**Abstract**

Technological progress in the world has unarguably improved the quality of life for the average person in many ways. The age of the automobile has shaped the way in which work, play and live our lives. Roadways, buildings, cities and entire countries have been designed to accommodate motor vehicles. As automobile technology has advanced making cars faster and capable of more advanced maneuvers, so has our concern with the safety of these vehicles. Entire disciplines such as traffic management are devoted to optimizing numerous factors to ensure the safe and efficient movement of people and goods. As we move into the age of data, all stakeholders in the automobile industry must effectively collect and utilize the wealth of information available to better meet their goals if progress is to continue. In this project, we take the position of a law enforcement agency, the New York City Police Department, as they seek to best utilize their resources in the context of responding to traffic collisions in the city.

**Background**

At the end of 2017 in New York City, there were 1,923,041 cars registered to residents of the city. (<https://nyc.streetsblog.org/2018/10/03/car-ownership-continues-to-rise-under-mayor-de-blasio/>) This already-significant number does not include the heavy flow of vehicles of those who visit the city or are simply passing through. By contrast, the New York City Police Department (NYPD) budgets for a headcount of 35,822 uniformed officers (<http://council.nyc.gov/budget/wp-content/uploads/sites/54/2017/03/056-NYPD.pdf> - page 4), distributed across 77 police precincts (geographic divisions of the city). On-duty officers/traffic enforcement agents are allocated to each precinct to enforce traffic laws and handle emergency and administrative response to traffic incidents (such as collisions). NYC has been collecting traffic data, including specific data on vehicle collisions since 2014 to support “Vision Zero” , a traffic safety initiative which has the goal of eliminating traffic fatalities. (<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data>)

**Objective**

The objective of our analysis is to develop a supervised, prediction model using Machine Learning techniques and the CRISP-DM framework (cite textbook) on the available collision data to predict whether there will be a collision in a specified police precinct at a specified time. The intent of predicting this data is to inform the NYPD’s optimal assignment of limited officers and resources across the 77 police precincts.

**Data Analysis**

The data set that supports this analysis is source from the NYC Open Data project. The title of the data set is “Motor Vehicle Collisions – Crashes”. It contains entries for every collision recorded within New York City limits by NYPD agents beginning July 1st, 2012 up to the present day. There are approximately 1.65 million entries in the data set.

**Initial Data Exploration and Cleaning**

Based on what we know about the data set from the specifications at NYC Open Data and the data dictionary, we have decided to perform some initial cleaning steps.

Our analytics problem is to predict whether there will be a collision at a specific time (time including a time of the day, day of the year and calendar year). In this context, we will first look at the "CRASH.DATE" graph. Thinking about the scheduling of police resource, we assume that this happens in advance, on a hour-by-hour and day-by-day basis. We assume that resources are not scheduled on a year-by-year basis due to uncertainty in staffing, budget, etc. We therefore examine the data to see whether we should include the year at all. Including the year would treat the data set as a time-series, years ranging from 2012-2020. Alternatively, we could drop the year and group all occurrences on the same day in the same bin, possibly enhancing our prediction.

To decide, we plot the dates and look for trends. If trends repeat annually, we will drop the year as this trend will be preserved when we combine. If the trend does not repeat annually (extends over the whole range of dates) then we will not combine year as we will lose this information when dropping year.

[plot of all years]

The vertical black lines in the “All Years” plot represent the start of each year. As you can see from the plot, there is a noticeable repeating trend in each year (between the black lines) with a decrease in collisions at the start of each year, followed by various other increase/decreases.

As we are more interested in capturing this repeating annual trend than year-over-year changes, we will combine all data into a representation of one year.

Additionally, we will drop years 2012 and 2020 (the first and last years in the data set) to avoid over/under-representing specific months in the combined-year data set. This leaves us with the following “Years Combined” data set, plotted below.