Predicting Traffic Accidents in New York City

analog: dramesh1, mlunghi, rhill6, tliu46

Goal

Traffic accidents seriously injure approximately 3,000 people and kill approximately 200 people in New York City each year. These types of accidents are the leading cause of death for children under the age of 14. We aim to predict the number of accidents in NYC during any period of time given weather conditions, traffic speeds, location, and time of day. By doing so, we hope our results can be used by government programs, like Vision Zero NYC, to improve traffic safety by providing insights on factors that elevate the risk of accidents occurring.

Data

We collected NYC accident and traffic speed data from NYC OpenData. We decided to use only data points during the years 2018 and 2019. The rest were discarded. For the traffic speed dataset, this left ~26 million data points (~76% of total). For accident data, this left ~450 thousand data points (~27% of total). The main columns from the accident dataset were: time of accident, borough of accident, and number of people injured or killed. The exact coordinates of the accidents were given for less than 10% of the samples, so location grouping was by borough rather than exact location. The traffic speed dataset provided roughly hourly speed measurements on major arterials and highways in each borough of NYC. These data points were used to calculate the average speed in each borough during every hour of every day during 2018 and 2019. To supplement these two datasets, we also acquired hourly weather data for each borough during 2018 and 2019 using Dark Sky's API. The accident, traffic speed, and weather datasets were joined based on borough, date, and hour.

Model and Evaluation

Our goal is to predict the number of accidents in NYC in a specified hour given weather conditions, traffic speeds, location, and time of day. Since traffic flow is dependent on characteristics of a given city such as population density, availability of public transportation, and road infrastructure, we do not expect the results of our model to generalize well to other cities. However, we would expect to see features interact in similar fashions across cities. For example, if our model finds a positive relationship between snow and number of traffic accidents for NYC, we would expect this to also be the case for other cities (though the magnitude of the coefficient will likely vary). Our joined dataset consists of ~80 thousand data points. We used an 80/20 split for training and testing. Since features and results are not dependent on other time periods, we can split the dataset randomly. We train using a multiple linear regression model where we feature-scaled our quantitative columns and one-hot-encoded our qualitative columns. To evaluate performance, we created a baseline model that predicts by summing up all accidents during 2018 and 2019 and dividing by the total number of hours in these two years. We compared our model to this baseline by comparing MSE.

Results and Analysis

Claim #1: The multiple linear regression model outperforms the baseline model.
Support: When comparing our model to the baseline model, both train and test MSE drop by over 50%. The R-Squared value of our model is 0.553.

	Training MSE	**Testing MSE**
Baseline	10.552	10.752
Full Model	4.720	4.817

Claim #2: Weather data had a negligible impact on the model's MSE and R-Squared.
Support: When we remove weather data from our model, the R-Squared value doesn't change and the MSE only changes slightly. Furthermore, only visibility (p-value 0.04) and cloudiness (p-value of 0.00) were significant. The condition of snow (p-value 0.616) was insignificant.

	Training MSE	**Testing MSE**	**R-Squared**
Full Model without Weather Features	4.723	4.821	0.552
Full Model	4.720	4.817	0.553

Claim #3: Splitting examples by borough significantly improves model performance.
Support: Evaluating a multiple linear regression model without the one-hot encoded borough features, the model performs with an R-Squared of 0.286 and a MSE of ~7.5 for both train and test sets. This is in comparison to the full model, which is over 85% higher in R-Squared and around 33% lower in MSE.

Full Model Features

1. Note: each sample represents a one hour block in a specific borough of NYC. Bolded features are continuous (feature scaled), italicized features are one-hot encoded features (clear, Staten Island, hour 23, and weekend were arbitrarily chosen to be left out of the model due to reduced linear dependency)
2. Full Model Features: **precipitation intensity, wind speed, visibility, average traffic speed,** *rain, cloudy, snow, Bronx, Brooklyn, Manhattan, Queens, hour 0, hour 1, …, hour 22, weekday*