Intersection Congestion Prediction

Shuang Song (ss2627), Ken Jianhao Wu (jw2585), Pranshu Gupta (pg475)

1. Project Goals and Applications

The goal of our project is to predict congestion at intersections in the major cities in US. The results of our project could be used to:

- Reduce the stop-and-go stress of commuters
- Improve safety, optimize operations and find opportunities for infrastructure challenges

2. Dataset Description

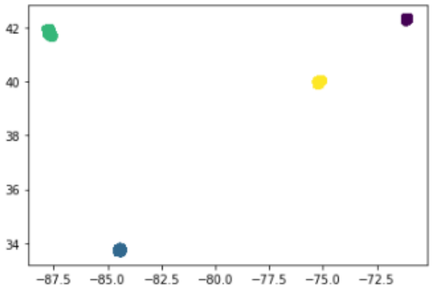
The dataset we use contains trip-logging metrics by commercial vehicles from about 4,800 unique intersections in four major cities. It contains 27 features and more than 857k entries. These features include city, intersection ID's, coordinates, entry and exit street names, hour of day, weekend or not, month, direction of entries and exits and percentiles for total time stopped, distance from first stop and time from first stop. Most of the time records are zeros. We have much more intersection data for city 2 and 3, less for city 0 and 1. The dataset contains missing values only for the two features containing street names.

3. Data Exploration, Feature Transformations and Handling Missingness

We analyse how many roads are linked to a particular intersection. Some roads might be one way and so number of entry and exits are counted separately. All intersections have at least one entry and at least one exit. Also, by thinking intuitively, an intersection should have more than just one entry street and one exit street. However, there are 414 intersections with one entry and exit. Most of these were entries with missing street names and are thus replaced with the name Unknown for now. The number of entry and exit streets for intersections are added to our training set.

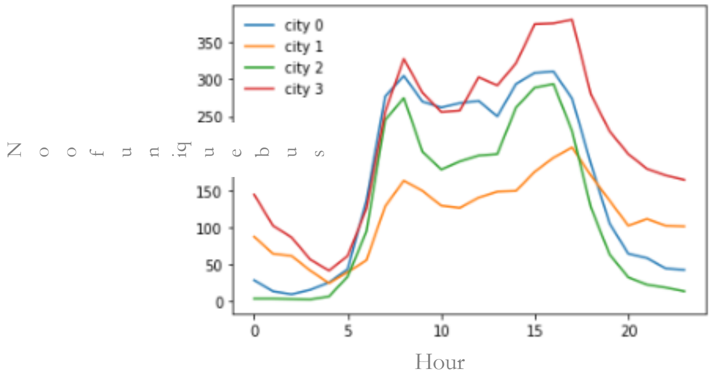
Since the directions are related to each other, they are encoded to numerical values by including a value for the North-South direction and another value for East-West direction. (e.g. W = (-1, 0), SW = (-2^0.5, -2^0.5), N = (0, 1)).

Although there is a cities feature, we check if there are roads in between the cities etc. or mislabelled city name in the data. By using K-means method on the longitude and latitude, we easily cluster the data into four groups, and the graph shows that they are perfectly clustered. This means that we do not have to deal with issues like roads between cities etc.



]

It is observed that even with the separation of time, the data is highly unbalanced. So we try to find the busy streets first. We first define busy when the 20th percentile of total time stopped is greater than five and the 50th percentile of total time stopped is greater than ten. However, this led to slightly biased results and so we try a more strict definition of "busy", this time we restrict it to have at least 30 minutes waiting time on average.



