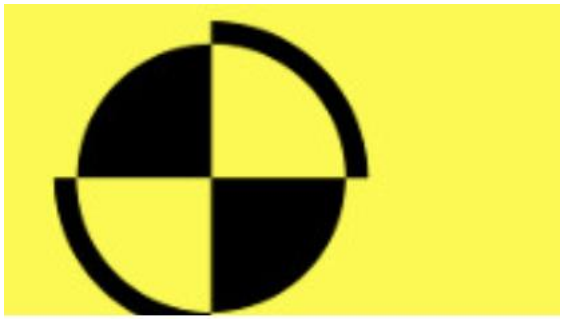
Predicting Primary Car Crash Cause

Phase 4 Project.

Datasets.

Traffic Crashes - Vehicles

Traffic Crashes - People



Traffic Crashes - Crashes

CHICAGO DATA PORTAL

Datasets.

CAR CRASHES - CRASHES

CAR CRASHES - VEHICLES

CAR CRASHES-PEOPLE

Traffic Crashes - People

- Dataset size: 1877321x29.
- Driver’s age, sex, physical condition.
- Passengers, Pedestrians, Cyclists.
- vehicle_id, crash_record_id.

CHICAGO
DATA PORTAL

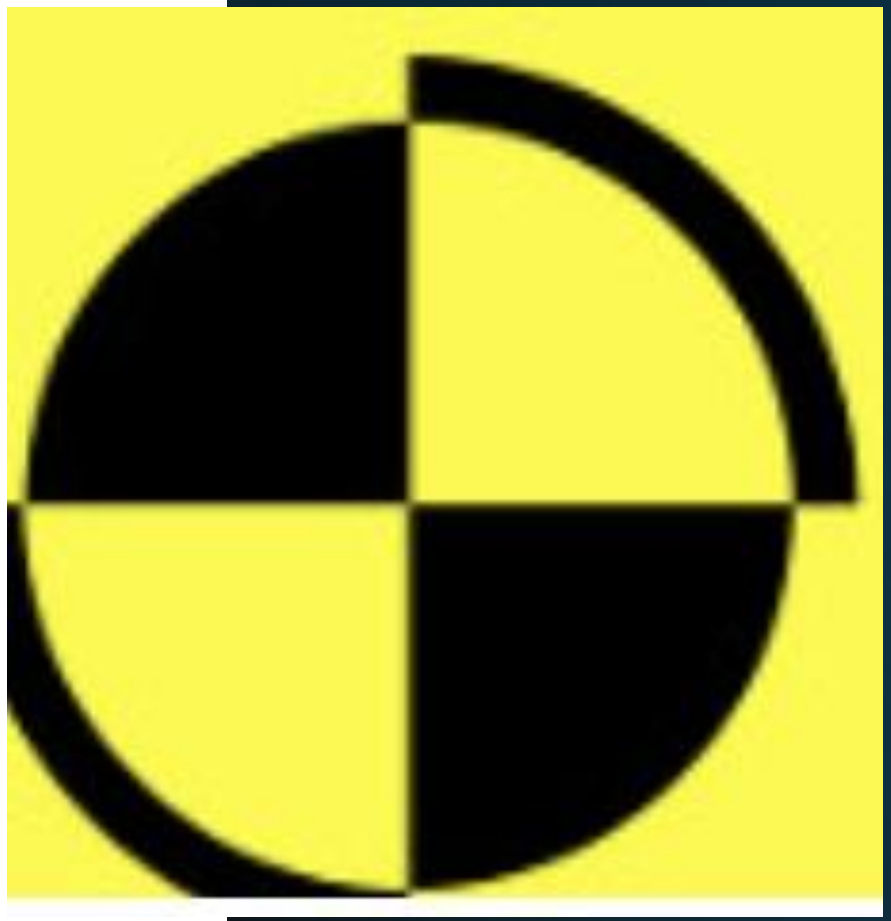
Traffic Crashes - Vehicles

- Dataset size: 1743922 x 71.
- Vehicle Type, Make, Model.
- Vehicle Use.
- vehicle_id, crash_record_id.

CHICAGO
DATA PORTAL

Traffic Crashes - Crashes

- Dataset Size: 854910 x 48.
- Crash scene related information: weather, roadway, signals, etc..
- Primary and Secondary Crash Causes.
- crash_record_id.



CHICAGO
DATA PORTAL

Dataset issues.

Inconsistent Data Logging.

Missing Information.

Number of features.

How are we going to approach the problem?

1. Identify relevant to Primary Contributory Cause columns in each dataset.
2. Explore and clean relevant information.
3. Merge Crashes-People and Crashes-Vehicles on vehicle_id.
4. Merge Crashes-Crashes on crash_record_id.
5. Perform EDA.

Transforming Datasets

People:1877321x29.

Vehicles: 1743922 x 71.

Crashes: 854910 x 48.

Resulting Dataset: 1541625 x 60.

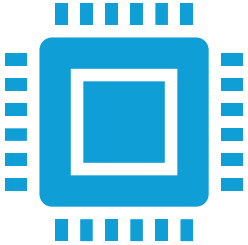
Exploratory Analysis.

- Most of the crashes occur in clear weather, have no injures, and are in 30-35 mph speed limit zone.
- Most of the crashes involve 2 cars, followed by one car.
- Most frequent type of crash is parked vehicle followed by rear-ending.
- Most typical environment is an undivided road without traffic controls.
- 2/3 of cashes in our dataset have Unknown or Not Applicable primary crash cause.
- Women are 2x as likely to be labeled “distressed”.
- Men are 2.5x more impaired.

Preparing Dataset for ML.

Reducing both features and primary contributory causes: 619627 samples x 22 features, with target containing 6 classes.

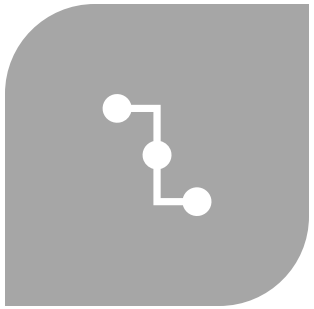
Use 15% of the data as a holdout set.



Use 30% of the data for training and testing (70%-30%).

Modelling Results.

XGBOOST (71%).



KNN (70%).

RANDOM FOREST (70%).

DEEP NEURAL NETWORK (69%).

XGBoost

- Accuracy 72%.
- Most important feature: first_crash_type_Rear End