Chicago Traffic Crashes

Multiclass Classification

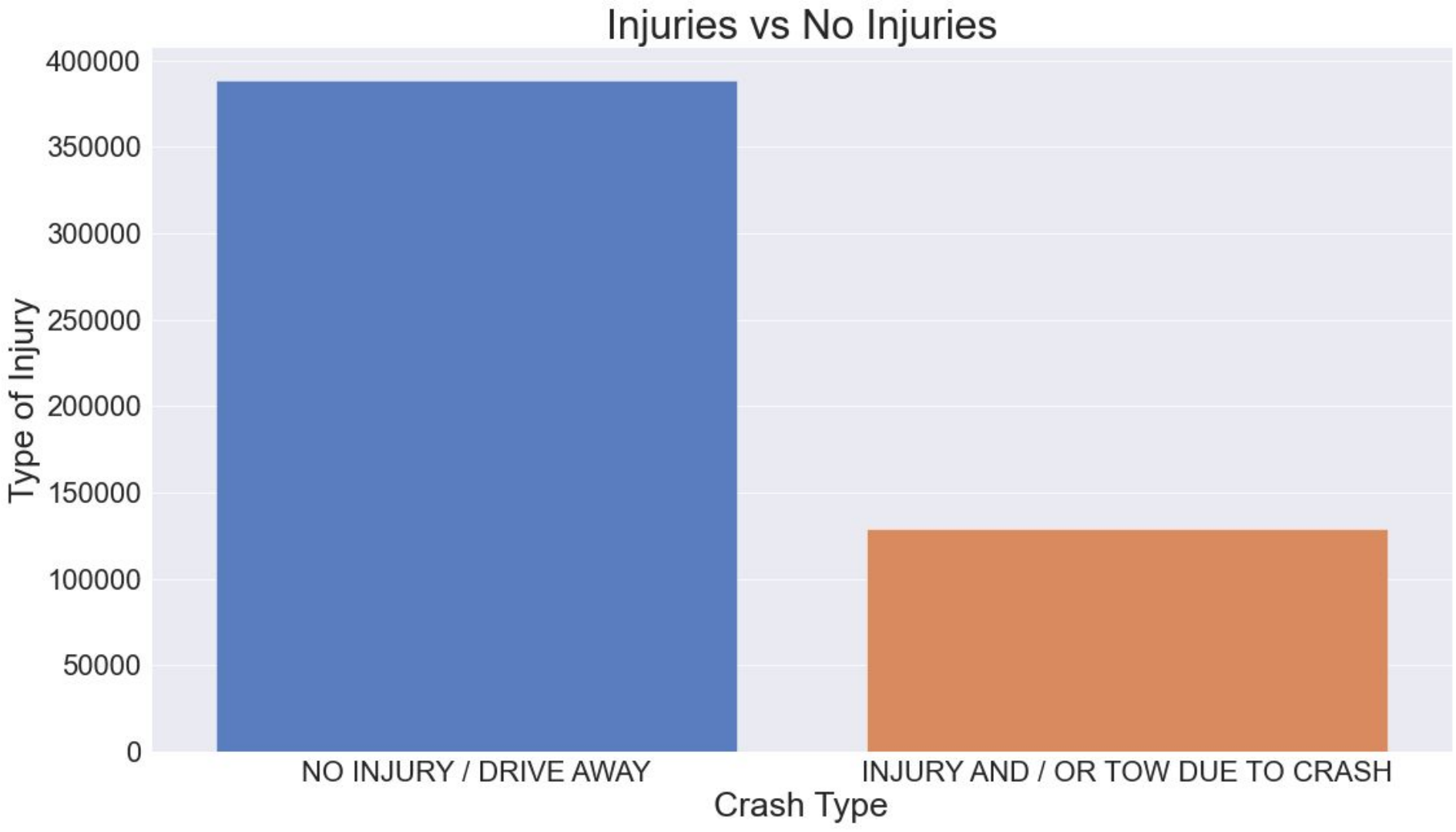
Samuel Rahwa

Can we predict what causes injuries vs no injuries in traffic crashes?

What could the city of Chicago fund/
educate it's residents about when it comes to traffic crashes?

