# USA Car Accidents Severity Prediction

IBM Data Science Capstone Project

Nikita Kohli

## Introduction

- Motivation:
  The economic and societal impact of traffic accidents cost U.S. citizens hundreds of billions of dollars every year. And a large part of losses is caused by a small number of serious accidents. Reducing traffic accidents, especially serious accidents, is nevertheless always an important challenge. The proactive approach, one of the two main approaches for dealing with traffic safety problems, focuses on preventing potential unsafe road conditions from occurring in the first place. For the effective implementation of this approach, accident prediction and severity prediction are critical. If we can identify the patterns of how these serious accidents happen and the key factors, we might be able to implement well-informed actions and better allocate financial and human resources.
- Objective:
  The first objective of this project is to recognize key factors affecting the accident severity. The second one is to develop a model that can accurately predict accident severity. To be specific, for a given accident, without any detailed information about itself, like driver attributes or vehicle type, this model is supposed to be able to predict the likelihood of this accident being a severe one. The accident could be the one that just happened and still lack of detailed information, or a potential one predicted by other models. Therefore, with the sophisticated real-time traffic accident prediction solution developed by the creators of the same dataset used in this project, this model might be able to further predict severe accidents in real-time.

**Dataset Overview**

US-Accident dataset is a countrywide car accident dataset, which covers 49 states of the United States. It contains 3 million cases of traffic accidents that took place from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. In this project, only the data of accidents that happened after February 2016 and were reported by MapQuest was finally used in exploration analysis and modeling so that irrelevant factors can be eliminated to the greatest extent. Details about features in the dataset:

- Traffic Attributes (12):
  - ID: This is a unique identifier of the accident record.
  - Source: Indicates source of the accident report (i.e. the API which reported the accident.).
  - TMC: A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.
  - Severity: Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).
  - Start_Time: Shows start time of the accident in local time zone.
  - End_Time: Shows end time of the accident in local time zone.
  - Start_Lat: Shows latitude in GPS coordinate of the start point.
  - Start_Lng: Shows longitude in GPS coordinate of the start point.
  - End_Lat: Shows latitude in GPS coordinate of the end point.
  - End_Lng: Shows longitude in GPS coordinate of the end point.
  - Distance(mi): The length of the road extent affected by the accident.
  - Description: Shows natural language description of the accident.
- Address Attributes (9):
  - Number: Shows the street number in address field.

- Street: Shows the street name in address field.
- Side: Shows the relative side of the street (Right/Left) in address field.
- City: Shows the city in address field.
- County: Shows the county in address field.
- State: Shows the state in address field.
- Zipcode: Shows the zipcode in address field.
- Country: Shows the country in address field.
- Timezone: Shows timezone based on the location of the accident (eastern, central, etc.).

- Weather Attributes (11):
  - Airport_Code: Denotes an airport-based weather station which is the closest one to location of the accident.
  - Weather_Timestamp: Shows the time-stamp of weather observation record (in local time).
  - Temperature(F): Shows the temperature (in Fahrenheit).
  - Wind_Chill(F): Shows the wind chill (in Fahrenheit).
  - Humidity(%): Shows the humidity (in percentage).
  - Pressure(in): Shows the air pressure (in inches).
  - Visibility(mi): Shows visibility (in miles).
  - Wind_Direction: Shows wind direction.
  - Wind_Speed(mph): Shows wind speed (in miles per hour).
  - Precipitation(in): Shows precipitation amount in inches, if there is any.
  - Weather_Condition: Shows the weather condition (rain, snow, thunderstorm, fog, etc.).
- POI Attributes (13):
  - Amenity: A Point-Of-Interest (POI) annotation which indicates presence of amenity in a nearby location.
  - Bump: A POI annotation which indicates presence of speed bump or hump in a nearby location.
  - Crossing: A POI annotation which indicates presence of crossing in a nearby location.

- Give_Way: A POI annotation which indicates presence of give_way sign in a nearby location.
- Junction: A POI annotation which indicates presence of junction in a nearby location.
- No_Exit: A POI annotation which indicates presence of no_exit sign in a nearby location.
- Railway: A POI annotation which indicates presence of railway in a nearby location.
- Roundabout: A POI annotation which indicates presence of roundabout in a nearby location.
- Station: A POI annotation which indicates presence of station (bus, train, etc.) in a nearby location.
- Stop: A POI annotation which indicates presence of stop sign in a nearby location.
- Traffic_Calming: A POI annotation which indicates presence of traffic_calming means in a nearby location.
- Traffic_Signal: A POI annotation which indicates presence of traffic_signal in a nearby location.
- Turning_Loop: A POI annotation which indicates presence of turning_loop in a nearby location.

• Period-of-Day (4):

- Sunrise_Sunset: Shows the period of day (i.e. day or night) based on sunrise/sunset.
- Civil_Twilight: Shows the period of day (i.e. day or night) based on civil twilight.
- Nautical_Twilight: Shows the period of day (i.e. day or night) based on nautical twilight.
- Astronomical_Twilight: Shows the period of day (i.e. day or night) based on astronomical twilight.

Due to the limit of computer capacity, I am focusing on Montgomery County in the State of Pennsylvania. I will only select a few features I believe are more relevant to severity. Categorical data will be treated

with Pandas get_dummies method. Rows with missing values will be dropped.
Data source: https://www.kaggle.com/sobhanmoosavi/us-accidents