**Predictive modeling of motor accidents and weather in NYC**

Logan Herrera

September 20, 2020

1. **Introduction**
   1. **Background**

People all over the world are driving day-to-day. Whether you are the driver of a vehicle, a passenger, or even just a pedestrian traffic safety is important for all parties. Different factors play a role in traffic safety (weather, time of day, etc.) and we are interested in predicting different accidents’ severity and if it relates to weather. The Federal Highway Administration (FHWA), published Figure 1 depicting “Weather-Related Crash Statistics (Annual Averages):

|  |  |  |
| --- | --- | --- |
|  | **Weather-Related Crash Statistics** | |
| **Weather-Related\* Crashes, Injuries, and Fatalities** | **10- year Average (2007-2016)** | **10-year Percentages** |
| 1,235,145 crashes | 21% of vehicle crashes |
| 415,005 persons injured | 19% of crash injuries |
| 5,376 persons killed | 16% of crash fatalities |

* “Weather-Related” crashes are those that occur in the presence of adverse weather and / or slick pavement conditions.

*Figure 1.* Table of Weather-Related Crash Statistics. Adapted from “U.S. Department of Transportation: Federal Highway Administration” by Road Weather Management Program, *How Do Weather Events Impact Roads?* Retrieved September 15, 2020, from <https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm>

Looking at the data you can see that weather clearly influences accidents.

* 1. **Problem**

Driver safety is of utmost importance not only for the driver and any passengers but also for the other drivers and passengers on the road as well. Driving requires the full attention of a driver. Certain external factors may play a role in accident severity. Developing a model to examine accident severity based on current weather conditions may be able to change driver consciousness and provide a safer driving environment for all. The objective is to predict accident severity based on weather conditions.

* 1. **Interest**

Public decision-makers and anyone that interacts with motorized vehicles (drivers, passengers, pedestrians) would be extremely interested in an accurate prediction. This could help influence policies from local decision-makers but in a more generalized sense, this could also improve driver performance and increase awareness on all ends.

1. **Data acquisition and cleaning**
   1. **Data sources**

I will be examining data in New York City (NYC) for this instance. NYC Open Data (<https://opendata.cityofnewyork.us/>) provides open access to a multitude of datasets associated with NYC and all five boroughs. I will be using the “ NYPD Motor vehicle Collision – Crashes” dataset (<https://data.cityofnewyork.us/NYC-BigApps/NYPD-Motor-Vehicle-Collisions-Summary/m666-sf2m>) to examine a detailed list of accidents and their associated crash date, crash time, brough, and more detailed information about location and results (injuries or deaths). Historical weather data will be extracted from World Weather Online (<https://www.worldweatheronline.com/>). I am using this site because I have prior experience using it and know it provides accurate, well-organized data. The combination of both these should provide for the backing to create a model for predicting accident severity.

The crash data contains the following columns: crash date, crash time, brough, zip code, latitude, longitude, location, on street name, cross street name, off street name, number of persons injured, number of persons killed, number of ped. Injured, number of ped. Killed, number of cyclists injured, number of cyclists killed, number of motorists injured, number of motorist killed, contributing factor vehicle 1, contributing factor vehicle 2, colloniid, and vehicle types.

A screenshot of a social media post

Description automatically generatedWeather data is more straightforward as we know we are working with time, temperature, humidity, wind, wind speed, pressure, precipitation, and other conditions.

*Figure* A screenshot of a cell phone

Description automatically generated*2. Data frame information about weather data*

* 1. **Data cleaning**

A close up of a device

Description automatically generatedThe NYPD accident data has 1,711,577 different rows and 29 columns. I also checked the data types of each object to ensure I was interpreting everything correctly and could make changes if needed. A missing null value % chart was created to determine where the nulls were located and how it would affect the dataset. This helped determine categories such as off-street name, contributing factor vehicle 3, contributing factor vehicle 4, contributing factor vehicle 5, vehicle type code 3, vehicle type code 4, and vehicle type code 5s high null value percentages could be attributed to no additional vehicles being involved in the accident. Besides this, borough and zip code had 30.57% null values. This probably occurred from human error such as data entry error, or simply no entry. I also created heatmap of the null values to better visualize this information.

*Figure 3. Heatmap of null values in accident data frame*

* 1. **Feature selection**

A screenshot of a cell phone

Description automatically generatedAfter the initial data cleaning and joining of the tables, all columns were further inspected. The motor vehicle collision dataset required a lot of cleaning and manipulating. I dropped a few different columns (ZIP CODE, LATITUDE, LONGITUDE, LOCATION, ON STREET NAME, CROSS STREET NAME, VEHICLE TYPE 1, VEHICLE TYPE 2, VEHICLE TYPE 3, VEHICLE TYPE 4, and VEHICLE TYPE). After evaluation, due to the purpose of this analysis these columns were dropped. This was due to the excess amount of NaN in these columns. For most of it the data was not necessary in the report or it was not input correctly. I also chose to filter the crash data from 01/01/2015 to 01/01/2020 due to the robust amount of data going as far back as 2012.

