NYC Green Taxis with Neural Network

Brittany Nicholls
Department of Computer Science
and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, Maryland 21250
Email: brn1@umbc.edu

***Abstract*—NYC Green TAxi. Use Neural Net. Kaggle Competition. RMSE vs RMSLE**

I. Introduction

This paper explores using a neural network in order to predict the duration of a taxi ride for Green Taxis in New York City. The model is built only using information that would be available at the start of the ride which means that the trip route is not known.

Most people are familiar with the iconic New York Yellow Taxis; however, there are a second type of taxis in New York City called Green taxis which might also be referenced as Boro taxis. Green taxis were implemented in 2013 after the government realized that 95% of Yellow taxi pickups were in central Manhattan, LaGuardia Airport, and JFK Airport [1]. This meant that a large subsection of the population was underserved by the existing taxi services. In response to this discovery, the local government decided to issue licenses for Green Taxis, and as a compromise to the existing Yellow taxis, the Green taxis would specifically only pickup people from anywhere except in the exclusionary zone [1]. You can see a map of the green and yellow zones in Figure 1.

Figure 1 shows us that the rides for Green taxis are very different than the rides of Yellow taxis. Yellow taxis for the most part stay around Manhattan, while the Green taxis are all over the city. This means that, for example, the way the drivers find passengers, the traffic the drivers deal with, the routes that drivers take and the number of stoplights the drivers deal with are different between Yellow taxis and Green taxis. In addition, the number of trips that Green taxis and Yellow taxis make in a day are very different. Doing a basic comparison between the 2016 taxi datasets available at NYC Open Data, we see that there are 16.4 million records for Green taxis from January 1, 2016 to June 30, 2016 [3] while there are 133 million records for Yellow taxis for the whole year of 2016 [4].

There has been previous work on these datasets which has focused on both the datasets for Yellow and Green taxis; however, given the obvious differences in taxi trip details as well as data size, this report documents my attempt to model trip duration for only the Green taxi data. In this case, we found that using a basic feed-forward neural net to minmize the root mean squared error (RMSE) of the rounded duration, in minutes, was optimal. Furthermore, we explore why RMSE

Fig. 1. Map of NYC with the green taxi pickup areas colored in green. [2]

was chosen as the loss function instead of the root mean squared log error (RMSLE). ¡AND THE ENSEMBLE¿

A. Motivation

Originally, I was going to try to work on the problem of taxi trip duration prediction introduced in a Kaggle challenge [5]. However, as I explored existing work surrounding this problem, I realized that when most papers look at the NYC taxi data, they either combine Yellow and Green taxi data [6], or only focus on Yellow taxis [5][7]. Not to mention, most work available has been done by students or through kaggle.

Predicting taxi trip duration is important for multiple reasons. First, trip duration can be used by a customer to schedule their ride or to estimate the fare for a trip. Second, it could allow taxi companies to start making, or improve, their ability to chain together rides to increase profit for a driver. Lastly, having a baseline for what the expected trip duration would allow taxi companies to detect when drivers are taking their passengers on the ”scenic route” in order to charge the passengers more.

II. Related Work

As I mentioned above, the work on this dataset has been done informally through school projects or Kaggle competitions; however, there is other work related to taxi data.

A. Existing Work on NYC Taxi Trip Duration Prediction

In July 2017, Kaggle issued a challenge to Kagglers to work on a pre-selected NYC taxi trip dataset in order to predict the trip duration; however, the focus of the competition was to encourage collaboration, so the top publically available kernels are focused on exploring the data in a clear way to benefit the group [5]. In addition, the few publically available codebases were not from top performers [8] [9] [10]. Since it was still a competition to predict the trip duration, Kaggle still provided the means for evaluating the models using RMSLE. After evaluating this loss function, I decided to go with RMSE as my loss function instead. I will go into more detail in section VI-A.

Additionally, other students have used this dataset to predict trip duration and made their work publically available; however, none of them used neural nets, which are the new "hot" thing to do because of their ability to fit to a lot of datasets fairly well. In one paper, they used linear regression and random forests in order to build their model to predict the number of minutes the trip will take and achieved a low RMSE of 5.24 [11]. Another project achieved an RMSE of 4.87 by using a gradient boosting regressor [7]. I used the second paper that achieved a RMSE of 4.87 as a baseline, which will be discussed in section V.

