Capstone Sprint 1

By : John Narvaez

Date: July 28, 2021

Overview:

- What is my dataset ?

- The problem and ideas of addressing the problem.

- Preliminary EDA and findings.

- Concerns.

- Next steps.

Data Set:

- It is a United States country wide collection from 49 states. It has continuously collected starting from February 2016.
- Gathering is done by several providers and multiple Application Programming Interface (API) that provide streaming of data.
- As of March 2023 it has gathered around 7.7 M accidents.
- It has about 47 features.
- Taken from Kaggle and was uploaded by a Scientist Sobhan Moosavi.

Problem:

- Based on the research of the National Highway Traffic Safety Administration (NHTSA) that in 2019.

  - **$340 Billion** was the estimated cost of motor vehicle crashes.

  - **Estimated of $1k** for each of the 328 Million United States people.

  - **1.6 % of the $21.4 trillion** U.S gross domestic product.

  - **Societal harm** was nearly **$1.4 trillion.**

    Example:

    - Medical Cost, lost of productivity, emergency services, legal and court cost, congestion cost, property damage and alike.

Idea of solving the Problem:

Find a way to reduce or mitigate the problem.

How?

- Looking in the data for any trend and patterns.

- Study the statistical records.

- Use **machine learning** and modelling to predict future incidents.

Preliminary EDA

- The raw data consist of features about
  - Date of incident.
  - Geo location.
  - State, city and some address information.
  - Weather.
  - Presence description of the accident area.

- Concerns:
  - File size is 2GB and some irrelevant features.
  - Missing data and how to address them.
  - Duplicate rows.

Step taken / Next steps

- Identify **the target or feature to predict** which is the **“Severity”.**

- **Presence of an attribute** nearby the incident as the dependent variables.

- Dropped irrelevant columns (ex. End Time, End Lat & Lng) & formatting the data.

- Next step look into or analyze the target and remaining features.

- Further conversion of some Categorical data to numeric data??

- What would be the appropriate machine learning tool?

Thank You !