Title: Capstone Project-Car Accident Severity Prediction

1. Introduction:

1.1 Background:

- As road accidents, injuries & deaths is the global problem that every country is facing. Throughout the world, roads are shared by cars, buses, trucks, motorbikes, mopeds, pedestrians, animals & other travelers.
- **Travel made possible by motor vehicles supports economic & social development** yet each year, vehicles are involved in crashes are responsible for millions of deaths and injuries.
- Road accidents/injuries are estimated to be 8th leading cause of death globally; it is more than from HIV/AIDS.
- **Road traffic injuries place a huge burden on low and middle income countries.**
- India, as a developing country, I'd like to work on this data to provide some practical solutions.

1.2 Problem:

Data might contribute to determine severity of road accidents, according to different factors like Road conditions, light conditions, weather condition, and speeding.

This project aims to predict severity of road accidents according to severity code **3-Fatality, 2b-Serious injury, 2-injury, 1-property damage, 0-unknown.**

1.3 Interest:

People travelling, goods carriers will be interested to know the traffic condition so as to save the time.

Different government project contractors will be interested.

Cities in the country will be smart to implement these ideas.

2. Data Acquisition & Cleaning:

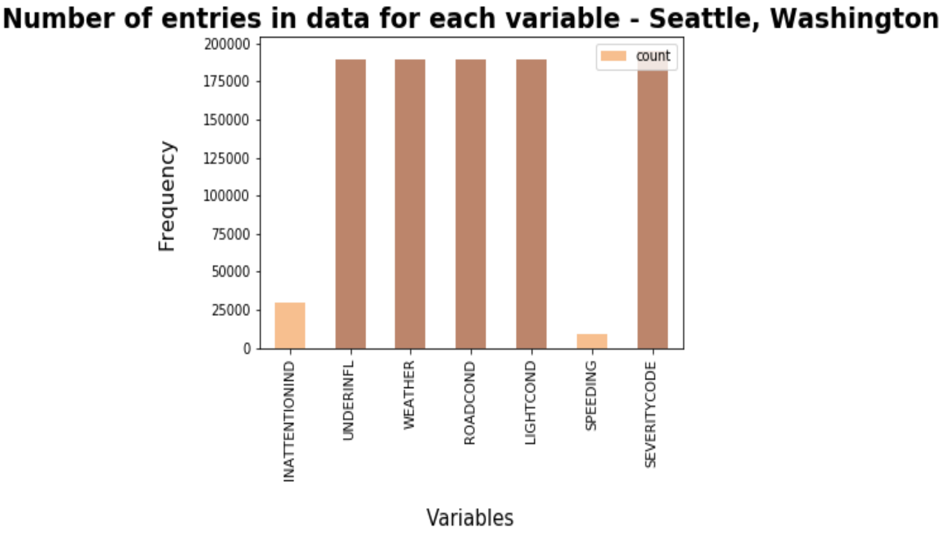
2.1 **Data Sources**:

Data used- Data_Collisions.csv provided in the course itself.
Data given in Capstone Project Metadata pdf.

2.2 Cleaning:

I downloaded the given data,

- Imported basic libraries like numpy, pandas.
- Used basic concepts of EDA to get acquainted with data
- There were a lot of missing values. Replaced them with mode (most frequent value)
- Deleted features due to redundancy (like SEVERITYCODE.1)



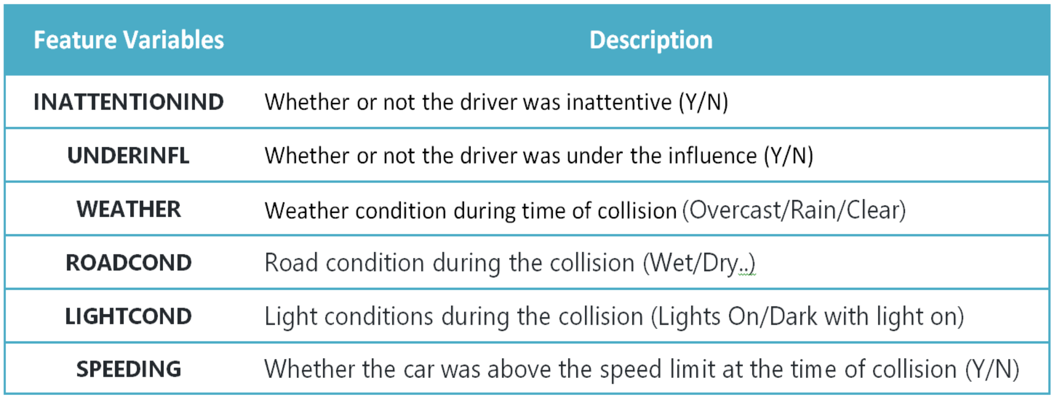
n order to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had *Other* and *Unknown* in them. This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.

3. Exploratory Data Analysis:

3.1 Feature Selection:

Target variable- y= SEVERITYCODE

Independent variables- X= INATTENTIONIND, UNDERINFL, WEATHER, ROAD CONDITION, LIGHT CONDITION, SPEEDING



Feature Variables	Description
INATTENTIONIND	Whether or not the driver was inattentive (Y/N)
UNDERINFL	Whether or not the driver was under the influence (Y/N)
WEATHER	Weather condition during time of collision (Overcast/Rain/Clear)
ROADCOND	Road condition during the collision (Wet/Dry..)
LIGHTCOND	Light conditions during the collision (Lights On/Dark with light on)
SPEEDING	Whether the car was above the speed limit at the time of collision (Y/N)

It is seen that the dataset is *supervised* but an *unbalanced* dataset where the distribution of the target variable is in almost 1:2 ratio in favor of property damage. It is very important to have a balanced dataset when using machine learning algorithms. Hence, **SMOTE** was used from imblearn library in order to balance the target variable in equal proportions in order to have an unbiased classification model which is trained on equal instances of both the elements under severity of accidents.

4. Modeling:

- **Logistic Regression:** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.

In Logistic Regression, independent variables, if categorical they should be converted into dummy or indicator coded.

- **Decision Tree Analysis:** The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

5. Results:

5.1 Decision Tree

	Precision	recall	f1-score	support
0	0.64	0.72	0.68	33903
1	0.44	0.34	0.39	21348
Accuracy			0.58	55251
macro avg	0.54	0.53	0.53	55251
weighted avg	0.56	0.58	0.56	55251

Accuracy of Decision tree is- **0.5760076740692476**

5.2 Logistic Regression

	Precision	recall	f1-score	support
0	0.72	0.67	0.69	38445
1	0.35	0.41	0.38	16608
Accuracy			0.59	55251
macro avg	0.53	0.54	0.53	55251
weighted avg	0.61	0.59	0.60	55251

Accuracy of Logistic Regression- **0.5888219217751715**

Accuracy is better for **Logistic Regression** model.

6. Recommendations:

After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders. The developmental body for Seattle city can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.