A Deep Learning Approach for TNC Trip Demand Prediction Considering Spatial-Temporal Features

Preprint

Yi Hou, Venu Garikapati, Joshua Sperling,
Alejandro Henao, and Stanley Young

National Renewable Energy Laboratory

A Deep Learning Approach for TNC Trip Demand Prediction Considering Spatial-Temporal Features

Preprint

Yi Hou, Venu Garikapati, Joshua Sperling, Alejandro Henao, and Stanley Young

National Renewable Energy Laboratory

NOTICE

Introduction

Ride-hailing or transportation network companies (TNCs), such as Uber, Lyft, DiDi Chuxing, or RideAustin, are emerging as a new and disruptive on-demand mobility service in recent years. However, the methods for developing predictive analytics to explore the nature and dynamics of TNCs across cities in the United States are still nascent due to the lack of publicly available data. Recently available public datasets on TNCs by RideAustin offer a unique opportunity to examine spatial, temporal, environmental, and special event factors associated with TNC trip demand. This study explores the use of a deep learning approach—Long Short-Term Memory (LSTM)—to predict TNC trip demand at the ZIP Code level using data from Austin, Texas. The analysis includes key predictive factors such as time of day, day of week, precipitation, and temperature, indicating their corresponding associations with TNC trip demand. Results from initial analysis show that LSTM is able to predict the TNC trip demand for the upcoming hour accurately. LSTM, when compared to other prediction methods, such as historical average and instantaneous trip demand, reduces the mean absolute error (MAE) of the model predictions by 37% and 24%, respectively. This novel method offers significant potential for scaling up and/or replicability across cities where data are available for understanding TNC trip demand to inform emerging mobility system operators. Predicting trip demand can help TNC drivers make informed decisions on how to be more efficient by maximizing passenger pickups and minimizing wait times and deadheading.

This study utilizes data from RideAustin to explore the spatial and temporal distribution of TNC trip demand, identify factors that are relevant to TNC trip demand, and develop a TNC trip demand prediction model using deep learning techniques. Before TNCs arrived as an alternative transportation mode, taxis were the most popular alternative transportation mode to travel distances that are hard to cover by bike or foot. There has been considerable interest in predicting the demand for taxi trips both from the operator perspective (to optimally position the taxis with respect to demand) and also the city's perspective (to understand the amount of travel made by taxis). Chang et al. (*1*) mined historical data to predict taxi demand distributions using clustering algorithms. Moreira-Matias et al. (*2*) applied time series techniques to forecast taxi passenger demand. Gong et al. (*3*) proposed a machine learning model, XGBoost, to predict New York City taxi demand. However, studies on TNC trip demand prediction are limited. Recently, Ke et al. (*4*) introduced the fusion convolutional long short-term memory network (FCL-Net) to forecast passenger demand for on-demand ride services in Hangzhou, China, using real-world data provided by DiDi Chuxing.

This paper adopts a deep learning model, called Long Short-Term Memory (LSTM), to predict the spatio-temporal features of TNC trip demand. LSTM is a special recurrent neural network architecture designed to learn time series data with long time spans and high dimensions. By presenting the available data and methods for predictive analytics, this paper aims to lay the groundwork for new early-stage research, analyses and adaptable decision support for cities and TNCs.

Methodology

LSTM is a special recurrent neural network architecture that was first introduced by Hochreiter and Schmidhuber (*5*). LSTM has the same chain-like structure as a recurrent neural network, but the repeating module has a different structure. Instead of a single neural network layer, LSTM has

four layers interacting in a very special way. The calculation steps in LSTM repeating module are performed as follows.

1. A sigmoid layer σ called the "forget gate layer" makes a decision on what information to throw away from the cell state. It looks at the output of the previous step, h_{t-1}, and current inpu,t x_t, and outputs a binary decision for each number in the previous cell state, C_{t-1}. It is calculated as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

2. LSTM decides what new information to store in the cell state. A sigmoid layer σ called the "input gate layer" decides which values to update. Then, a tanh layer outputs an array of new candidate values, $\tilde{C}_t$, to add to the previous cell state, C_{t-1}. It is calculated as:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

3. The previous cell state, C_{t-1}, is updated by the equation below.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

