Traffic Flow in NYC

Nicolaus Gagliano

STATUN3106 Applied Machine Learning

3/13/2025

**Introduction**

New York City is known for one thing when the clock strikes 4 pm, traffic. Ranked 22nd in the global traffic rankings, New York residents suffer from the same traffic problem as every city does in the world. (<https://trafficindex.org/>) With how dense and yet large the city is, any number of factors contribute to this issue. A theory known as Braess’ Paradox, is when an additional road is added to a two-path network, connecting the faster portions of those two paths, the traffic flow of the system decreases rather than increases. NYC is made up of hundreds of facets of faster or shorter paths to reach a destination. Looking at this on a large scale, the roads that are causing any of this can be viewed when they are closed for construction or maintenance; or they are only issue roads during certain time periods of the day.

Traffic flow in NYC is tracked by the Department of Transportation for NYC through their Open Datasets. Using this information for the hour-by-hour analysis of traffic counts on roads in NYC over a large time period, problematic roads can possibly be identified through machine learning models, and then used to either eliminate such roads, or help NYC residents avoid them in general, even if it seems like those roads are faster from an individual standpoint.

**Problem**

(State the problem. Needs to be a clear AML problem.)

(Highlight most impressive result)

**Data**

The data set Traffic Volume Counts comes from the NYC Department of Transportation Open Data using Automated Traffic Recorders (ATRs). The data is made up of 42756 unique road/day combinations over the span of 10 years. Each row in the data set consists of identifier columns and the counts over each hour of the day.

(Explain your approach and alternatives)

(Explain the data at hand)

**Results**

(Explain why your approach was best)

**Conclusion**

(Talk about risks/limitations. Assess whether the ML method answered the problem, it can be possible that it does not.)

(Restate the conclusion)

**The following is simply an outline of the sections required for the final project:**

**Introduction**

(Introduce the topic and background info. Name hypothetical audience.)

You must identify an audience and articulate the potential value for that audience in the report.

Your audience can be an individual or an institution but should have some level of detail like “recent college grad” or “think tank for public policy on housing” rather than a hypothetical person/institution.

Value can be realized in the form of saved time, decreased uncertainty, or increased reward. I’m open to hearing other forms of value definition.

**Problem**

(State the problem. Needs to be a clear AML problem.)

Possible datasets:

<https://data.cityofnewyork.us/browse?Dataset-Information_Agency=Department+of+Transportation+%28DOT%29&sortBy=relevance&pageSize=20&limitTo=datasets>

<https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts/btm5-ppia/about_data>

Data mining:

Generate a non-obvious insight from the data that can be leads for further investigations (similar to the observation step in the scientific method). The most common way to do this is by merging datasets from different sources.

A metric or (ranking) algorithm that quickly filters or sorts large amounts of data for people.

Large-scale decision making:

Anomaly detection

Auto-labeling/tagging

A model that is inferring relationships between variables is not machine learning. An exception exists for machine learning models embedded in a causal inference framework.

(Highlight most impressive result)

**Data**

(Introduce data and/or model. Data must be big enough to warrant ML.)

Your data should be big, i.e. at least 2 datasets or a single large dataset for your project.

If you have two datasets you should articulate how the datasets complement each other, for example

One dataset could provide complementing features (e.g. one dataset may have the activities and another may have the user demographic data).

One dataset could provide more resolution (e.g. detailed surveys tend to be done less frequently so they’re often complemented with surveys that are less detailed but collected more frequently).

One dataset could provide more data from a different population (e.g. a system change has made historical patent data to be stored in a different system).

You should not call a dataset collected from a different time point but from the same source as a different dataset.

For single datasets that are truly large or complex, you should articulate why it is considered large/complex relative to the standards your audience is used to.

(Explain your approach and alternatives)

Please engineer at least one feature and evaluate its usefulness in the context of your project.

You must perform some sort of exploratory data analysis that checks the completion and/or quality of the data or the existence of the problem (e.g. market size assessment).

You must identify the source of the data and summarize the datasets.

(Explain the data at hand)

Ideally, the data source should come from the entity that manages the data. Management is defined by the entity responsible if the data has quality issues. For example, data aggregators like Kaggle do not manage the data it posts on its challenges but Twitter manages Twitter data. Some exceptions exist like platforms like NYC OpenData allow you to report data quality issues and will be considered a manager of the data.

This exploration should perform basic checks for quality (e.g. returns should have positive and negative values, population values should match expectations etc). This step should convince the reader that the data is correct.

You must use a machine learning algorithm (this does not have to be one we covered)

**Results**

(Explain why your approach was best)

Please have at least one graphical summary that highlights how your algorithm is working.

You must quantitatively evaluate how your algorithm fits to the data. For example, correlations are meaningful if they’re close to -1 or 1, clustering and supervised learning methods all have different metrics associated with them.

Please verify the results whether is due to chance or likely a real pattern.

For example, you may have to subset the data to see if the result is due to an outlier. You can also quote external sources to help understand the data.

Please make sure you quantitatively address the question “if you had a different dataset, how robust are the results from the algorithm?” (CV can address this)

You should compare to a “dumb yet reasonable” approach

**Conclusion**

(Talk about risks/limitations. Assess whether the ML method answered the problem, it can be possible that it does not.)

(Restate the conclusion)