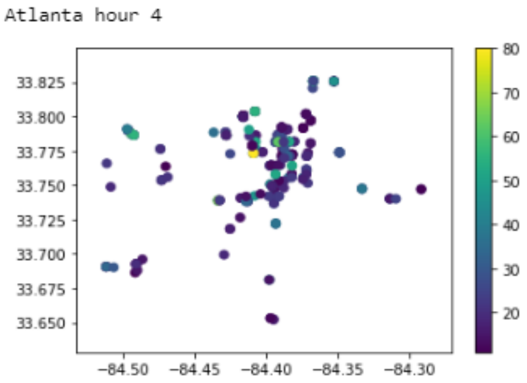
Some interesting observations from the following data and chart:

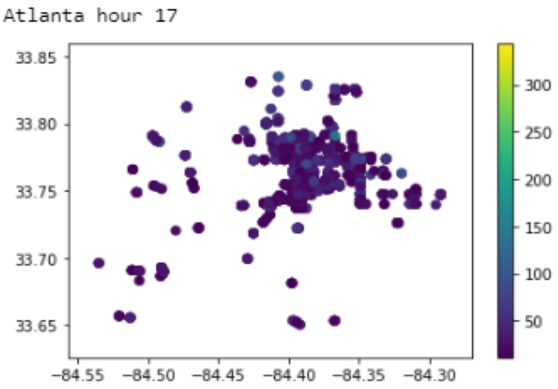
- Even though for city 0 we only have 973 unique intersections in our data, we see in its peak time, about a third of its intersections are busy. For city 1 we only have 377 unique intersections and at its peak time about half of them are busy
- Now that we have a stricter definition of 'busy', the number of busy streets are actually getting closer to each other despite the number of total intersections in that city

Additionally, we try to understand the relationship between time from first stop, distance from first stop, and our total stop time. For this, we first look at the correlation matrix. The time from first stop is almost perfectly correlated to the total stop time. The distance from first stop has a correlation coefficient of ~0.6, which indicates strong correlation as well.

The number of entry street is mostly correlated to the number of exits streets, this lowers the possibility of congestion due to more number of entries than exits. Also the total stop time is uncorrelated with number of exit and entry streets.

Visualizing the busy intersections in Atlanta at hour 4 and hour 17:





4. **Searching for Models**

We first select a few numerical variables as our features. These include intersection ID, latitude, longitude, city, hour, weekend, month, number of entry streets, number of exit streets, entry direction and exit direction. The target is the 50th percentile of the total time stopped.

For a preliminary analysis, we run a linear regression against all selected variables and obtain a mean squared error of 236.

We further try a KNN Regressor using the same variables as we did in regression. This was done for both uniform and distance weights and for number of neighbours ranging from four to nine. We obtained a score of 0.4 using the test set for the best KNN model obtained.

Additionally, we also tried Gradient Boosting with number of estimators as 100, learning rate as 0.1, maximum depth as 3 and least squares as our loss function. This gave a mean squared error of 214. This analysis also provided us with feature importance, and it showed that the latitude, longitude of the intersection, the hour of day and the number of streets linked to the intersection were the most important features in predicting congestion and that the other features were less important.

5. **Model Effectiveness Testing**

We separated our training data set into 80/20 for training and validation. Then we will employ k-fold validation and the test set to test the effectiveness of our model.

6. **Plans to Avoid Overfitting**

   a. Regularization: we will try both L1 (Lasso) and L2 (Ridge) regularization methods in our models to prevent overfitting.
   b. Choosing less robust algorithms: instead of using linear regression with least mean square error loss, we can use Huber loss instead, which is less sensitive to outliers and thus less likely to overfit than least mean square error loss.
   c. Using ensemble methods: through ensemble methods, we can achieve lower variance and average out the bias. We will first use ensemble methods like random forest, then after some classifiers are built, we will ensemble those classifiers together and form our results through averaging or weighted averaging.

7. **Upcoming Plans**

   a. Cluster the intersections within the same city. We can try different number of clusters later. This feature might be helpful to determine the potential traffic of other unseen intersections based on their cluster of the streets. (Alternatively, we can use k-nearest-neighbours for predictions of unseen intersections)
   b. Congestion is also widely known to be dependent on weather and we plan to add another feature including weather data in our model
   c. Use a more robust loss function for Gradient Boosting as our data involves large outliers which cannot be ignored. Also use Parameter Grid to obtain better parameters for our fit.
   d. Further exploration of intersections which always have zero as their total time stopped for all percentiles