OmniTaxis

Predicting NYC Taxi Fares, Trip Distance, and Activity

Authors: Boxiao Pan, Varun Nambiar, Paul Jolly
{bxpan, vnambiar, pjolly94}@stanford.edu

Introduction

For a New York City taxi driver, being in the right place at the right time is often what makes or breaks a day. One may naively assume that the right spot corresponds to a place simply where demand (or activity) is high. However, taxi drivers might find it more lucrative to be in a slightly lower activity location where people are demanding shorter trips that are worth more.

To assist drivers in this decision, we explored different models to predict the **activity, fare amount** and **trip distance** given input features **location**, the **day of the week**, and the **time of the day**.

Data and Features

Raw Data:

New York City Taxi and Limousine Commission (TLC) provides a large amount of trip data from 2014 to 2018 including the following information:

a) Date and time of trip
b) Pickup location
   - Mid-2014 – Mid-2016: Latitudes and longitudes
   - Mid-2016 – Mid-2018: Location IDs
c) Fare amount
d) Trip distance

Data Pre-Processing:

- **Label Bucketing:**

Instead of using exact values for the labels (activity, fare and trip distance), we discretized them by creating buckets. This was done by inspecting the distribution of the labels over the data and selecting realistic bucket ranges (table 1). Not only did this enhance the model performance, but more importantly, it proves more useful in application, since we are presenting estimated ranges (i.e., buckets) to the driver as opposed to exact numbers, which is what drives care more about.

- **Creating Location Clusters with K-means:**

To increase the granularity of the newer trip data (i.e., post Mid-2016), we created clusters using the latitude / longitude data from the older dataset (i.e., pre Mid-2016) and distributed the newer data into "cluster IDs" based on the distribution within each location ID obtained from K-means (figure 1).

Subset Description:

- Data set split: 90% / 5% / 5%
- Training Set:
  - Fare and trip distance: **1.8 million**
  - Activity: **0.27 million**
- Validation and test set:
  - Fare and trip distance: **0.2 million**
  - Activity: **0.03 million**

Bucket ID	Activity (# Trips)	Fare ($)	Trip Distance (miles)
0	< 2	< 0	< 0.5
1	2 - 5	0 - 5	0.5 – 1.0
2	5 - 7	5 - 10	1.0 – 1.5
3	7 - 10	10 - 15	1.5 – 2.0
4	10 - 15	15 - 25	2.0 – 3.0
5	15 - 25	25 - 50	3.0 – 5.0
6	25 - 35	50 - 60	5.0 – 10.0
7	35 - 45	> 60	> 10.0
8	> 45	-	-

Table 1: Label classes created through bucketing

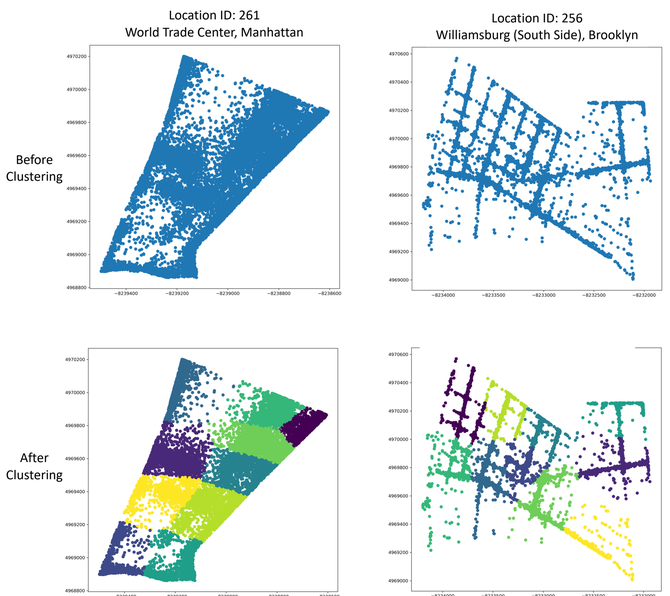

Figure 1: Two examples of location clustering based on trip data through K-means. The top row shows trips before clustering and the bottom row shows trips classified into clusters.

