Predicting the possibility of getting into a car accident and the severity of the accident

2. Data and data preprocessing

2.1 Data source

The dataset was obtained from an online website.[1] The data was collected from February 2016 to June 2020 for the Contiguous United States. The number of observations in the data is about 3.5 million and each observation represents a traffic accident. The original data contains one label, i.e severity based on traffic delay, where 1 indicates the least impact on traffic and 4 indicates a significant impact on traffic. It also consists of 48 attributes.

2.2 Feature selection and data cleaning

There were several steps necessary to clean the dataset and make it usable for the problem statement:

First, not all of the attributes in the dataset are relevant for the problem statement covered in this report. Therefore 14 features were removed from the report entirely. Source just gave information on the source of the accident report, which is not important for the problem statement at hand. TMC and Description both feature a more detailed description of the event, which is not relevant as well. Distance gives information on the length of the road extent affected, which does help predict the likelihood of an accident happening. Airport Code denotes an airport based weather station closest to the accident, which is not important for the targeted prediction in this report. The same goes for Weather Timestamp, which is almost a duplicate of

[1] Data obtained from https://smoosavi.org/datasets/us_accidents, accessed September 25, 2020. Please also note:

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (2019).
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

the Start Time of the accident. Number, Street, Side, City, County, State, Zip Code and Country are duplicate data to the Latitude and Longitude of the accident location and was removed.

Second, the Boolean attributes Amenity, Bump, Crossing, Give Way, Junction, No Exit, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Signal and Turning Loop were assigned the object type and needed to be changed. Start Latitude, Start Longitude, End Latitude, End Longitude, Precipitation, Wind Chill, Temperature, Humidity, Pressure, Visibility and Wind Speed were also falsely assigned an object type and were changed to float. Finally, Start time and End time were assigned to a datetime type.

Third, the attribute Wind Direction included five duplicate values (out of the 24 unique values in total excluding NaN), so I combined each of those with their duplicating values to achieve a consistent taxonomy.

Fourth, 16 out of the remaining 35 attribute columns had missing values. The features End Latitude, End Longitude of the accident as well as Precipitation and Wind Chill were dropped entirely, because more than half of the values in the dataset were missing. Sunrise_Sunset, Civil Twilight, Nautical Twilight and Astronomical Twilight were all missing for the same exact 116 events of the dataset, so I dropped these accident events from the data set. For Temperature, Humidity, Pressure, Visibility and Wind Speed each missing value was replaced by the mean value of the attribute. For Wind Direction, Weather Condition and Timezone, missing values were replaced by the most frequent unique values. Finally, for Timezone, the missing values were derived from the Start Latitude and Start Longitude values.

In the end, the number of attributes was therefore reduced to 30 with 3,513,624 accident events. The label,severity based on traffic delay, remains the same.

The remaining attributes are as follows:

#	Attribute	Description	Example value
1	ID	A unique identifier of the accident record.	A-1
2	Start_Time	Shows start time of the accident in local time zone.	2016-02-08 05:46:00
3	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	2016-02-08 11:00:00
4	Start_Lat	Shows latitude in GPS coordinate of the start point.	39.865147
5	Start_Lng	Shows longitude in GPS coordinate of the start point.	-84.058723
6	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	US/Pacific
7	Temperature(F)	Shows the temperature (in Fahrenheit).	36.9
8	Humidity(%)	Shows the humidity (in percentage).	91.0
9	Pressure(in)	Shows the air pressure (in inches).	29.68
10	Visibility(mi)	Shows visibility (in miles).	10.0
11	Wind_Direction	Shows wind direction (e.g. SW).	SW
12	Wind_Speed(mph)	Shows wind speed (in miles per hour).	8.218939
13	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Light Rain
14	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	True
15	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	True
16	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	Trze

17	Give_Way	A POI annotation which indicates presence of give way in a nearby location.	True
18	Junction	A POI annotation which indicates presence of junction in a nearby location.	True
19	No_Exit	A POI annotation which indicates presence of no exit in a nearby location.	True
20	Railway	A POI annotation which indicates presence of railway in a nearby location.	True
21	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	True
22	Station	A POI annotation which indicates presence of station in a nearby location.	True
23	Stop	A POI annotation which indicates presence of stop in a nearby location.	True
24	Traffic_Calming	A POI annotation which indicates presence of traffic calming in a nearby location.	True
25	Traffic_Signal	A POI annotation which indicates presence of traffic signal in a nearby location.	True
26	Turning_Loop	A POI annotation which indicates presence of turning loop in a nearby location.	True
27	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Day
28	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	Day
29	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	Day
30	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	Day