Data Source

- Chicago Data Portal
  - Traffic Crashes shows information about each traffic crash
  - Most of the data was determined by the reporting police officer
- Snapshot of the data
  - 539,468 Rows
  - 49 Columns
- We attempting to predict the CRASH TYPE
  - Crash Type data is Imbalanced
  - NO INJURY / DRIVE AWAY
    - 74.63%
  - INJURY AND / OR TOW DUE TO CRASH
    - 25.37%


Injuries vs No Injuries
400000
350000
300000
250000
200000
150000
100000
50000
0
Type of Injury
NO INJURY / DRIVE AWAY
INJURY AND / OR TOW DUE TO CRASH
Crash Type

Description of Data Columns Used

- **Posted Speed Limit**
  - Posted speed limit
- **TRAFFIC_CONTROL_DEVICE**
  - Traffic control device present at crash location
- **DEVICE_CONDITION**
  - Condition of traffic control device
- **WEATHER_CONDITION**
  - Weather condition at time of crash
- **LIGHTING_CONDITION**
  - Light condition at time of crash
- **FIRST_CRASH_TYPE**
  - Type of first collision in crash
- **TRAFFICWAY_TYPE**
  - Trafficway type
- **ALIGNMENT**
  - Street alignment at crash location
- **ROADWAY_SURFACE_COND**
  - Road surface condition
- **ROAD_DEFECT**
  - Road defects
- **REPORT_TYPE**
  - Administrative report type
- **CRASH_TYPE**
  - A general severity classification for the crash
- **DAMAGE**
  - A field observation of estimated damage
- **PRIM_CONTRIBUTORY_CAUSE**
  - Most significant factor in causing the crash
- **SEC_CONTRIBUTORY_CAUSE**
  - Second most significant factor in causing the crash
- **STREET_NO**
  - Street address number of crash location
- **STREET_DIRECTION**
  - Street address direction (N,E,S,W) of crash location
- **STREET_NAME**
  - Street address name of crash location
- **BEAT_OF_OCCURRENCE**
  - Chicago Police Department Beat ID
- **NUM_UNITS**
  - Each unit represents a mode of traffic with an independent trajectory.
- **CRASH_HOUR**
  - The hour of the day
- **CRASH_DAY_OF_WEEK**
  - The day of the week
- **CRASH_MONTH**
  - The month component of CRASH_DATE.
- **LOCATION**
  - The crash location

EDA

- Used the Pandas Profile for some exploration
- Dropped 13 Columns with 60% or more the missing Data
- Created a Clean Dataset, so it could be Transformed and Standardized for use in all my models
- Created some Plots for Data Visualization and Feature Importance

Models

- **F1 is the metric of measurement for our models**
  - Used if we need to seek a balance between Precision/Recall AND if there is an uneven class distribution
- **XGB Boost**
  - XGBoost makes a decision without looking ahead to see if it is the absolute best choice in long term
  - Initial Model Results:
    - Injury, F1 → 72%
    - No Injury, F1 → 91 %
  - The Improved Model Results:
    - Injury, F1 → 66 %
    - No Injury, F1 → 90 %
- **Random Forest**
  - The model considers only a small subset of features rather than all of the features of the model
  - Initial Model Results:
    - Injury, F1 → 63 %
    - No Injury, F1 → 89%
  - The Improved Model Results:
    - Injury, F1 → 52 %
    - No Injury, F1 → 89 %