B. Other Work on Taxi Data

Ferreira et. al. have also used the NYC Taxi data in order to discuss how to best visualize and query datasets such as the NYC taxi data [12]. While their work is interesting and does use the NYC taxi data, it does not attempt to make any predictions on the trip duration.

It should be noted that in 2015 there was another Kaggle competition to predict the trip duration for taxi data from Porto, Portugal [13]. One of the top ranking results were published by a group from IBM in which they discuss how they would predict the final destination based on the beginning trip trajectory and would use that to predict the trip duration [14]. However, they were predicting the trip duration based on the initial route of the taxi after picking up a customer, which contains very different data than is available for the NYC taxi data.

C. Other Work on Trip Duration Prediction

Work has been done to predict when a bus will arrive at its stop using an algorithm based on Kalman Filtering [15]; however, their work was focused on data collected in India and they needed to be able to analyze the data real time. First, traffic in India is probably different from traffic in NYC. Second, bus routes and stops are well defined while taxi routes and start/end locations are not pre-determined until a customer gets into a taxi.

More work on bus arrival prediction has been done by Biagoni et. al. in which they use a smartphone that is placed on a bus to predict when the bus will arrive at its next stop [16]. Once again, this is bus data and relies on real time data for a specific route with specific stops.

Additionally, there has been quite a bit of work on predicting travel time on freeways [17] [18]. However, driving in a city is very different from driving on a freeway.

III. Data and Data Challenges

A. NYC Open Data

New York City made the data for January 2016 - June 2016 of the trip records for the Green taxi data publically available [3]. There are approximately 16 million rows. This data contains the following information:

- Vendor ID : One of two vendors
- Pickup and Dropoff Date time: Time and Date when customer was picked up and dropped off
- RateCode ID: Rate (and the code) change based on where the customer is going
- Pickup and Dropoff Location: Contains the lat/long for both of these locations
- Passenger count: Driver recorded number of passengers in the vehicle for the ride
- Information about the fare: Base rate, taxes, tips

I enriched this dataset with two other datasets: weather and sunrise/sunset times.

B. Weather

Since traffic and taxi use can be heavily influenced by weather, I pulled down some basic stats about the weather for the taxi data being analyzed. The weather for April and March was retrieved from a site run by NOAA [19]. There was one record per day containing:

- Minimum/Maximum/Average temperature
- Precipitation: The amount of rain that fell. Note that a value of T stands for trace
- New Snow: The amount of new snow that fell
- Snow depth: How deep the snow was

Note that both the New Snow and Precipitation columns contained a few values of "T", which stands for trace. This means that very little snow/rain fell and it was not able to be measured. Records that had a value of "T" were assigned the value of 0.00001 based on a recommendation in the Journal of Service Climatology [20].

C. Sunrise/Sunset

Lastly, the taxi data was enriched with sunrise and sunset information. Note that the sunrise and sunset can affect traffic because the sun may get into people's eyes, causing them to slow down. The sunset and sunrise times were retrieved from a site owned by the US Navy [21].

IV. Features Used

V. Baseline Model

VI. Neural Network

A. Loss Function

I debated between using two loss functions, RMSLE and RMSE. Utlimately, I choose to use RMSE and will explore why in the following paragraph.

Kaggle decided to use RMSLE:

$\epsilon = \sqrt{\frac{1}{n}\sum (log(p_i + 1) - log(a_i + 1))^2}$

with a little bit of manipulation of the logs, this transforms into:

$\epsilon = \sqrt{\frac{1}{n}\sum log(\frac{p_i+1}{a_i+1})^2}$

This shows that, in essence, the model is heavily punished when $p_i < a_i$ as compared to when $p_i > a_i$. This means that this model will try to avoid predicting under the actual value, but the loss isn't as severe if it goes over. Thus, minimizing the RMSLE loss is better for finding a model that will predict the minimum trip duration. Furthermore, because of it's use of the log function, RMSLE would be good for numbers that are going to be large. For instance, if we were to predict the trip duration in milliseconds or seconds, then RMSLE might be a better choice; however, I choose to focus on predicting the trip duration in minutes, not seconds.

RMSE is calculated as follows:

$\epsilon = \sqrt{\frac{1}{n}\sum (p_i - a_i)^2}$

which shows that the model will get equally punished for predicting too large of a value as it will too small of a value. Thus, I choose to minimize the RMSE when building my model.

B. Neural Network Settings

C. Testing

D. Transfer Learning

Attempt on March Data

VII. Future Work and Ideas

Combining results from models trained on different loss functions? Additional features. Model to predict route. Geo clustering.

