What is the problem you are attempting to solve?

Traffic and congestion on New York City Streets is a real and present problem. I want to create a model that can predict not only if a street will have traffic but at what speed the traffic is travelling at. If crashes do occur, the slower a vehicle is moving, the likelihood of it being a fatal crash decreases. It is in the interest of the safety of the public for the city to be able to properly manage and predict the speeds at which vehicles are moving on its streets.

How is your solution valuable?

The use of this model could assist in helping prepare the Department of Transportation to be ready and able to respond when congestion hits an especially large rate. Other applications could be for the assessment of strategies for fixing or upgrading roads. If the city were to think about the possible impact of adding or removing a lane to a road, or the possible impact of construction on a certain section of road.

What is your data source and how will you access it?

I will be using three main sources of data. First, the real-time traffic speed data from NYC's open data program. Secondly, I want to pull in the speed limit data for each section of road that is used in the real-time traffic data. Finally, to gain more insights about the features about each road, I will pull in the LION dataset, which catalogues each street in New York City and each detail about it.

What techniques do you anticipate using?

Preprocessing:

I will use the geographic features of each of the datasets to connect the features to each other. This will require using the shapely python library, along with Geopandas dataframes. I will also examine the relation of time with the data, graphing with seaborn and matplotlib how each section of road performs.

Supervised Learning:

I will create both a classification and a regression outcome variable to model with. The regression will focus on traffic speed, whereas the classification will focus on the speed in relation to the road's speed limit. I will run these on a linear regression, random forest, KNN, decision tree, and gradient boosted decision tree.

Unsupervised Learning:

As a next step, I will run the data through a KMeans model and then examine what I can learn from the clusters it creates. Organizing the clusters into charts that compare the outcomes to the actual data.

Tensorflow & Keras:

Finally I will build a convolutional neural network using Tensorflow and Keras and run the traffic data through it. I will be approaching this as a supervised problem and use the classification as my outcome variable.

What do you anticipate to be the biggest challenge you'll face?

I anticipate connecting all of the data and gathering all of the features to be a particularly difficult challenge. I also foresee the data possibly having some problems learning from an unsupervised perspective.

513-885-2574 / thomascbriggs@gmail.com Linkedin / GitHub