Capstone 1 Proposal

Predicting Accidents in New York City

Gene Hopping

What is the problem I want to solve?

Road traffic accidents resulted in the death of over 37,000 people in the US in 2017 alone.[1] The cost associated with traffic accidents in 2010 (from a report published in 2015) was $242 billion, or 1.6% of the gross domestic product for that year.[2] Companies that make their living in transportation would benefit from information regarding where accidents are more likely to occur, so they can avoid those areas and decrease their risk of accidents. Rideshare companies, such as Uber, rely on getting from A to B quickly and *safely* in order gain customer satisfaction. Uber provide insurance for its drivers[3], bearing the costs of accidents so is at risk for more than loss of customer satisfaction if involved in an accident. This project will focus on predicting the increased risk of accident at specific locations throughout New York city at specific times during the day. Companies like Uber could use this information when generating routes to avoid areas of increased risk of accident.

What data are you using? How will you acquire that data?

The data I will use is the NYPD Motor Vehicle Collisions Data provided by the City of New York.

- https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

The data is provided as a CSV file and also via an API. Initially I plan to develop code using the CSV data with the aim to acquire the data through the API for the final model.

Briefly outline how you'll solve this problem

The data consist of the time of accident, the geographical coordinates, and other information such as cause of accident, vehicles involved, fatalities, etc. Data will be cleaned to ensure each entry contains valid coordinates and timestamps (it appears as though 00:00 is used as a placeholder for missing time data. This is important for time slices shorter than 1 day). At this stage I plan to represent the data as a raster to determine areas of increased risk of accident, at specific times during the day. A 2D matrix visualized over time is basically an image recognition problem, so I plan to investigate this with deep learning. There is precedence for an approach such as this.[4]

---

[1] https://www.nhtsa.gov/press-releases/us-dot-announces-2017-roadway-fatalities-down
[2] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013
[3] https://www.uber.com/drive/insurance/
[4] https://arxiv.org/pdf/1710.09543.pdf

What are your deliverables?
Deliverables include a jupyter notebook containing code, a report and a blog post.