Conclusion

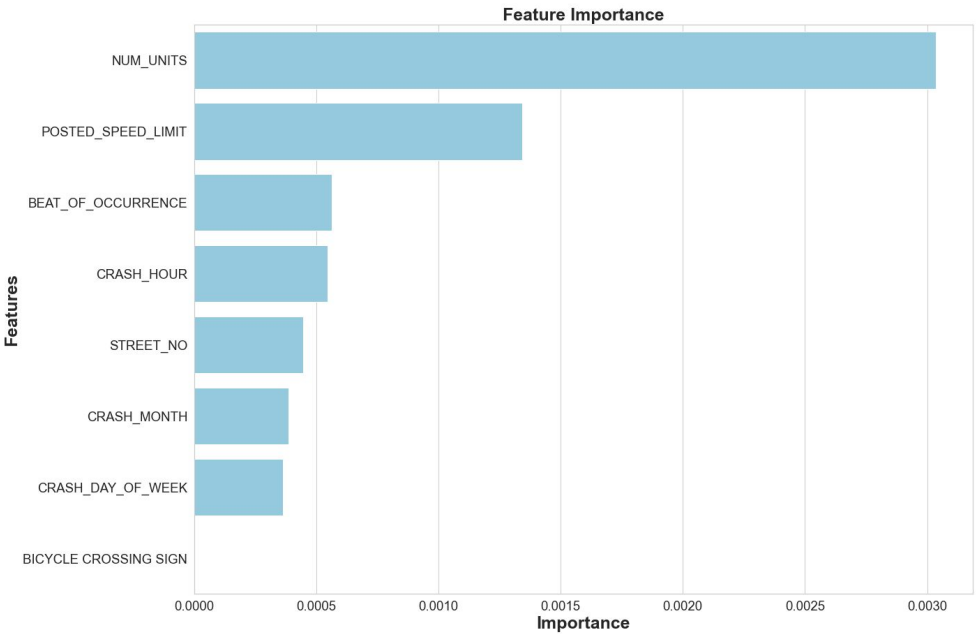
- **Improved XBG Boost Model Predicted these Top 3 Factors in Predicting Injuries vs No Injuries**
  - 1.) NUM_Units
    - Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.
  - 2.) CRASH_HOUR
    - The hour of the day component
  - 3.) POSTED_SPEED_LIMIT
    - Posted speed limit, as determined by reporting officer
- **Improved Random Forest Model Predicted these Top 3 Factors in Predicting Injuries vs No Injuries**
  - 1.) NUM_UNITS
    - Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.
  - 2.) POSTED_SPEED_LIMIT
    - Posted speed limit, as determined by reporting officer
  - 3.) BEAT_OF_OCCURRENCE
    - Chicago Police Department Beat ID. Boundaries available at https://data.cityofchicago.org/d/aerh-rz74

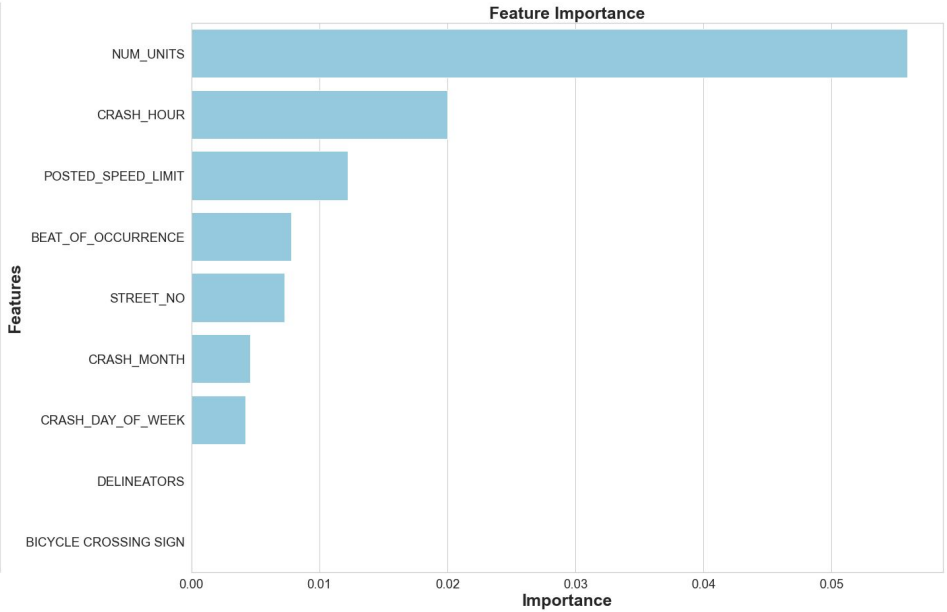
	Features	Gini-Importance
0	NUM_UNITS	0.003034
1	POSTED_SPEED_LIMIT	0.001342
2	BEAT_OF_OCCURRENCE	0.000562
3	CRASH_HOUR	0.000548
4	STREET_NO	0.000446
5	CRASH_MONTH	0.000385
6	CRASH_DAY_OF_WEEK	0.000362
7	BICYCLE CROSSING SIGN	0.000000

