Predicting Texas automobile accidents

Abstract:

Reducing road accidents is a critical issue for public safety. However, the majority of traffic accident analysis and prediction studies have used small-scale datasets with limited coverage, limiting their impact and applicability; and existing large-scale datasets are either private, old, or do not include important contextual information such as environmental stimuli, limiting their impact and applicability (weather, points-of-interest, etc.).US-Accidents now contains information around 1 million traffic accidents that occurred within the contiguous United States in the previous three years. Location, time, natural language description, weather, time of day, and points of interest are all intrinsic and contextual elements of each accident report. In this study, we offer this dataset as well as a variety of conclusions obtained from it regarding the spatiotemporal aspects of accidents. The goal of the project is to predict the severity of the accident.

Data source and Description:

We have taken the data source from Kaggle (https://www.kaggle.com/sobhanmoosavi/us-accidents). The dataset covers 49 states of the US. The below following table shows the description of the dataset.

No	Attribute	Description
1	ID	This is a unique identifier of the accident record.
2	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).
3	Start_Time	Shows start time of the accident in local time zone.
4	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.
5	Start_Lat	Shows latitude in GPS coordinate of the start point.
6	Start_Lng	Shows longitude in GPS coordinate of the start point.
7	End_Lat	Shows latitude in GPS coordinate of the end point.
8	End_Lng	Shows longitude in GPS coordinate of the end point.
9	Distance(mi)	The length of the road extent affected by the accident.
10	Description	Shows natural language description of the accident.
11	Number	Shows the street number in address field.
12	Street	Shows the street name in address field.
13	Side	Shows the relative side of the street (Right/Left) in address field.
14	City	Shows the city in address field.
15	County	Shows the county in address field.
16	State	Shows the state in address field.
17	Zipcode	Shows the zipcode in address field.
18	Country	Shows the country in address field.
19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).
20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.
21	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).
22	Temperature(F)	Shows the temperature (in Fahrenheit).
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
24	Humidity(%)	Shows the humidity (in percentage).

25	Pressure(in)	Shows the air pressure (in inches).
26	Visibility(mi)	Shows visibility (in miles).
27	Wind_Direction	Shows wind direction.
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).
29	Precipitation(in)	Shows precipitation amount in inches, if there is any.
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
31	Amenity	Shows presence of amenity in a nearby location.
32	Bump	Shows presence of speed bump or hump in a nearby location.
33	Crossing	Indicates presence of crossing in a nearby location.
34	Give_Way	Indicates presence of give_way in a nearby location.
35	Junction	Indicates presence of junction in a nearby location.
36	No_Exit	Indicates presence of no_exit in a nearby location.
37	Railway	Indicates presence of railway in a nearby location.
38	Roundabout	Indicates presence of roundabout in a nearby location.
39	Station	Shows presence of station in a nearby location.
40	Stop	Shows the presence of stop in a nearby location.
41	Traffic_Calming	Indicates presence of traffic_calming in a nearby location.
42	Traffic_Signal	Indicates presence of traffic_signal in a nearby loction.
43	Turning_Loop	Indicates presence of turning_loop in a nearby location.
44	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
45	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
46	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.

Goal of the Project:

The goal of the project is to predict the severity of the accident using Classification and find the hotspot areas where the accidents are happening so that emergency services can be routed/dispatched. We hoped that the analysis with the data exploration and model building will answer the following questions:

Data Exploration – We will use summary statistics and visualization tools to examine the following questions:

1. What is the trend in total accidents year on year in different severity categories?
2. The relationship between the severity of the accident (column: Severity) and various attributes, such as the weather, the location, time of the day, POI annotations on bumps, crossings, give way, stop signs and junctions, etc.

Prediction Model – We will apply different classification models to predict the severity of the severity level of the accident.

1. Data Understanding and Preparation: Here we can perform data cleaning, merging, exploring, normalization, removing/fixing missing values, feature engineering, etc.
2. Modeling: Apply classification algorithms to predict the severity of the accidents (column: Severity) based on available and important attributes learnt from data exploration.
3. Validation: Validate the model performance on the test data, and generate performance metrics.
4. Conclusion: What do we learn from the model in car accident prediction? How do we apply the model to guide driving in different conditions?