APPLIED DATA SCIENCE CAPSTONE PROJECT

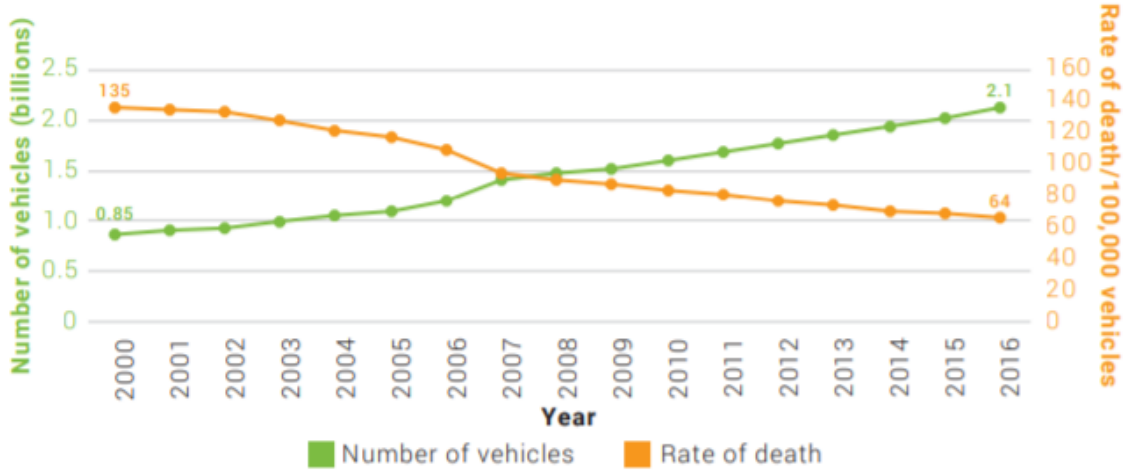
PREDICTION OF CAR ACCIDENTS SEVERITY PREDICTION IN US

By. Sivasankari Balasubramanian

Introduction Business Understanding:

Road accidents are serious concern for most of the nations around the world because accidents can cause severe injuries and fatalities. Traffic accidents are the leading causes beyond death. According to the World Health Organization's Global status report on Road Safety 2018, number of road accident deaths are continued to climb, reaching 1.35 million in 2016. Road traffic injuries are the eighth leading cause of death for all age groups. According to the report Road traffic injuries are currently the leading cause of death for children and young adults aged 5–29 years and the second leading cause of death worldwide amongst children ages 5-14. Shockingly, crashes account for 2.2% of all deaths around the world. The World Health Organization predicts that at the current rate, car accidents are likely to become the fifth leading cause of death globally by 2030.

Figure1: Number of motor vehicles and rate of road traffic death per 100,000 vehicles: 2000–2016



Sources: GLOBAL STATUS REPORT ON ROAD SAFETY 2018

According to The National Safety Council, more than 40,000 people in the U.S. are killed in car crashes every year. They also estimate that 4.57 million people were sustained seriously enough injuries require medical. The total cost to society for these crashes comes to about $413.8 billion a year.