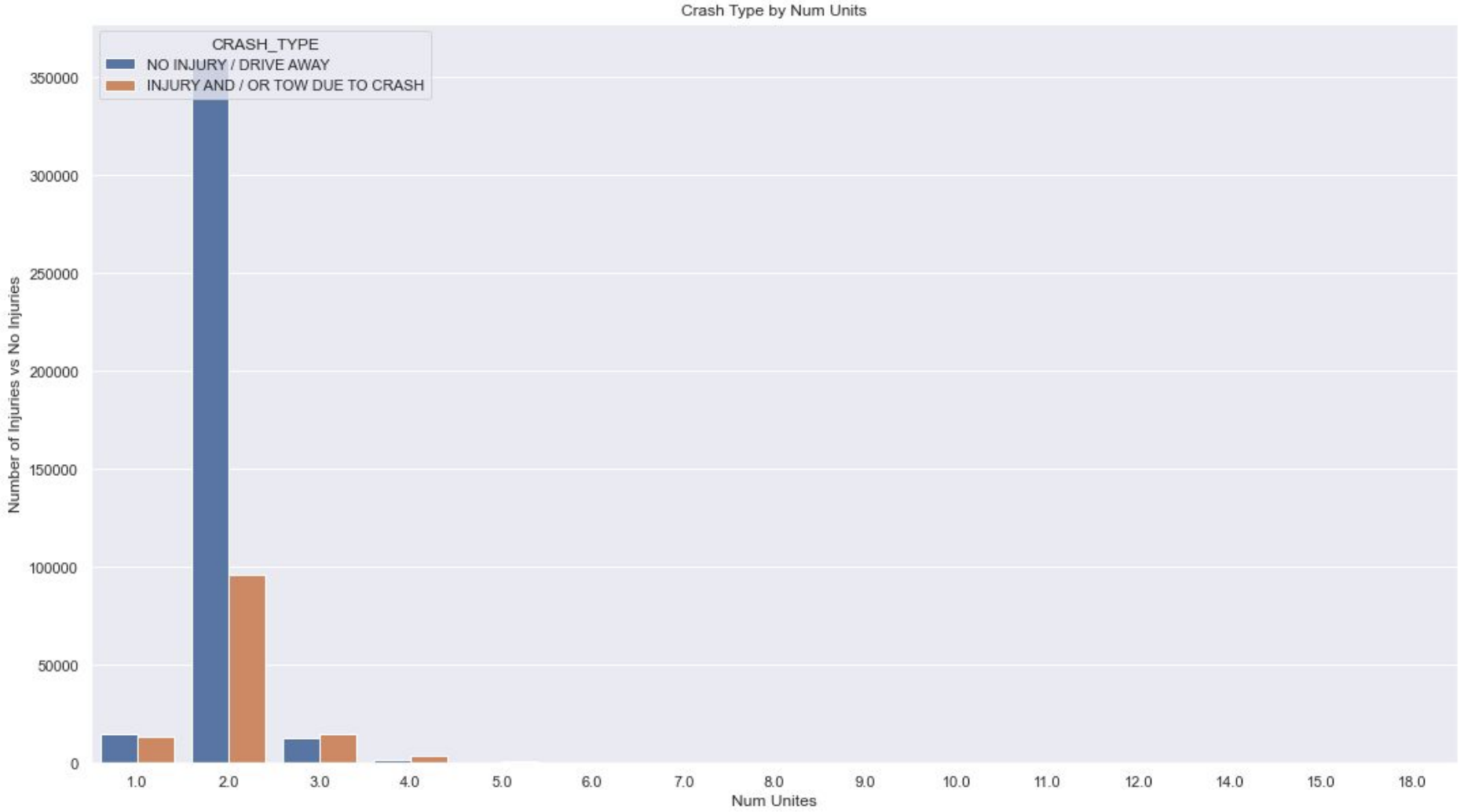
← Hyper Tuned XGB Boost Model

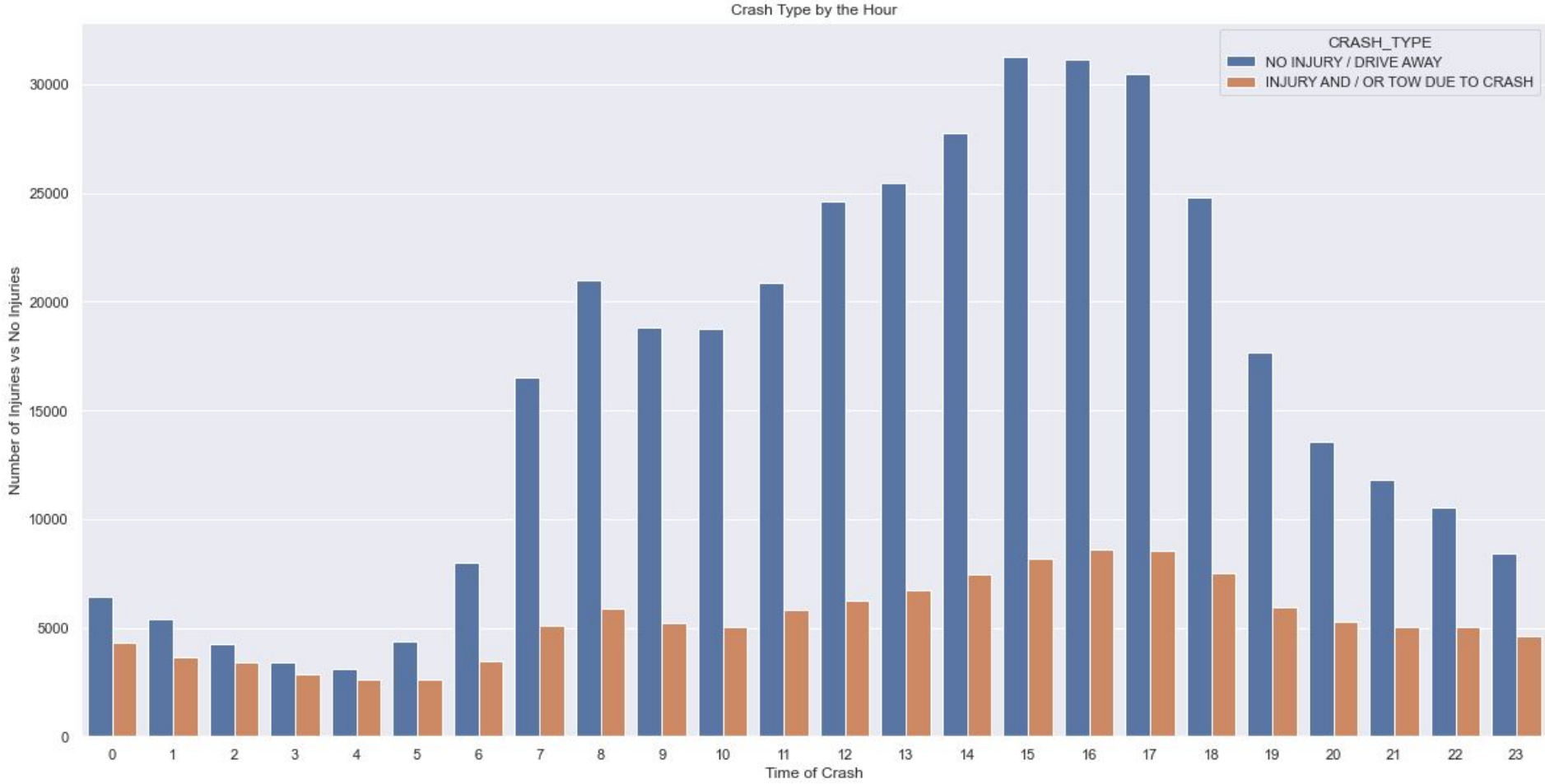
Hyper Tuned Random Forest Model →

	Features	Gini-Importance
0	NUM_UNITS	0.003034
1	POSTED_SPEED_LIMIT	0.001342
2	BEAT_OF_OCCURRENCE	0.000562
3	CRASH_HOUR	0.000548
4	STREET_NO	0.000446
5	CRASH_MONTH	0.000385
6	CRASH_DAY_OF_WEEK	0.000362
7	BICYCLE CROSSING SIGN	0.000000






Crash Type by Num Units
CRASH_TYPE
NO INJURY / DRIVE AWAY
INJURY AND / OR TOW DUE TO CRASH
Number of Injuries vs No Injuries
350000
300000
250000
200000
150000
100000
50000
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
14.0
15.0
18.0
Num Unites


Crash Type by the Hour
CRASH_TYPE
NO INJURY / DRIVE AWAY
INJURY AND / OR TOW DUE TO CRASH
Number of Injuries vs No Injuries
0
5000
10000
15000
20000
25000
30000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Time of Crash

Next Steps

- I could include the Vehicle Data and to Driver/Passenger Data to the Traffic Crashes
- Find a solution why oversampling the minority class (INJURY AND / OR TOW DUE TO CRASH) led to producing zero's across the Test/Train Split
- Increase time spent feature engineering, instead of a reliance on OneHotEncoder to simplify the process
- Prepare more time for modeling due to lack my computational capabilities

Thank you!

Thank you to the City of Chicago for the Dataset

Email: Samuelaaronrahwa@gmail.com

GitHub: @SamuelRahwa

LinkedIn: https://www.linkedin.com/in/samuelrahwa