Multimodal Location-aware Taxi Demand Prediction

Abhishek Taur

Department of Electrical and Computer Engineering, Texas A&M University
{abhishektaur123}@tamu.edu

Abstract—Accurate forecasting is important in case of time-series data. Many applications where the data is time-dependent forecasting is based on events which directly or indirectly influence the data. Quite often models try to use large temporal data but they don't include data which is in the form of unstructured text and contains valuable information regarding the time-series data. Even though the text data is often not directly relatable to the time-series data it contains contextual information regarding the time-series data. In this project, I have shown how using unstructured data from the Barclays Center we can predict the taxi demand in Brooklyn zones 97, 25, 181, 189. A neural network model was used to make the predictions and word embedding of the event descriptions was used along with time-series data of the taxi. The results show a significant decrement in the Mean Absolute Error and Root mean square Error when we use MLP or LSTM convolution layer.

Index Terms—Deep learning, Textual data, Taxi demand, Special events, Data fusion, Time series forecasting

1 Introduction

Taxies are a major mode of transportation, especially in a densely populated city. People use a taxi for commuting to the work or a specific location. The demand for taxies is heavily decided by various factors like the location where the pickup is, traffic in that area, the number of peoples using it at that time, the drop off location, toll rates, events at the pickup or drop site, weather, and many others. Some of these factors are recurrent *i.e.* important on a daily basis for influencing the taxi demand. For example, for a group of companies where employees commute daily using taxies a trend of taxi demand is set. Deciphering travel behavior is an important research topic for developing efficient and effective intelligent transportation system. Many approaches use recurrent mobility patterns or short-term correlation models to capture habitual travel behavior [1], [2], [6] . [8] studies the driver's decision to pick up from the airport or to cruise in the city for pick up based on policies implemented by the airport authority. It provides an important tool to suggest policy recommendations to improve ground access at JFK international airport. This model doesn't take into account event data to improve the decision of the driver. It uses simple logistic regression to make the prediction. While these approaches are good for predicting mobility patterns in non-eventful areas like housing neighborhood or airport they are prone to areas having huge events like music concerts, games, sports, dance performance, etc. These models are not prepared for demand surges in a taxi which leads to increased fare prices and bad user experience. [8] is restricted to study of taxi demand in JFK airport.

In order to fully encapsulate the effect of events, one can exploit the vast event databases that are present on the Web. However, a lot of this information is in the form of unstructured text and cannot be directly related to taxi demand in a zone. Solution to this multi-domain data fusion problem becomes the key motivation for understanding the mobility demands in events area. This approach can be extended to other fields of study too like finance [4], elections [3]. Deep learning has shown promising results in case of natural language processing and therefore suitable for extracting meaningful data from unstructured text. Previous work [7] have shown that deep learning models are already successful in predicting taxi demand and outperform traditional approaches. However, none of them have used unstructured text data for taxi prediction and are not location aware. This project aims to explore the deep learning model and use multi-modal data fusion to predict the taxi demand in New York Brooklyn zones. It can be easily extended to make taxi demand predictions in the whole of New York or any other city and tries to solve the general problem of multi-modal data fusion. The modal is compared against a baseline model which uses the last few taxi demands in that particular area. The New York zone is given as an input to the model and based on any live or past events dynamically extracted from the Web it is able to predict the taxi demand in that particular area. Fig. 1 shows the time series plot for the number of taxi pickups in zones 97, 25, 181, 189 of New York. It can be seen that the there is an event in zone 181 which lead to increased taxi demand in the zone 181. This event occurred on Friday and has impacted the taxi demand.

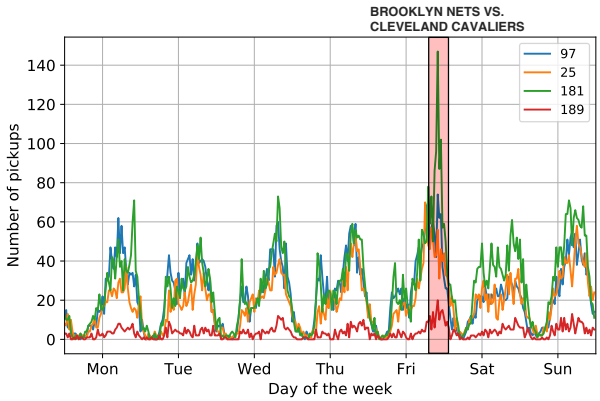


Fig. 1: Example of Taxi demand in Brooklyn zones

The remainder of this paper is organized as follows. In the next section, I present the proposed neural network architectures and the experimental results are presented in Section 3. The paper ends with the conclusions and Future Work(Section 4).

