \phi_p)(1 - \Phi_1 - \Phi_2 - \cdots - \Phi_p)}$$

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - \sum_{j=1}^{p} \phi_j L^{j \times s})(1 - L^s)^D w_t - \mu$$

$$= (1 + \sum_{i=1}^{q} \theta_i L^i)(1 + \sum_{j=1}^{Q} \theta_j L^{j \times s}) a_t$$

1. The variance of the shocks is constant or time-invariant.
2. The order of an AR component process is solely determined by the order of the last lagged auto-regressive variable with a non-zero coefficient (i.e. w_{t-p}).
3. The order of an MA component process is solely determined by the order of the last moving average variable with a non-zero coefficient (i.e.a_{t-p}).
4. In principle, you can have fewer parameters than the orders of the model.

With the help of this model, we have done the cab allocation prediction for some of the localities considered for the experiment and visualized them in the form a graph. This graph gives a comparative view of the allocation that shall be made proportionally.

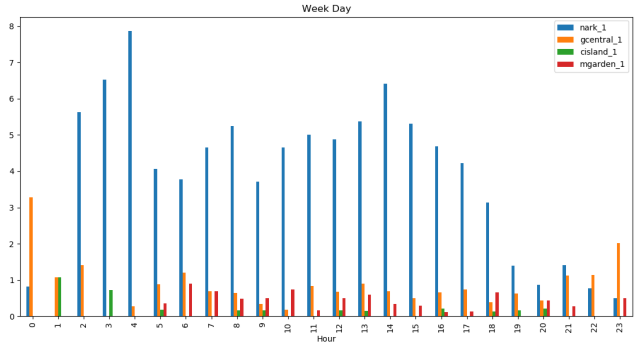

Fig 8 Cab allocation for Weekdays

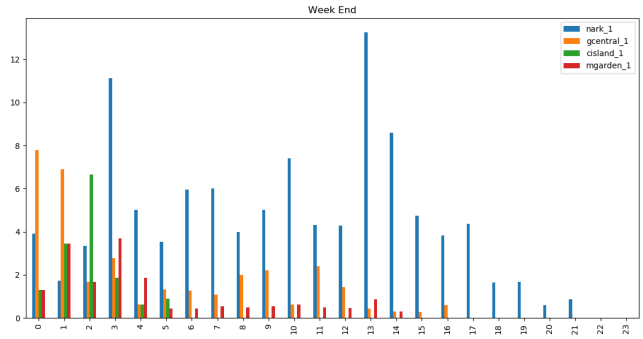

Fig 9 Cab allocation for Weekends

5 CONCLUSION AND FUTURE WORK

The prediction approach used in the system seems to be promising. We developed polygons with the latitude and longitude points to find out the bookings that fall under a certain locality.

There are numerous questions for further line of research in the topic. First, is the approach used scalable? Are we able to generalize our predictive model considered with 6 different localities in New York. If so then how can we estimate or predict the demand from these taxi stands? Solving this problem has the potential to better match taxis to customers, thereby saving fuel, reducing the number of taxis required to serve the city and increasing the overall quality of service provided by the taxi Aggregator system. Second

is that can our model be extended to other cities too. If so then how it behave in that new environment is to be analysed.

Our research is a first step towards a real time system to balance the supply of taxis at some popular localities in New York. Our future work would be to include public holidays such as Independence day, Christmas, New Year etc. and then do allocation based on the past booking history of those days. Providing adequate transportation to passengers is a problem faced worldwide, and we aim to implement our methods and algorithms in a commercial system to meet the demand.

References

[1] Gerardo Berbeglia, Jean-Francois Cordeau, Irina Gribkovskaia, and Gilbert Laporte. Static pickup and delivery problems: a classification scheme and survey. Top, 15(1):131, 2007.

[2] Jean-Franccois Cordeau, Gilbert Laporte, and Stefan Ropke. Recent models and algorithms for one-to-one pickup and delivery problems. In The vehicle routing problem: latest advances and new challenges, pages 327357. Springer, 2008.

[3] Vincent P Crawford and Juanjuan Meng. New york city cab drivers labor supply revisited: Reference-dependent preferences with rationalexpectations targets for hours and income. American Economic Review, 101(5):191232, 2011.

[4] Joya A Deri, Franz Franchetti, and Jose MF Moura. Big data computation of taxi movement in new york city. In Big Data (Big Data), 2016 IEEE International Conference on, pages 26162625. IEEE, 2016.

[5] Chenguang Zhu and Balaji Prabhakar. Reducing inefficiencies in taxi systems. In Decision and Control (CDC), 2017 IEEE 56th Annual Conference on, pages 63016306. IEEE, 2017.

[6] Nivan Ferreira, Jorge Poco, Huy T Vo, Juliana Freire, and Claudio T Silva.Visual exploration of big spatio-temporal urban data: A study of new York city taxi trips.

IEEE Transactions on Visualization and Computer Graphics,19(12):21492158, 2013.

[7] Yi-Shian Lee and Lee-Ing Tong. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. Knowledge-Based Systems, 24(1):6672, 2011.

[8] Eleanor A Maguire, Richard SJ Frackowiak, and Christopher D Frith. Recalling routes around london: activation of the right hippocampus in taxi drivers. Journal of neuroscience, 17(18):71037110, 1997.

[9] Martijn Mes, Matthieu van der Heijden, and Peter Schuur. Look-ahead strategies for dynamic pickup and delivery problems. OR spectrum, 32(2):395421, 2010.

[10] Umang Patel and Anil Chandan. Nyc taxi trip and fare data analytics using bigdata. Retrieved June, 9:2017, 2010.

[11] Austin W Smith, Andrew L Kun, and John Krumm. Predicting taxi pickups in cities: which data sources should we use? In Proceedings of the 2017ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium onWearable Computers, pages 380387. ACM, 2017.

[12] Charles Vidich. New York Cab Driver and His Fare. Routledge, 2017.

[13] Jian Yang, Patrick Jaillet, and Hani Mahmassani. Real-time multivehicle truckload pickup and delivery problems. Transportation Science, 38(2):135148, 2004.

[14] M Anil Yazici, Camille Kamga, and Abhishek Singhal. A big data driven model for taxi drivers airport pick-up decisions in new york city. In Big Data, 2013 IEEE International Conference on, pages 3744. IEEE, 2013.

[15] Wei Yuan, Pan Deng, Tarik Taleb, JiafuWan, and Chaofan Bi. An unlicensed taxi identification model based on big data analysis. IEEE Transactions on Intelligent Transportation Systems, 17(6):17031713, 2016.

[16] Henry S Farber. Reference-dependent preferences and labor supply: The case of new york city taxi drivers. American Economic Review, 98(3):106982, 2008.