MOTOR VEHICLE CRASHES IN THE STATE OF NEW YORK

TEAM MEMBERS

1. Abigail Allen – A02053771 (Undergrad)
2. Josh Urry – A02081283 (Grad Student)
3. Trevor Jex – A02283455 (Undergrad)

DESCRIPTION OF THE PROJECT

We are going to analyze data concerning motor vehicle crashes for the state of New York. We feel this information can be used to benefit the state of New York to determine what factors influence the type and severity of traffic accidents that occur.

We are going to start out by manually identifying patterns/correlations in the data. We will use our discoveries to inform our training of a machine learning model that can predict the type and severity of a collision based on road conditions, lighting conditions, time of day, and other background information about the crash.

We will build the machine learning model using scikit-learn. We will start out with logistic regression and move to other classification models/combinations of classification models based on our results.

DATASET

We are going to use a dataset that contains data collected by the State of New York about motor vehicle crashes. This dataset contains data starting in January 2017 all the way until the end of 2021.

The crash data collected by the State of New York includes a record for each individual crash, containing details, including but not limited to; whether property damage was caused, whether there was a death, lighting conditions, collision type, county name, road description, whether pedestrians were involved, etc. This data is available to us in CSV format through the following link:
https://data.ny.gov/Transportation/Motor-Vehicle-Crashes-Case-Information-Three-Year-/e8ky-4vqe

IMPLEMENTATION PLAN

We will start with data preprocessing and cleaning. A lot of the columns in the dataset are text descriptions. So part of the preprocessing will require us to code the columns in a numerical way that will allow the data to be fed into the model. We are tentatively planning to accomplish preprocessing by March 15th. This will allow us to get everything in order to start training the model by the time the first progress report is due.

We want to have our machine learning training/prediction mostly finished by April 15th, so we have time to write our report, decide what details we are going to highlight in our presentation, polish up the code, and create and submit our video by the due date of April 30th.

ROLE OF MEMBERS

Abigail Allen will be the project leader. She will be helping guide the project as well as help with the initial analysis of the data in preparation for our Machine Learning analysis and correlation.

Josh Urry will help with data cleaning and descriptive analyses. He will also assist Trevor with the machine learning model.

Trevor Jex will be the resident sklearn/Machine Learning guru. He took CS 4320 and is familiar with how to do data preprocessing, encoding, training and testing/validation using scikit-learn.

All members will aid in analyses, hypotheses, report and presentation creation.