*Figure 4. table examining the missing value % for each column*

Weather data was uploaded in the form of csv and contains columns such as: PRCP, SNOW, SNWD, TAVG, TMAX, TMIN, WSF5 that were kept. The columns WT01, WT02, WT03, WT04, WT05, WT06, WT07, WT08, WT09 were all dropped for containing predominately Nan values. I was also sure to check the data types and convert the date columns in both datasets to datetime. The motor vehicle collision data and the weather data were joined by matching weather conditions on coinciding accident dates. Following this the data was cleaned once more to ensure only necessary data columns were present.

*Figure* A screenshot of a cell phone

Description automatically generated*5. Bar chart of contributing factors*

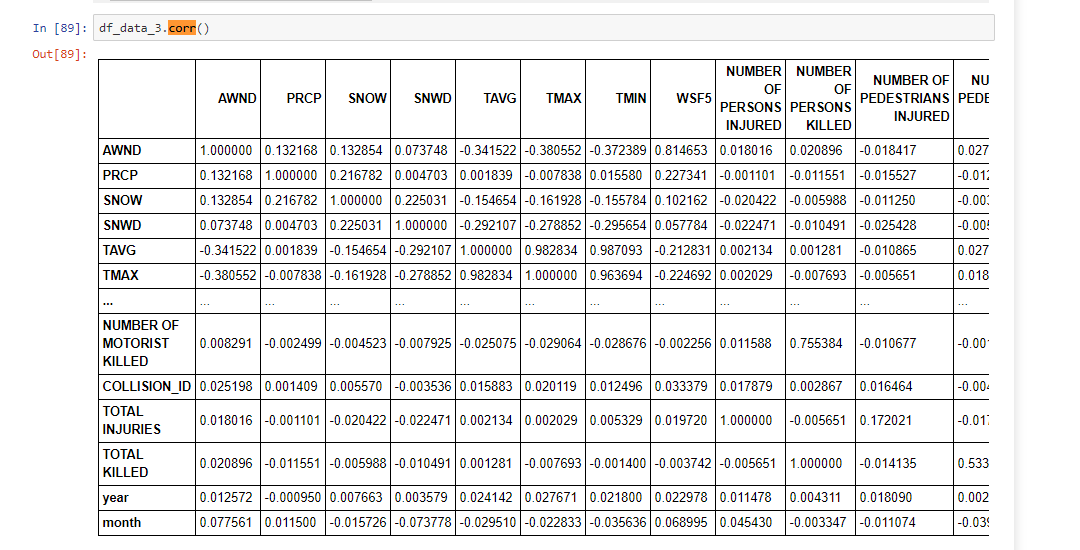
1. **Exploratory Data Analysis**
   1. **Calculation of target variable**

A screenshot of a cell phone

Description automatically generatedThe columns: NUMBER OF CYCLIST INJURED, NUMBER OF MOTORIST INJURED, NUMBER OF PEDESTRIANS INJURED, NUMBER OF CYCLIST KILLED, NUMBER OF MOTORIST KILLED, NUMBER OF PEDESTRIANS KILLED were summed up in their own respect and put in a condition statement that further evaluated whether the accident was “slight”, “serious” or “fatal. This was based on the condition that if it was slight no one was injured or killed, serious if it was only injuries and no fatalities, and fatal if there was at least one fatality. This is the variable we will used to model and determine accident severity and if there is any correlation to weather.

*Table 1. Analysis of accidents for each time frame based on data*

* 1. **Relationship between weather and accidents**

The weather data was further manipulated into more columns. Rain was split into light, moderate and heavy based on historical rain data in NYC as well as classifications for heavy-light rain fall. NYC has a continental climate and because of that I wanted to further split the dates into seasons. From the crash dates I further categorized the dates into seasons by separating the month number from the crash date, and writing a condition for a new columns that reads between 3(for March) and 5(for May) and classifies as Spring. This was completed for spring, summer, autumn, and winter. Correlation testing was run against all variables to get an initial understanding of the data. Nothing stuck out most variables had a negative correlation. This dataset is showing no relationship between the time of the year and accident severity.

*Figure 6. correlation with each variable*

* 1. **Relationship between accidents and wind**

A screenshot of a cell phone

Description automatically generatedLike the rain data the wind was split into categories based on historical average wind data in NYC. Heavy rain is classified as being >= 11 inches in a day, moderate as <= 11 inches and > 4 inches, and < inches as light rain. Splitting the data into subcategories made it visible there is little to not relationship between accident severity and wind conditions. I started by using “df.loc” to examine accidents that were fatal and coinciding wind conditions that were heavy. Of the records it returned 6 accidents matching the above conditions. On the other hand, when accident severity was “serious” and wind conditions were “moderate” we returned over 350 records. This was more common as compared to the previous result.