4. Finally, LSTM decides what to output. First, a sigmoid layer σ called the "output gate layer" decides what part of cell state to output. Next, cell state is passed through a tanh layer and multiplied by the output of "output gate layer". The mathematical form of the calculations is ((5) and (6)):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

Findings

The seven ZIP Codes considered for the TNC demand prediction analysis are shown in red in Figure 1. The proposed models predict TNC trip demand in all zones for the coming hour. The target variable for the study is the number of TNC trips originating from any of the seven ZIP Codes within the coming hour. From an extensive data exploration exercise, the TNC trip demand of the previous hours of the day was found to be a significant predictor of the trip demand for the coming hour. In addition, four other relevant factors, namely time of day, day of week, precipitation, and temperature, were also considered as independent variables in the model. After data aggregation, a total of 4,745 observations (~ 6 months of observations = 24*180) were extracted from the dataset. A total of 3,990 observations (~85% of the data) from July 24, 2016, to January 6, 2017, were used for model training. A total of 755 (~15% of the data) observations from January 7 to February 6, 2017, were used to test the model.

2
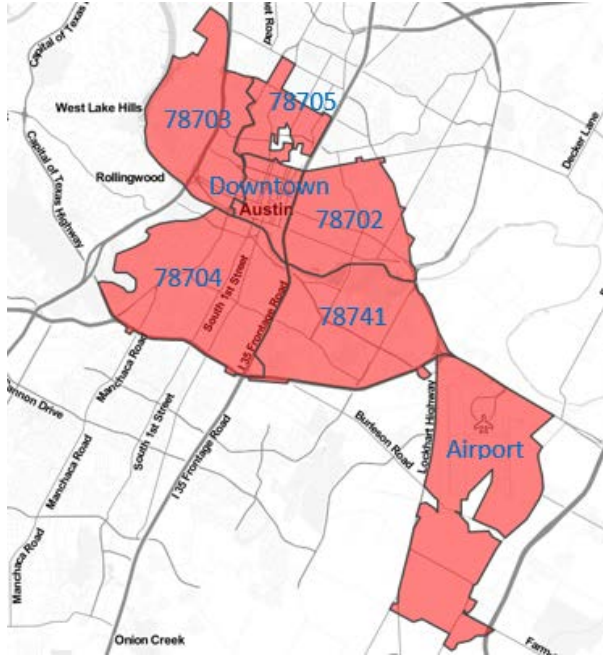
FIGURE 1 The seven ZIP Code zones for trip demand prediction.

An exhaustive trial-and-error process was adopted for tuning hyperparameters. During the training process, mean square error (MSE) was chosen as the loss function. In every epoch, 10% of the training data was used for validation. LSTM with 6 hours of "look-back" interval, one LSTM layer of 128 hidden neurons, 30 epochs, and a batch size of 32 yielded the best model accuracy with lowest MSE for the validation dataset.

Two baseline methods, namely instantaneous trip demand and historical average, were used for comparison against the LSTM results. Instantaneous trip demand uses the trip demand of the most recent hour as a proxy for the trip demand of the next one hour. The historical average is conditioned on hour of the day, day of the week, and ZIP Code. All methods were evaluated using MAE and root mean square error (RMSE) (formulations shown in (7) and (8)).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|d_i - \widehat{d_i}| \qquad (7)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \widehat{d_i})^2} \qquad (8)$$

where d_i is the observed trip demand, $\widehat{d_i}$ is the predicted trip demand, and N is the total number of observations.

Prediction results for the test data are presented in Table 2. From the results, it can be observed that the historical averaging performs the worst of the three methods considered, followed by the instantaneous trip demand method. LSTM emerges as the clear winner with the lowest MAE as

3

well as RMSE values. Compared to the historical average and instantaneous trip demand, LSTM reduces the MAE by 37% and 24%, and RMSE by 51% and 45%.

**TABLE 1 One Hour Ahead – Trip Demand Prediction Results**

| Model	MAE	RMSE
Historical Average	11.5	25.0
Instantaneous Trip Demand	9.5	22.5
LSTM	7.2	12.3

Units: number of trips per hour.

