Capstone 1 Project Proposal

Sara Satti

The problem

Car accidents in the US average 6million every year, with 3million injuries and over 30,000 fatalities. In fact they are the leading cause of death among healthy Americans. As a result, millions of dollars are spent in medical bills, rapid response teams deployment, and destroyed property.

In order to explore some of the causes, answers to the following questions will be sought:

The main question I wish to answer in this project is 'What factors determine the severity of an accident?' Severity here is a measure of the impact of the accident on traffic, where 1 is minor delays, and 4 causes closure of the road/highway.

Along the way, other information will be gleaned, such as: which states have the highest number of road accidents? Is the state with the highest number of accidents also the one with the most severe? Which time of day do those severe accidents happen? Is inclement weather a factor? Which of the recorded weather conditions have the most effect (e.g.high winds, temperature, pressure, precipitation etc.). Are accidents more prevalent on the left or right side of the street relative to the driver? What role do points of interest (amenities, bumps, traffic lights, junctions etc.) play?

Clients: department of transportation at the state and national levels would be interested in terms of road safety and budget allocations. Law enforcement, first responders and individuals would benefit from a model that can predict accident severity in terms of weather conditions, time of day and vicinity to points of interest.

Dataset: the dataset contains close to 3million accident records from 49 states in the United States. This Kaggle dataset is compiled by Sobhan Moosavi etal*, and is intermittently updated with new data. The version used for this study covers the period February 2016 untill December 2019, it was collated from several data sources, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.

The dataset contains a rich variety of data types to be experimented with. The figures below show data types present, and some statistics on the numerical data.

	count	mean	std	min	25%	50%	75%	max
TMC	2246264.0	207.831632	20.329586	200.000000	201.000000	201.000000	201.000000	4.060000e+02
Severity	2974335.0	2.360190	0.541473	1.000000	2.000000	2.000000	3.000000	4.000000e+00
Start_Lat	2974335.0	36.493605	4.918849	24.555269	33.550402	35.849689	40.370260	4.900220e+01
Start_Lng	2974335.0	-95.426254	17.218806	-124.623833	-117.291985	-90.250832	-80.918915	-6.711317e+01
End_Lat	728071.0	37.580871	5.004757	24.570110	33.957554	37.903670	41.372630	4.907500e+01
End_Lng	728071.0	-99.976032	18.416647	-124.497829	-118.286610	-96.631690	-82.323850	-6.710924e+01
Distance(mi)	2974335.0	0.285565	1.548392	0.000000	0.000000	0.000000	0.010000	3.336300e+02
Number	1056730.0	5837.003544	15159.278074	0.000000	837.000000	2717.000000	7000.000000	9.999997e+06
Temperature(F)	2918272.0	62.351203	18.788549	-77.800000	50.000000	64.400000	76.000000	1.706000e+02
Wind_Chill(F)	1121712.0	51.326849	25.191271	-65.900000	32.000000	54.000000	73.000000	1.150000e+02
Humidity(%)	2915162.0	65.405416	22.556763	1.000000	49.000000	67.000000	84.000000	1.000000e+02
Pressure(in)	2926193.0	29.831895	0.721381	0.000000	29.820000	29.980000	30.110000	3.304000e+01
Visibility(mi)	2908644.0	9.150770	2.892114	0.000000	10.000000	10.000000	10.000000	1.400000e+02
Wind_Speed(mph)	2533495.0	8.298064	5.138546	0.000000	4.600000	7.000000	10.400000	8.228000e+02
Precipitation(in)	975977.0	0.020495	0.235770	0.000000	0.000000	0.000000	0.000000	2.500000e+01

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2974335 entries, 0 to 2974334
Data columns (total 49 columns):
ID                      object
Source                  object
TMC                     float64
Severity                int64
Start_Time              object
End_Time                object
Start_Lat               float64
Start_Lng               float64
End_Lat                 float64
End_Lng                 float64
Distance(mi)            float64
Description             object
Number                  float64
Street                  object
Side                    object
City                    object
County                  object
State                   object
Zipcode                 object
Country                 object
Timezone                object
Airport_Code            object
Weather_Timestamp       object
Temperature(F)          float64
Wind_Chill(F)           float64
Humidity(%)             float64
Pressure(in)            float64
Visibility(mi)          float64
Wind_Direction          object
Wind_Speed(mph)         float64
Precipitation(in)       float64
Weather_Condition       object
Amenity                 bool
Bump                    bool
Crossing                bool
Give_Way                bool
Junction                bool
No_Exit                 bool
Railway                 bool
Roundabout              bool
Station                 bool
Stop                    bool
Traffic_Calming         bool
Traffic_Signal          bool
Turning_Loop            bool
Sunrise_Sunset          object
Civil_Twilight          object
Nautical_Twilight       object
Astronomical_Twilight   object
dtypes: bool(13), float64(14), int64(1), object(21)
memory usage: 853.8+ MB
```

Approach: An Exploratory Data Analysis will be conducted to aid in determining the relationships between the various attributes available in the dataset. Once these relationships are understood, a predictive model will be created to predict the severity of the accident.

Since latitude and longitude data is available, maps will be plotted to show states with highest accidents rates, and for a selection of states, also the specific areas within those states where the severest accidents happen.

Deliverables: the code will be stored in a Github repository, as well as a slide deck containing project methodology and summary.