*Figure 7. Count of accident severity based on wind*

It was also found that when accident severity was light and when conditions were heavy there were over 600 accidents. Taking this into consideration along with the p-value will be beneficial moving forward.

* 1. **Relationship between accidents and season**

Climate data was divided into further columns based on seasons: “Spring”, “Summer”, “Autumn”, and “Winter”. The accident date was further divided into month and year, and by doing this I was able to condition sets of months to output seasons. Like the wind, there is no significant relationship between accident severity and the season of the accident. The incidents were dispersed, and no real inference can be made. I added this as an extra column I added during the analysis to see if/any relationship existed between the two. This step was part of the exploratory analysis process and determining what variables if/what variables stood out.

* 1. **Relationship between accidents and snow**

A picture containing drawing

Description automatically generatedAverage snowfall in NYC over the past five years was examined to create parameters. The 5-year maximum average in a day was 11.6 inches so “heavy” snow is greater than > 11 inches, moderate is between 11 inches and >= 4 inches, and light < 4 inches. Snow had no direct correlation with accident severity and this could be due to several reasons. One being that individuals may be more inclined to stay indoors and not drive at all during the winter.

*Figure 7. Count of accident severity*

* 1. **Relationship between accidents and crash time**

A picture containing drawing

Description automatically generated Crash time was further categorized into “morning” between 03:00 and 12:00, “afternoon” between 12:01 and 17:00, “evening” between 17:01 and 22:00, and “night” between 22:01 and 03:00. Splitting the time into categories enabled me to determine most accidents that occurred in the morning or the evening. This is likely due to a few reasons, one being these are peak traffic times you have individuals commuting to and from work. Also, as you would expect most of the accidents are serious, not fatal. This is also New York City, so it is certainly different than other locations. Speed limits are much lower here compared to other areas around the U.S. These lower speed limits were designed to avoid more fatal injuries if they are to occurs. Within you five boroughs you will find most local road having a speed limit of 30 miles per hour(mph).

*Figure 9. Count of time frame in which accident occurred*

1. **Predictive Modeling**

It was determined classification works best with our data. I’ll be using machine learning algorithms using Sckit-learn. As stated earlier, we want to differentiate between accident severity. We start by examining different algorithms and which works the best.

* 1. **Regression models**

Further analysis was performed to determine what tests may work for the dataset. Logistics regression was not feasible what the available data, so this was not an option. Training and test sets for created from the accident/weather dataset. Algorithms for the decision tree, logistics regression, and k-nearest neighbors was inapplicable for the dataset.

* 1. **Neural Network for accident severity**

The neural network was the preferred method here as it provided the opportunity for the most accurate prediction. Based on my research multi-layer perceptron would have likely worked based on the data. Due to constraints on testing capabilities and time spent on other testing methods I was unable to create this model. Being involved in traffic planning and engineering the question we examined piqued my interest and what a model like this can provide to others. I am currently in the process of improving this dataset so that I can use neural network estimation for accident severity.

Classification uses a multi-layer perceptron algorithm that trains data using backpropagation. MLP can fit a non-linear model to the training data. Feedforward processing follows, and this is a computation that produces an output value. Finally, followed up by backpropagation, so we are just moving in the reverse direction to effectively train the network.

I will be working on accident severity following this certification. I’m eager to create a fully functioning neural network based on the data.

1. **Conclusions**

Our data was not ideal for creating a model to predict accident severity. This could have been due to several reasons: filtering the data to only be between 2015- current may not have been as beneficial as using data between 2012-current. The problem here, too many missing values within the data from 2012-2015 to effectively use within the analysis. NYC is also unique in that it does not always follow statistical trends you see around the U.S. Machine learning can be extremely beneficial, especially in this case. It may require a vaster amount of data or more training to better predict accident severity. Having done this once now and working in transportation, I will be interested in running further models on this here in NYC and other locations.

1. **Future directions**

The machine learning did not go as expected. But at the there was a lot to learn from this experience and how the process may be improved. I was working in the right direction; the model could be improved with a more refined dataset. First, I think a more comprehensive dataset is overall much better. Living in NYC firsthand and examining the data, the input/model should be modified. Referring to the motor vehicle accident data there was a lot of missing data or incorrectly input data. A problem like this could be corrected in something as small as redesigning the database that the information is being inputted. By this, I mean rules can be implemented on the backend of the database to eliminate some of these input errors. The data was still comprehensive and able to be analyzed. But something as simple as this would have made for a better predictive model.

Additional information about the accident such as age-associated with the driver, and simply simplified list of reasons for the accident. This is not to say the weather is not related to accident severity, as I think it is clear from other studies that it certainly is a factor. Numerous academic studies suggest weather conditions do play a role in local/highway accident severity.

The analysis starts with sound and quality data. Our local government's data and even weather station data that is available is not at par. As data is continually incorporated in decision making the emphasis needs should be placed on the data collection stage of an organization. For proper analysis of the highest standard, it requires clean and well-organized data.