A robust model must be able to predict accurately under a variety of conditions (i.e., peak hour, non-peak hour, weekdays, and weekends). Therefore, the developed LSTM was evaluated by comparing the predicted TNC trip demand with that of observed data for a span of one full week. A one-week time window from Monday to Sunday was randomly selected from the time range of test data. Figure 2 shows the predicted results of LSTM for each of the ZIP Codes in the one-week period from January 16 to January 22, 2017. The blue line in the figure indicates predicted trip demand. The red dashed line represents observed trip demand. Figure 2 shows that LSTM is effective in predicting TNC trip demand on both weekdays and weekends. The observed TNC trip demand clearly indicates that there are two distinct peaks on Friday and Saturday nights for all zones except the airport. It can be observed from the figure that LSTM was able to capture these peaks. In addition to periodical peaks, LSTM is able to provide adequate prediction for special events.
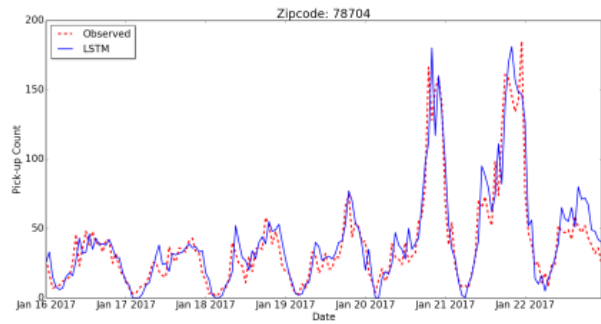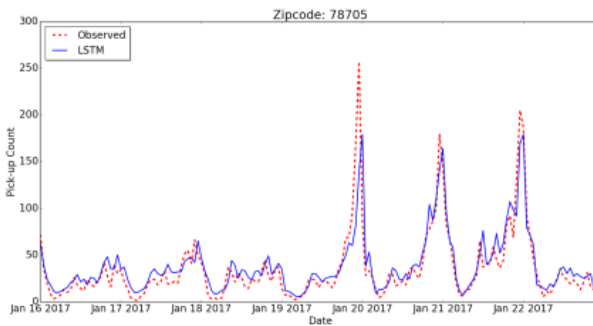
4
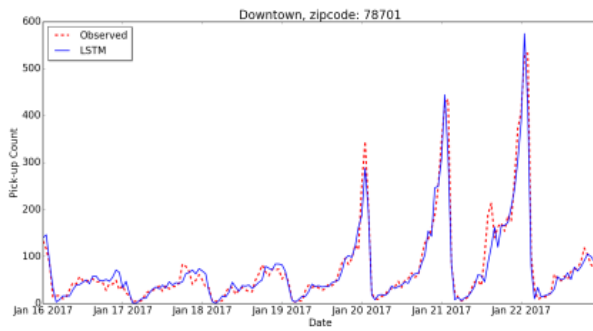
(a) Downtown
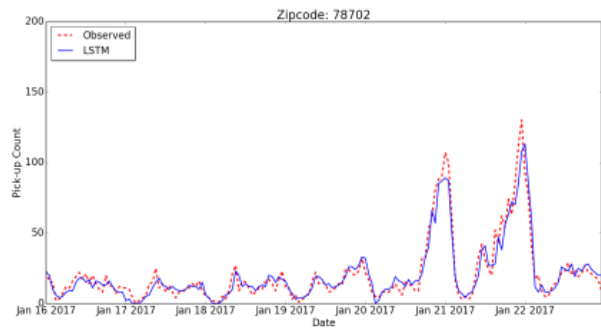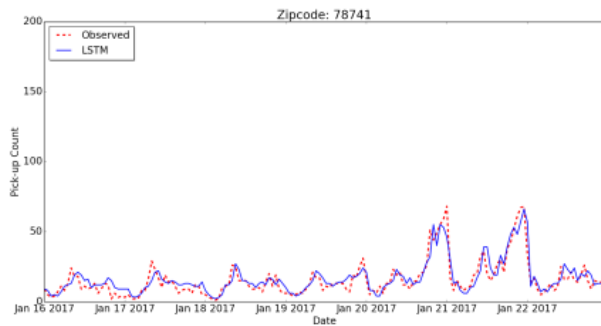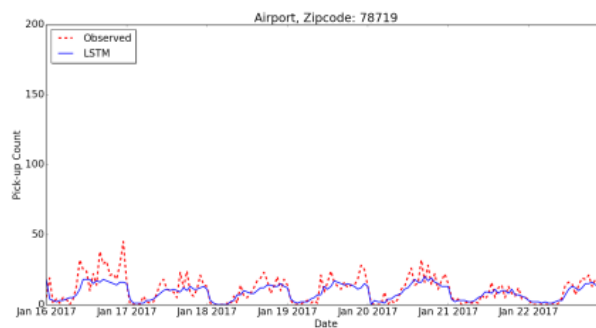
(b) ZIP Code 78702
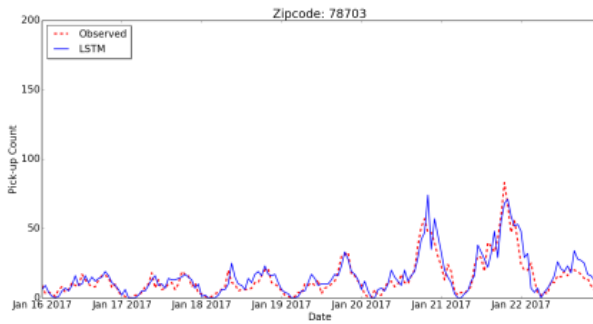
(c) ZIP Code 78703