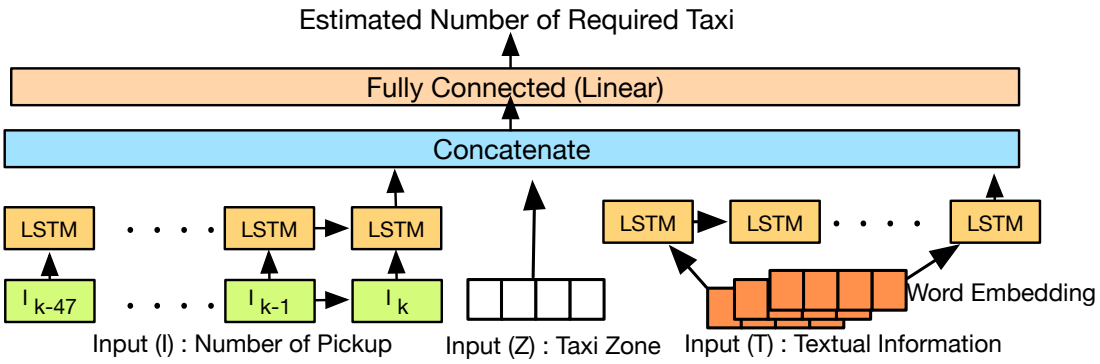


Fig. 2: Neural Network Architecture for taxi demand prediction

2 Methodology

In this section, a general-purpose methodology for exploiting event data from the Web to do time-series forecasting for taxi demand prediction is explained.

2.1 Text and Taxi data Pre-processing

The New York taxi data was extracted from [5] and the events data was extracted from https://www.barclayscenter.com/events/event-calendar which has the list of all events happening in Barclays center. Selenium WebDriver and BeautifulSoup were used to scrap events data from Web. From the scraped data event descriptions was extracted. The taxi data was downloaded from the Web and was pre-processed into a CSV file to contain the year, month, day, hour, minute, zone, the number of pickups in an ordered fashion to represent the time series.

2.2 Neural Network Architecture

Numerous methods can be used to combine textual data with time-series data for forecasting. The neural network architecture has a front-end embedding layer which maps each event description to a lower dimensional space to keep only the relevant information in form of integer vectors. It basically maps the semantic relationship of the words to a geometric space. For example, ”bat” and ”bowl” are semantically associated to each other while ”bat” and ”world” are not and therefore the geometric distance between this two words should be more. In this paper, the word embeddings used are similar to those used in Tensorflow. A tensor is created which is then passed to the LSTM layer. The input lags which is the time-series data is combined with this LSTM to produce input to another convolution layer which can either be MLP or LSTM based on the configuration given at runtime. If it is MLP a linear activation function is applied to the output of the convolution network otherwise if it is LSTM then a ReLu activation function is applied. Adam optimizer is used for mean squared loss function for the neural network to converge. I tried using different convolution layers but there was not much difference in the output. More than one convolution layer produced almost the same output as one convolution layer. Fig. 2 shows the neural network architecture.

3 Evaluation

3.1 Data set and case study

The base dataset for our study comes from the NYC Taxi Limousine Commission [5] which consist of records for the yellow, green and FHV taxi from Jan 2009 to June 2018. The data that is used for this project consist of taxi trips from Jan 2017 to June 2018 as it contains the zone data along with Pickup location and drop off Location. Based on the top venues for events was the Barclays center and its official website has a complete set of events that have occurred from Jan 2017 to June 2018. Barclays center is modern multi-purpose arena which hosts a lot of live concerts, events and is the home to NBA teams Brooklyn Nets and the New York Islanders. It is situated in the zone 181 of as shown in Fig. 3 [5]. A zone is set of locations for taxi operations in New York. For the project the zones 97, 25, 181, 189 were used and all the events in Barclays center were used. Only the green cabs operate in the Brooklyn area of the New York. Therefore only the green cabs data was extracted from the [5]. The individual pickups were grouped according to the day and sampling of the data was done every half hour and a time-series was created.

