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 (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.

- **Creating Location Clusters with K-means:** To increase the granularity of the newer trip data, 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.

![Image](image.jpg)  
**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.

Table 1: Label classes created through bucketing

<table>
<thead>
<tr>
<th>Bucket ID</th>
<th>Activity (F Trips)</th>
<th>Fare ($)</th>
<th>Trip Distance (miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&lt; 2</td>
<td>&lt; 0</td>
<td>&lt; 0.5</td>
</tr>
<tr>
<td>1</td>
<td>2 - 5</td>
<td>0 - 5</td>
<td>0.5 - 1.0</td>
</tr>
<tr>
<td>2</td>
<td>5 - 7</td>
<td>5 - 10</td>
<td>1.0 - 1.5</td>
</tr>
<tr>
<td>3</td>
<td>7 - 10</td>
<td>10 - 15</td>
<td>1.5 - 2.0</td>
</tr>
<tr>
<td>4</td>
<td>10 - 15</td>
<td>15 - 25</td>
<td>2.0 - 3.0</td>
</tr>
<tr>
<td>5</td>
<td>15 - 25</td>
<td>25 - 50</td>
<td>3.0 - 5.0</td>
</tr>
<tr>
<td>6</td>
<td>25 - 35</td>
<td>50 - 60</td>
<td>5.0 - 10.0</td>
</tr>
<tr>
<td>7</td>
<td>35 - 45</td>
<td>&gt; 60</td>
<td>&gt; 10.0</td>
</tr>
<tr>
<td>8</td>
<td>&gt; 40</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Models

Random Forest Classification (RFC)
- **Loss Function:** Gini Loss -> \( L_{Gini} = \sum_i \bar{p}_i (1 - \bar{p}_i) \)
- Fully Connected Neural Network (FCNN)
- **Loss Function:** Cross Entropy Loss -> \( L_{Cross} = -\sum_i p_i \log \hat{p}_i \)
- 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)
- **Loss Function:** Cross Entropy Loss -> \( L_{Cross} = -\sum_i p_i \log \hat{p}_i \)

![Image](image2.jpg)  
**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.

![Image](image3.jpg)  
**References**