VIII. Conclusion

References

[1] The City of New York. (2017) Background on the boro taxi program. [Online]. Available: http://www.nyc.gov/html/tlc/html/passenger/shl_passenger_background.shtml

[2] ——. (2017) Your guide to boro taxis. [Online]. Available: http://www.nyc.gov/html/tlc/html/passenger/shl_passenger.shtml

[3] Taxi and Limousine Commission (TLC). (2017, Jun.) 2016 green taxi trip data. NYC OpenData. [Online]. Available: https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb

[4] ——. (2017, Sep.) 2016 yellow taxi trip data. NYC OpenData. [Online]. Available: https://data.cityofnewyork.us/Transportation/2016-Yellow-Taxi-Trip-Data/k67s-dv2t

[5] Kaggle. (2017) New york city taxi trip duration. [Online]. Available: https://www.kaggle.com/c/nyc-taxi-trip-duration

[6] S. Z. J. Z. Yunrou Gong, Bin Fang. (2016, Sep.) Predict new york city taxi demand. NYC Data Science Academy. [Online]. Available: https://nycdatascience.com/blog/student-works/predict-new-york-city-taxi-demand/

[7] P. J. T. S. Himanshu Jaiwal, Tushar Bansal, "Nyc taxi rides: Fare and duration prediction," University of California, San Diego, Tech. Rep., 2017. [Online]. Available: https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a077.pdf

[8] Currie32. (2017, Jul.) Nyc-taxi-trip-duration. [Online]. Available: https://github.com/Currie32/NYC-Taxi-Trip-Duration

[9] yukw777. (2017, Aug.) kaggle-nyc-taxi. [Online]. Available: https://github.com/yukw777/kaggle-nyc-taxi

[10] mk9440. (2017, Oct.) New-york-city-taxi-trip-duration. [Online]. Available: https://github.com/mk9440/New-York-City-Taxi-Trip-Duration

[11] A. F. A. J. Christophoros Antoniades, Delara Fadavi, "Fare and duration prediction: A study of new york city taxi rides," Stanford University, Tech. Rep., 2016. [Online]. Available: http://cs229.stanford.edu/proj2016/report/AntoniadesFadaviFobaAmonJuniorNewYorkC report.pdf

[12] H. T. V. J. F. C. T. S. Nivan Ferreira, Jorge Poco, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, Dec. 2013. [Online]. Available: https://vgc.poly.edu/ juliana/pub/taxivis-tvcg2013.pdf

[13] Kaggle. (2015) Ecml/pkdd 15: Taxi trip time prediction (ii). [Online]. Available: https://www.kaggle.com/c/pkdd-15-taxi-trip-time-prediction-ii

[14] A. P. Y. G. Hoang Thanh Lam, Ernesto Diaz-Aviles and B. Chen, "(blue) taxi destination and trip time prediction from partial trajectories," IBM Research - Ireland, Tech. Rep., 2015. [Online]. Available: https://arxiv.org/pdf/1509.05257.pdf

[15] R. S. L. Vanajakshi, S.C. Subramanian, "Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses," *IET Intelligent Transport Systems*, vol. 3, no. 1, Mar. 2009.

[16] T. M. J. E. James Biagioni, Tomas Gerlich, "Easytracker: Automatic transit tracking, mapping, and arrival time prediction using smartphones," in *ACM Sensys*, 2011.

[17] N. G. Mehmet Yildirimogl, "Experienced travel time prediction for freeway systems," in *2012 15th International IEEE Conference on Intelligent Transportation Systems*, 2012.

[18] D.-C. S. M.-H. C. Chun-Hsin Wu, Chia-Chen Wei and J.-M. Ho, "Travel time prediction with support vector regression," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 4, Dec. 2004. [Online]. Available: http://www.csie.nuk.edu.tw/ wuch/publications/2003-itsc-svr.pdf

[19] NOAA. Nowdata - noaa online weather data. National Weather Service Forecase Office. [Online]. Available: http://w2.weather.gov/climate/xmacis.php?wfo=okx

[20] M. A. Adnan Akyux, Karsten Shein, "Procedure for assigning a value for trace precipitation data without changing the climatic history," *Journal of Service Climatology*, vol. 6, no. 1, 2013.

[21] U. S. N. Observatory. [Online]. Available: http://aa.usno.navy.mil/data/docs/RS_OneYear.php