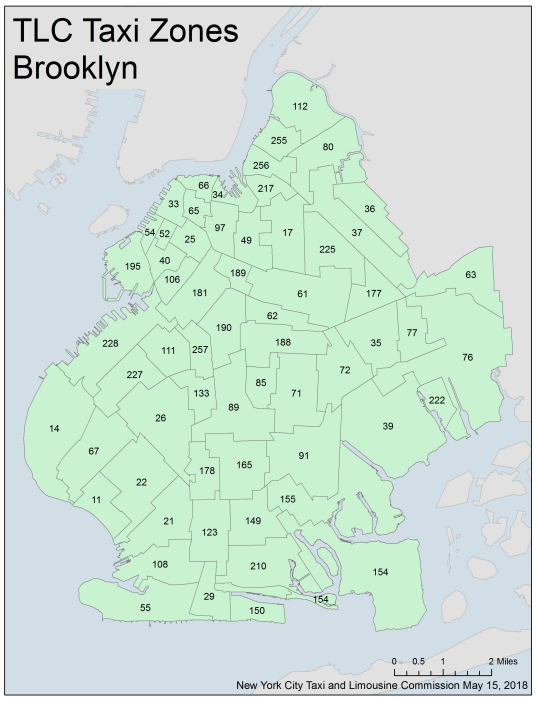


Fig. 3: Brooklyn taxi zones

	Baseline		MLP - L		LSTM - L		MLP - LC		LSTM - LC		MLP - LCE		LSTM - LCE	
zone	mae	rmse	mae	rmse	mae	rmse	mae	rmse	mae	rmse	mae	rmse	mae	rmse
97	13.093	16.227	13.854	17.231	12.041	15.056	12.037	15.025	12.014	15.011	11.770	14.723	11.645	14.607
25	10.373	12.754	10.861	13.553	10.354	12.964	10.414	13.030	10.289	12.936	10.195	12.671	10.059	12.543
181	14.158	18.143	13.114	16.543	12.230	15.541	12.135	15.354	12.111	15.400	11.974	15.102	11.852	15.087
189	2.393	3.301	1.874	2.462	1.757	2.321	1.822	2.472	1.758	2.301	1.693	2.275	1.677	2.266

TABLE 1: Model Results

3.2 Experimental setup and analysis

The dataset was divided into test, validation, training datasets. The training dataset consisted of 70% of the whole dataset, validation dataset was 10% of the dataset and 20% of the dataset was test dataset. No shuffling of the data was done to preserve the time-series. A baseline model was developed to compare. In the baseline model, I took the sum of previous few time-series data points and took mean of them. This mean was the predicted value of the baseline model.

In order to understand the contribution of the different sources I performed an incremental analysis. First, the model was trained without using any event-info and used MLP for training the time-series data. Then I used the location as input to the MLP and with this added information the Mean Absolute Error(MAE) and Root mean square Error(RMSE) reduced by a factor of 1 plus for all the zones. This shows that the new information of the zone gives the model a better idea about the taxi demand. Then I added the event info to the model and with this, we can see a reduced mae and rmse thus leading to improved results. TABLE. 1 shows the various results for MLP and LSTM. It can also be seen that LSTM performs better than MLP. LSTM performs better than MLP as it can capture long-term dependencies. The value predictions graphs are shown in Fig. 4 and as expected they closely resemble the actual values of the taxi demands.

4 Conclusion and Future Work

In this study, a neural network architecture was designed which either used MLP or LSTM as the convolution layer and made used of word embeddings which were derived from unstructured data from the Web. The project shows how the use of multi-modal design can improve the prediction of taxi demand prediction. The study included 4 zones in Brooklyn, New York and used one of the biggest event centers in New York that are the Barclays Center to prove the use of events data for taxi demand prediction. This study has a lot of potential in other fields like finance, economics *etc*. My design was a simple one which can be further extended to include all the zones in New York City. Also, the study focused on green cabs but yellow cabs, FHV can be also included. A similar approach can be used to study the effect of social media like Twitter on the taxi demand. This project was based on one event center but can be further extended to include more event centers. Weather data can also be combined similar to zones to make predictions.

