Analyzing Traffic Collisions in Montgomery County, Maryland

Jingjie Wan (iriswan@umich.edu)

Introduction & Motivation

- Traffic collisions remain a significant public safety issue globally. Understanding the dynamics and contributing factors behind traffic accidents is crucial for developing effective prevention strategies and policies.
- Leverage data mining techniques to analyze crash report data from Montgomery County, Maryland, to uncover underlying patterns and trends that contribute to traffic collisions.

Data

- Crash Reporting - Drivers Data in Montgomery County (2015-2023) by Automated Crash Reporting System (ACRS) of the Maryland State Police (https://catalog.data.gov/dataset/crash-reporting-drivers-data).
- 43 features including injury severity, crash date and time, road name, collision type, weather, driver substance abuse, location, vehicle information, etc.

Questions

- What are the primary contributing factors to the injury severity in traffic collisions in Maryland?
- What are the patterns the injury severity in traffic collisions in time series?
- Are injury severities predictable by related features & on time series?

## Key Features Affecting Injury Severity

1. One-hot encoding categorical features or transform them into numbers
2. Found Top 20 features most correlated with injury severity:
   a) None of the feature has high correlation (highest = 0.16) – the combination of many factors accounts for injury
   b) Top 3 features correlated: **whether the vehicle is a motorcycle** (positive correlated), **the speed limit** (positive correlated), **if the collision is a same-direction sidewipe** (negative correlated).
   c) If the drivers are not at fault, they surprisingly tend to have more severe injuries. Possible reason: unavoidable accident & less reaction time
3. With the top 20 features, build models to classify injury severity. XGBoost Classifier is successful in classifying the injury severity with the highest F1-score of 0.747 – injury severity is predictable by related features


- The **damage extent of a vehicle** has the highest importance in predicting the severity of injury – pay more attention to vehicle quality
- other important features including the **unknown equipment problems** and **substances** used by the driver.

Model (Lag = 3)	Mean Squared Error
Dummy Regression (baseline)	0.0119
ARIMA	0.0116
XGBoost Regression	0.0117

Injury Severity on Time Series
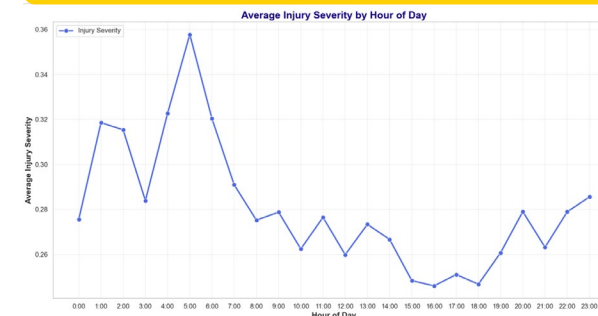
- Seasonal pattern appears on injury severity:
  - Hour of Day: more severe in early morning & Peak at 5 am – lack of attention
  - Day of Week: more on Saturday than other
  - Month: more in May-Sep – more active
- Educate the public & prepare more resources for emergency in these time periods
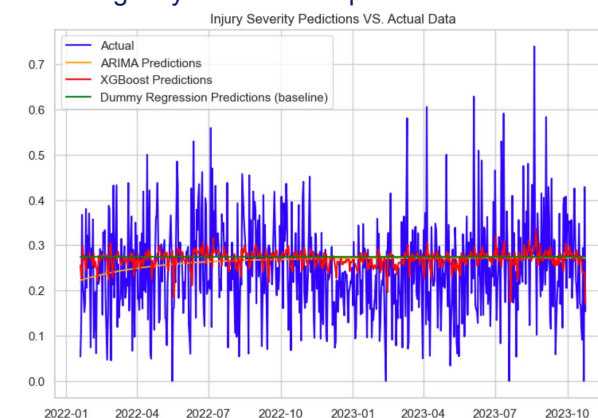
- Only using the time series itself (e.g. ARIMA model) is not sufficient for predicting injury severity in the future
- ARIMA and XGBoost Regression model performs better than the dummy baseline, but the predicted results are conservative and didn't capture the variations in the severity values enough.