(d) ZIP Code 78704

(e) ZIP Code 78705

(f) ZIP Code 78741

(g) Airport

FIGURE 2 Predicted values versus observed values.

5

Conclusions

The Long Short-Term Memory (LSTM) deep learning method is adopted in this paper to develop a TNC demand prediction model. LSTM is capable of simultaneously predicting trip demand at multiple locations in the city for the coming hour, thereby enabling drivers and future mobility system operators to optimize system performance using available TNC travel data. The model was evaluated in the seven ZIP Codes of the downtown core, surrounding areas of downtown, and the airport area (which cover about 70% of the TNC trip demand). Given a set of inputs, this model predicts the TNC trip demand at the ZIP Code level for the coming hour. To assess the predictive capability of the LSTM model, two baseline models namely Instantaneous Trip Demand and Historical Average, were developed. The LSTM outperformed the two baseline methods in terms of prediction accuracy, and performed extremely well in predicting the TNC trip demand for the coming hour across multiple ZIP Codes in Austin. The results show that LSTM predicts TNC trip demand with a mean absolute error (MAE) of 7.2 trips per hour, reducing the MAE by 37% and 24% compared to the historical average and instantaneous trip demand methods. Further examination of the predictions of LSTM in different ZIP Codes under a variety of conditions (i.e., peak hour, non-peak hour, weekdays, and weekends) reveals that LSTM is able to capture the trip demand variations not only in normal peak and off-peak hours, but also for special events and aberrant demand patterns. While the model predicts TNC trip demand well for most of the zones considered for analysis, it does not perform as well in predicting TNC trip demand at the Austin Airport. This is most likely because the TNC trip demand pattern for the airport is motivated by factors that are different from the rest of the zones (plane arrival/departure times, distance of the airport to various zones in the city, transit accessibility and timings). Developing TNC demand predictions models for special generator zones such as airports remains a direction for further exploration and future research.

Acknowledgments

References

1. Chang, H.-W., Y.-C. Tai, and J. Y.-J. Hsu. Context-Aware Taxi Demand Hotspots Prediction. *International Journal of Business Intelligence and Data Mining*, 2010. 5: 3–18.
2. Moreira-Matias, L., J. Gama, M. Ferreira, and L. Damas. A Predictive Model for the Passenger Demand on a Taxi Network. *Proc. 15th International IEEE Conference on Intelligent Transportation Systems*, Anchorage, Alaska, USA, 2012.
3. Gong, Y., B. Fang, S. Zhang, and J. Zhang. Predict New York City Taxi Demand. NYC Data Science Academy, 2016. Available: https://blog.nycdatascience.com/student-works/predict-new-york-city-taxi-demand/
4. Ke, J., H. Zheng, H. Yang, and X. Chen, "Short-Term Forecasting of Passenger Demand under On-Demand Ride Services: A Spatio-Temporal Deep Learning Approach," *Transportation Research Part C*, 2017. 85: 591–608.
5. Hochreiter, A., and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 1997. 9: 1735–1780.