Stats 581 - Project Proposal
Project: USA Car Accident Severity Prediction

Team member's name	NetID
Utsav Patel	upp10
Manan Vakta	mv651
Sneh Desai	sd1324
Harshal Sinha	hs1030

Motivation:

According to statistics from this website, there are 6 million car accidents on an average in the U.S.A. every year. More than 90 people die in car accidents everyday. 3 million people in the U.S. are injured every year in car accidents. Reducing this number would drastically help improve people's lifestyle. To achieve this, we want to find the main factors that affect the severity of an accident. By identifying and controlling these factors, we should be able to reduce the overall severity of an accident.

Goal:

To identify key factors that contribute to high severity accidents by predicting severity of accidents.

Data Description:

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2020, using multiple 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. There are about 1.5 million accident records in this dataset. Each row contains 47 different features. Description of each feature is given below:

Feature name	**Description**
ID	This is a unique identifier of the accident record.
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 the local time zone.
End_Time	Shows the end time of the accident in the local time zone.
Start_Lat	Shows latitude in the GPS coordinate of the start point.
Start_Lng	Shows longitude in the 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 was affected by the accident.
Description	Shows natural language description of the accident.
Number	Shows the street number in the address field.
Street	Shows the street name in the address field.
Side	Shows the relative side of the street (Right/Left) in the address field.
City	Shows the city in the address field.
County	Shows the county in the address field.
State	Shows the state in the address field.
Zip Code	Shows the zip code in the address field.
Country	Shows the country in the address field.
Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).
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 the 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.).
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 the presence of a junction in a nearby location.
No_Exit	A POI annotation which indicates the presence of no_exit sign in a nearby location.
Railway	A POI annotation which indicates the presence of a railway in a nearby location.
Roundabout	A POI annotation which indicates the presence of a roundabout in a nearby location.
Station	A POI annotation which indicates the presence of a station (bus, train, etc.) in a nearby location.
Stop	A POI annotation which indicates the presence of a 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.

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.

Planning for the next steps of the project:

1. Analyse input features and conduct EDA (Exploratory Data Analysis).
2. Feature Engineering (fill/remove Nans with different techniques).
3. Splitting the data into train, test and validation.
4. Train different classification models:
    a. Support Vector Machine
    b. Random Forest Classifier
    c. Gradient Boosting Classifier
    d. Ada Boosting Classifier
    e. Extreme Gradient Boosting Classifier
5. Select the best model from the above models.
6. Model evaluation.
7. Results:
    a. Accuracy, Average Categorical accuracy, confusion matrix for train, test and validation set.
    b. Identify key features of interest.