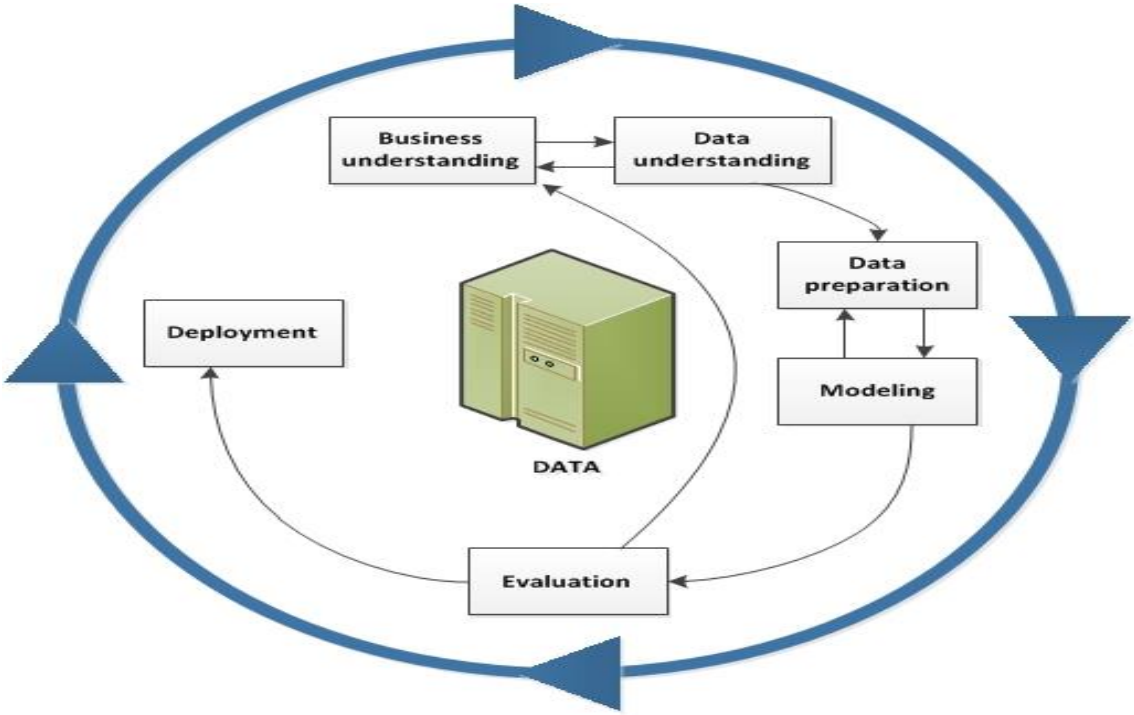
Reducing the traffic Incidents are always an important challenge. If we identify the patterns of how these severe accidents happen and the key factors, we might be able to implement the effective safety measures which are highly impact on human deaths, severe injuries and social economic loss. These safety measures will help the Traffic control Authorities, Transportation Department to impose and regulate the Traffic rules on Road safety and identify the Accident-prone areas.

Objective:

Our motivation is to predict the accident severity of any road, weather conditions, and the environment which are playing key roles in the collision. Our first aim is to identify the key factors causing the accidents as mentioned in second and third phases (i.e., Data Understanding and Data Preparation) of Data science methodology (Figure 2) and the second one is developing a model that can accurately predict the severity of the accidents.

Data cleaning is performed to identify and handle the corrupt and missing records. Further we are using some classification algorithms and evaluation methods to predict the severity. Here we are using SVM, LR, KNN and Decision tree classification models to predict the severity.

Figure 2: Six phases of Data science methodology. CRISP - DM



Source:
https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_crispdm_ddita/clementine/images/crisp_process.jpg

Data Understanding:

Dataset:

US accidents Dataset with 3.5 million records and 49 columns including weather conditions, Turning loop, wind speed etc collected from Kaggle is used in this project. It covers countrywide car accidents which includes 49 states of the USA. The accident data are collected from February 2016 to June 2020.

Dataset Link: https://www.kaggle.com/sobhanmoosavi/us-accidents

Attributes Explanation:

Traffic Attributes:

ID - Accident Record unique identifier.

Source - Source of the accident report (i.e. the API which reported the accident).

TMC - Traffic Message Channel code providing detailed description of the incident.

Severity - severity of the accident, a number between 1 and 4, where 1 is 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 - start time of the accident.

End_Time - refers to when the impact of accident on traffic flow was dismissed.

Start_Lat - Shows latitude in GPS coordinate of the start point.

Start_Lng - Shows longitude in GPS coordinate of the start point.

End_Lat - latitude in GPS coordinate of the end point.

End_Lng - longitude in GPS coordinate of the end point.

Distance(mi) - length of the road where the accident happens.

Description - Description of accident.

Address Attributes:

Number - Street number in the address field.

Street - Street name.

Side - Relative side of the street (Right/Left) in address field.

City - City name.

County - County name.

State - State name.

Zipcode - Zipcode in address field

Country - Country name.

Timezone - timezone based on the location of the accident (eastern, central, etc).

Airport_Code - airport-based weather station which is the closest one to location of the accident.

Weather Attributes:

Weather_Timestamp - timestamp of weather observation.

Temperature(F) - temperature (in Fahrenheit).

Wind_Chill(F) - wind chill (in Fahrenheit).

Humidity(%) - Humidity(in percentage).

Pressure(in) - air pressure (in inches).

Visibility(mi) - Visibility (in miles).

Wind_Direction - Wind direction.

Wind_Speed(mph) - wind speed (in miles per hour).

Precipitation(in) - precipitation amount (in inches), if there is any.

Weather_Condition - weather condition (rain, snow, thunderstorm, fog, etc.).

Point-Of-Interest 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 - period of day (i.e. day or night) based on sunrise/sunset.

Civil_Twilight - period of day (i.e. day or night) based on civil twilight.

Nautical_Twilight - period of day (i.e. day or night) based on nautical twilight.

Astronomical_Twilight - period of day (i.e. day or night) based on astronomical twilight.