Group 3: Arianna Dunham, RyeAnne Ricker, Lydia Teinfalt

DATS 6103

Dr. Jafari

March 18, 2021

Final Project Group Proposal

Problem Statement: What factors affect the occurrence of major injuries and fatalities incurred during a DC traffic accident?

As we start to look at the post-pandemic world and more Washingtonians are beginning to commute into DC area again, we want to use machine learning to understand factors that contribute to car crashes severity. There are 4.1 million cars registered in DC and surrounding Virginia and Maryland areas. [Source] With more cars coming back to the roads come more accidents. A report in 2010 by the NHTSA stated that the total economic cost of motor vehicle crashes in the United States was $242 billion. [Source] Additionally, the city of D.C. has implemented an initiative called Vision Zero where it plans to reach zero fatalities and serious injuries caused by traffic accidents in the year 2024. This is an ambitious goal, and our findings may help determine if the goal is reasonable. [Source]

Dataset:

We're using two datasets from the D.C. government—The District Department of Transportation (DDOT) and The Metropolitan Police Department (MPD). The first is titled *Crashes in D.C.* This dataset provides location for crashes that occurred in the District along with statistics about the crash including: the location of the crash, the number of injuries (major or minor), number of fatalities, types of road user involved (car, bike, pedestrian), if any of the involved parties were impaired, and if speeding was involved.

District Department of Transportation, Metropolitan Police Department, *Crashes in DC*, Open Data DC, (District of Columbia): Vision Zero Data Planning Work Group, 2020. Accessed on: Mar. 11, 2020. [online]. Available: **https://opendata.dc.gov/datasets/crashes-in-dc?geometry=-126.111%2C2.895%2C128.420%2C88.515**

The second dataset is also provided by DDOT and MPD and is a continuation of the *Crashes in D.C.* dataset. This second dataset—titled *Crashes Details Table*—provides more details for each crash including the age of those involved, the type of vehicle, the state the vehicle is registered in, and if a ticket was issued.

District Department of Transportation, Metropolitan Police Department, *Crashes Details Table*, Open Data DC, (District of Columbia): Vision Zero Data Planning Work Group, 2020. Accessed on: Mar. 14, 2020. [online]. Available: **https://opendata.dc.gov/datasets/crash-details-table**

Both datasets require some cleaning. Since the second dataset is a continuation of the first, many columns between the two sets are redundant and should be removed. Additionally, many of the columns in the *Crashes in D.C.* dataset provide the same information for use in different software (x,y coordinates, latitude/longitude, address, etc.). So, we'll only keep the columns that are relevant for us (latitude/longitude).

In addition to removing redundant columns, it was decided to perform a binary classification, using the groups "major injury/fatality" and "minor injury/no injury." Because of this, those columns must be combined.

Also, the erroneous and nuisance data must be cleaned out, such as the random data points from 1970 (prior to the start date of this data set) and reported times of accidents, as most all of them default to 5:00 am.

Lastly, any entries where the number of fatalities and injuries are unknown will be removed. Since fatalities and injuries are the main outcome we're looking at, we don't want to include entries with unknown values in those columns.

Models and Framework:

Logistic regression, random forest, k-means, DBSCAN.

Packages:

We plan on utilizing Numpy and Pandas for handling and combining the datasets. We will also use Matplotlib for data visualization. Finally, we plan on using Sklearn to implement our machine learning algorithms.

References:

L. Li, S. Shrestha and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques," 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, UK, 2017, pp. 363-370, doi: 10.1109/SERA.2017.7965753. Accessed Mar. 15, 2021.[Source]

A. S. Alotaibi, "Density-based clustering for road accident data analysis," http://www.science-gate.com/IJAAS.html, 13-Jun-2018. [Online]. Available: http://science-gate.com/IJAAS/Articles/2018/2018-5-8/14 2018-5-8-pp.113-121.pdf. Accessed March. 18, 2021. [Source]

James, G., Witten, D., Hastie, T., Tibshirani, R., *An Introduction to Statistical Learning - with Applications in R.* New York: Springer, 2013.

Barnet-Woods, Bryan, "How Roadway Capacity and Safety Intersect with Vision Zero Goals", Greater Washington, Oct. 12, 2020. Accessed Mar. 13, 2021. [Source]

Performance Analysis:

The performance will be judged using AUC. This metric considers false positives and false negatives to evaluate the model. An AUC closer to 1 indicates better performance, while one closer to 0 indicates poor performance.

Preliminary Schedule:

Week	Goal	Must Be Completed
March 21	Complete EDA and Combine	EDA
March 28	Data Preprocessing and Combine	Data Preprocessing
April 4	Catch-Up + Work of Written Portion	Finalized EDA, Preprocessing, Combining Scripts
April 11	Complete Data Mining and Results	Data Mining and Results
April 18	Complete Edits, Final Paper, Final Presentation	Edits, Final Paper, Final Presentation
April 25	Final Edits	Project + Presentation