References

[1] S. Krygsman, M. Dijst, and T. Arentze, "Multimodal public transport: an analysis of travel time elements and the interconnectivity ratio," *Transport Policy*, vol. 11, no. 3, pp. 265–275, 2004.

[2] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1393–1402, 2013.

[3] B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith *et al.*, "From tweets to polls: Linking text sentiment to public opinion time series." *Icwsm*, vol. 11, no. 122-129, pp. 1–2, 2010.

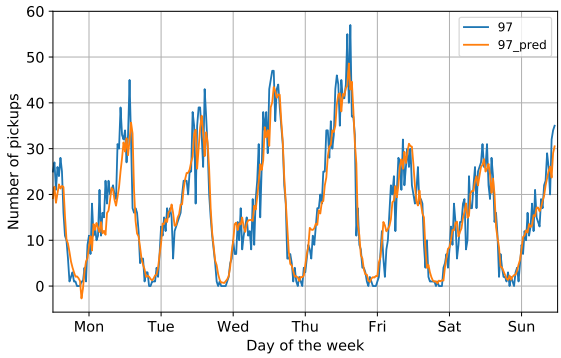
[4] X. Tang, C. Yang, and J. Zhou, "Stock price forecasting by combining news mining and time series analysis," in *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, vol. 1. IEEE, 2009, pp. 279–282.

[5] N. TLC, "Nyc taxi and limousine commission (tlc) trip record data," *URL http://www. nyc. gov/html/tlc/html/about/trip_record_data. shtml*, 2017.

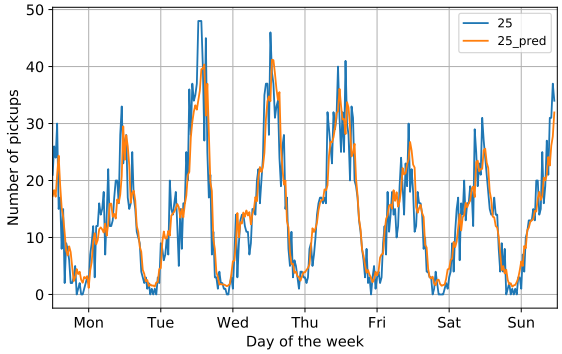
[6] N. Van Oort, T. Brands, and E. de Romph, "Short term ridership prediction in public transport by processing smart card data," *Transportation Research Record*, no. 2015, 2015.

[7] J. Xu, R. Rahmatizadeh, L. Bölöni, and D. Turgut, "Real-time prediction of taxi demand using recurrent neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 8, pp. 2572–2581, 2018.

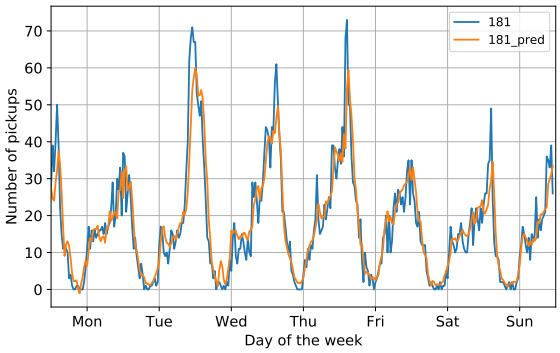
[8] M. A. Yazici, C. Kamga, and A. Singhal, "A big data driven model for taxi drivers' airport pick-up decisions in new york city," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 37–44.



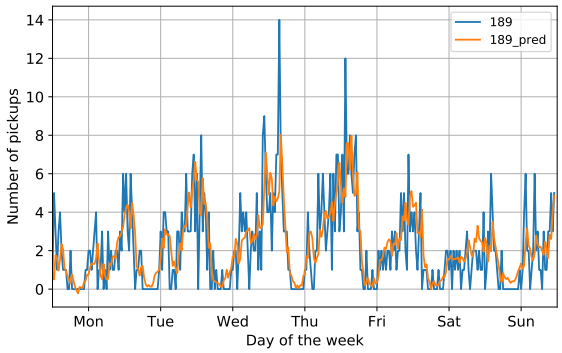
(a) Zone 97 prediction



(b) Zone 25 prediction



(c) Zone 181 prediction



(d) Zone 189 prediction

Fig. 4: Zone predictions