AI5000: Foundation of Machine Learning (FoML)

Hackathon

Deevanshu Gupta (SM21MTECH12014)

Shridharam Tiwari (SM21MTECH12003)

Data handling

1. First removing those features which non important and doesn't make sense for driver to be at fault

   'Report Number', 'Local Case Number', 'Agency Name', 'Off-Road Description', 'Municipality', 'Person ID', 'Vehicle ID', 'Vehicle Year', 'Vehicle Make', 'Vehicle Model', 'Equipment Problems'.

2. Checking the missing values

1.	Route Type	9.65
2.	Road Name	8.74
3.	Cross-Street Type	9.75
4.	Cross-Street Name	8.78
5.	Related Non-Motorist	92.92
6.	Collision Type	0.58
7.	Weather	7.82
8.	Surface Condition	11.42
9.	Light	1.37
10.	Traffic Control	15.10
11.	Driver Substance Abuse	17.66
12.	Non-Motorist Substance Abuse	93.43
13.	Circumstance	77.18
14.	Drivers License State	4.91
15.	Vehicle Damage Extent	0.77
16.	Vehicle First Impact Location	0.43
17.	Vehicle Second Impact Location	0.45
18.	Vehicle Body Type	1.52

19. Vehicle Movement	0.38
20. Vehicle Continuing Dir	2.19
21. Vehicle Going Dir	2.16

Features like Related Non-Motorist, Non-Motorist Substance Abuse, Circumstance have a lot of null values, but could be very important for the model to know the situation, so we can check for other feature selection prospective then, in the end we can reach to sensible decision.

3. Feature selection using variance threshold, We are not using sklearn, variance threshold for this since, because that needs data to be integers.

4. Now checking the correlation matrix

5. Checking data on catplot.

HANDLING MISSING VALUES(Encoding them)

1. Mode
2. And unknown

FINAL FEATURES

'ACRS Report Type',

'Related Non-Motorist',

'Collision Type',

'Surface Condition',

'Light', 'Traffic Control',

'Driver Substance Abuse',

'Vehicle Damage Extent',

'Vehicle First Impact Location',

'Vehicle Second Impact Location',

'Vehicle Movement',

'Vehicle Going Dir',

'Vehicle Continuing Dir',

'Speed Limit'

Classifier

XGBOOST

Its name stands for **eXtreme Gradient Boosting**, it was developed by Tianqi Chen and now is part of a wider collection of open-source libraries developed by the Distributed Machine Learning Community (DMLC). XGBoost is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit of memory and hardware resources for tree boosting algorithms.