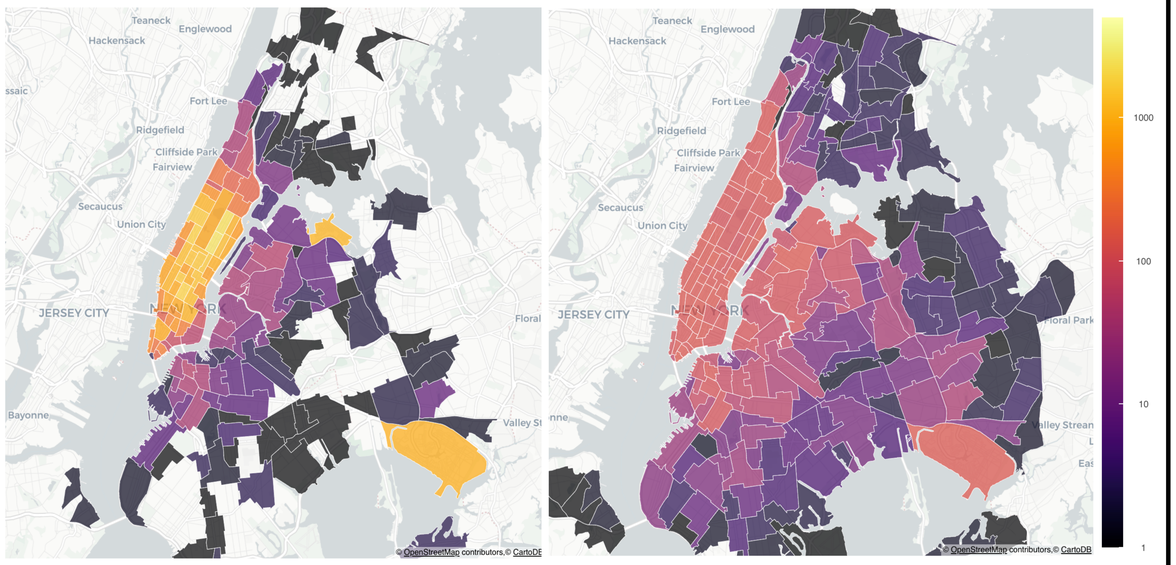

Figure 2: Heat map representation of ground truth activity (left) versus predicted activity (right). The lighter the color, the heavier the activity. Despite not being able to predict the exact magnitude of activity well, our model is able to capture the relative activity between different locations.

Models

Random Forest Classification (RFC)

- Loss Function:
  Gini Loss -> $L_{Gini} = \sum_c \hat{p}_c(1-\hat{p}_c)$

Fully Connected Neural Network (FCNN)

- Loss Function:
  Cross Entropy Loss -> $L_{Cross} = -\sum_c p_c \log \hat{p}_c$
- 4 hidden layers with 6, 10, 6, 12 neurons respectively and all with ReLU activation function.
- 1 output layer with Softmax activation function.

Long Short-Term Memory (LSTM) Network

- Loss Function:
  Cross Entropy Loss -> $L_{Cross} = -\sum_c p_c \log \hat{p}_c$

* The model parameters are for the case of activity prediction.

Experimental Results

Model	Activity	Fare	Trip Distance
RFC	51.02%	45.71%	30.08%
FCNN	35.67%	45.72%	23.16%
LSTM	26.65%	27.72%	23.10%

Table 2: Quantitative experimental results

For all three labels, Random Forest performs the best (or nearly) among the models, while all these three models perform poorly on this task. We think this should mainly be attributed to that the prediction task is too hard given the features we have. Figure 2 shows an example heat map output comparing predicted activity with the ground truth. While the exact numbers do not really match, our model captures the relative activity quite well.

Conclusion and Future Work

Our models can pick up the relative differences between different neighborhoods but does not perform well when trying to predict the exact numbers. This may be a result of using a single model to predict for the entire New York Area. Here are possible ways forward:

- Focus our model to Manhattan Island and the surrounding airports.
- Only include areas where Yellow Taxis officially serve.
- Add more data for the LSTM model. Our sparse sampling may be causing issues.
- Further tune network hyper-parameters.

References

"sklearn.cluster.KMeans," *KMeans - scikit-learn 0.19.2 documentation*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. [Accessed: 11-Dec-2018].