Data-Driven Analysis of Major Factors Influencing Traffic Crash Severity

DATS 6501: Data Science Capstone

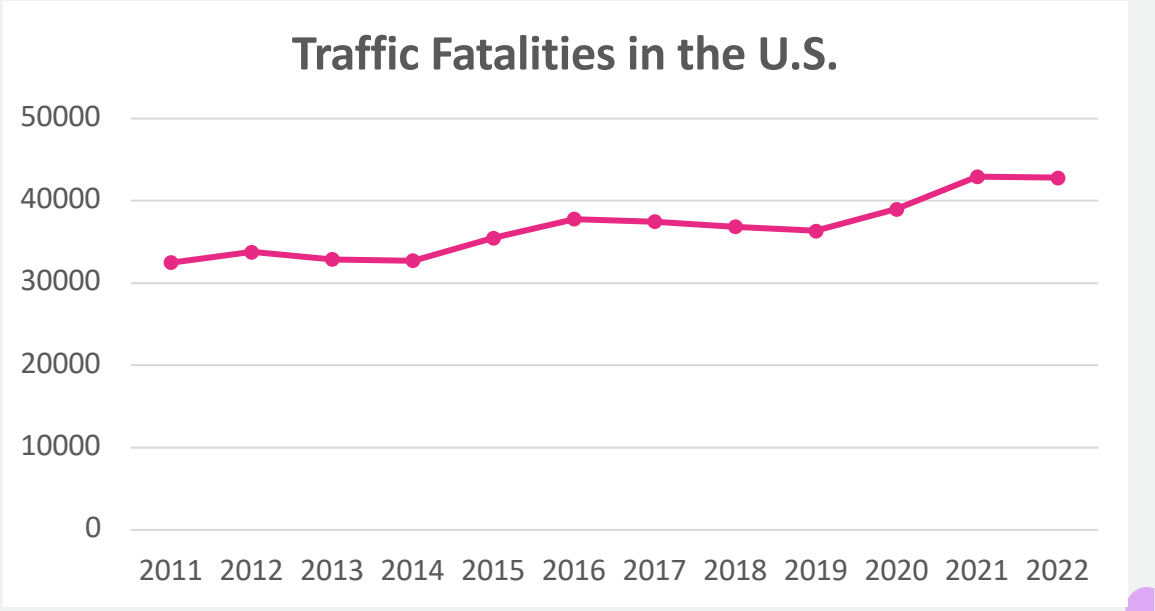
Mid-term

February 28, 2024

Kentaro Osawa, Bharat Premnath

Introduction

- Traffic crashes cause human suffering as well as economic losses
  - In the world, approximately **1.19 million people die** and **20 – 50 million people suffer non-fatal injuries** from road traffic crashes each year. [1]
  - In the U.S., estimated **47,795 people died** in traffic crashes in 2022. [2]
  - In the U.S., traffic fatalities have remained flat or increased. [2]
  - According to NHTSA, traffic crashes cost American society **$340 billion** in 2019. [3]



Introduction

- It is important that governments take **effective countermeasures** to decrease the number of traffic crashes.
- In this project, we will apply our knowledge and skills in data science to analyze **traffic crash data in Chicago**, with the aim of proposing effective countermeasures.

- The scope of this project is as follows:
  1. Finding **major factors** in the severity of traffic crashes
     - Creating and interpreting **ML models**
  2. Creating a **risk map** showing the expected traffic crash counts in each community area in Chicago
     - Creating **Generalized Linear Model (GLM)**

Literature Review (1)

- There are several previous research investigating significant factors in crash severity.
  - Building ML models from several feature subsets and considering the feature subset of the best model as the significant factors [4]
  - Building ML models and evaluating the feature importance of the best model [5, 6]
- **Decision Tree** and **Random Forest** are often used in those research.
- There are common key factors across multiple previous studies (e.g., **day of the week, geographical location**), but different factors have been suggested as the main ones depending on the study.
- Since the target variable, the traffic crash severity, can be highly imbalanced, it is important to handle this **imbalance** [6]

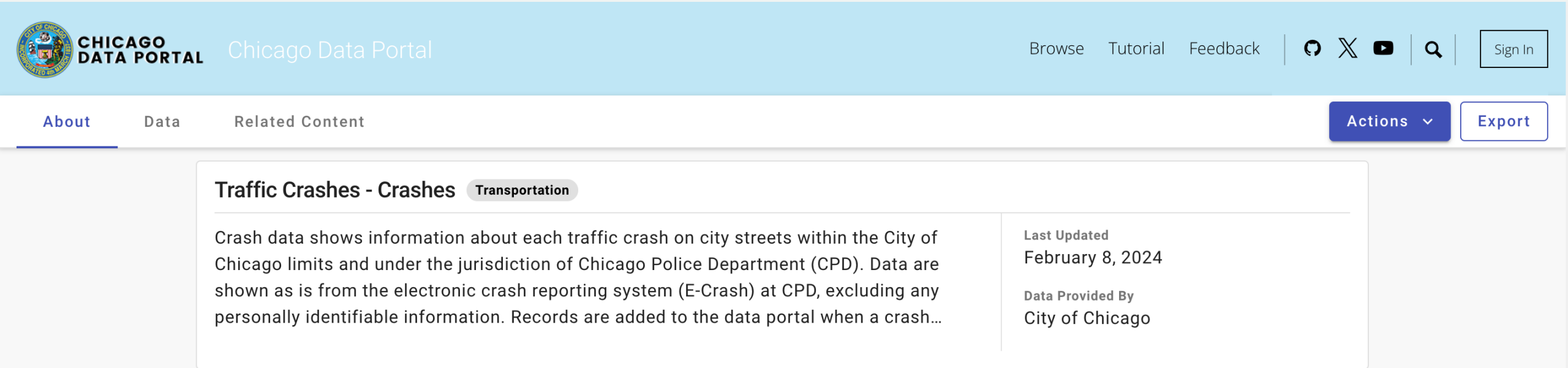
Literature Review (2)

- **Poisson** and **Negative Binomial (NB)** model have commonly used for modeling traffic crash count data. [7]
  - Poisson model: $\mu = var$
  - NB model: $\mu < var$
- Some crash frequency data show more 0 than expected from Poisson or Negative Binomial models. For these data, **zero-inflated models** (Poisson with added zeros) are used. [7]
  - Zero-inflated model: $P(n) = \begin{cases} \alpha + (1-\alpha)e^{-\lambda} ; & n = 0 \\ (1-\alpha)\, e^{-\lambda}\lambda^{n}/n! \; ; & n \geq 1 \end{cases}$
- Almost all crash data show over-dispersion ($\mu < var$), but some data show under-dispersion ($\mu > var$). For these data, **Conway-Maxwell Poisson** model is proposed. [8]

Methodology – Data Collection

- The dataset is acquired from **Chicago Data Portal** [9]
- The dataset is updated daily

Chicago Data Portal Website [9]



Methodology – Data Description

- **799,526** observations
  - Contains traffic crash details **from 2015 to present** (Jan 22, 2024)
  - **48** features
  - **17** numerical features
    - # of fatalities, # of total injuries
    - latitude, longitude
  - **31** categorical features
    - day of the week
    - weather condition, lighting condition

Methodology – Data Preprocessing

- Preparing two preprocessed data: for classification models and for GLMs

Data preprocessing for classification models

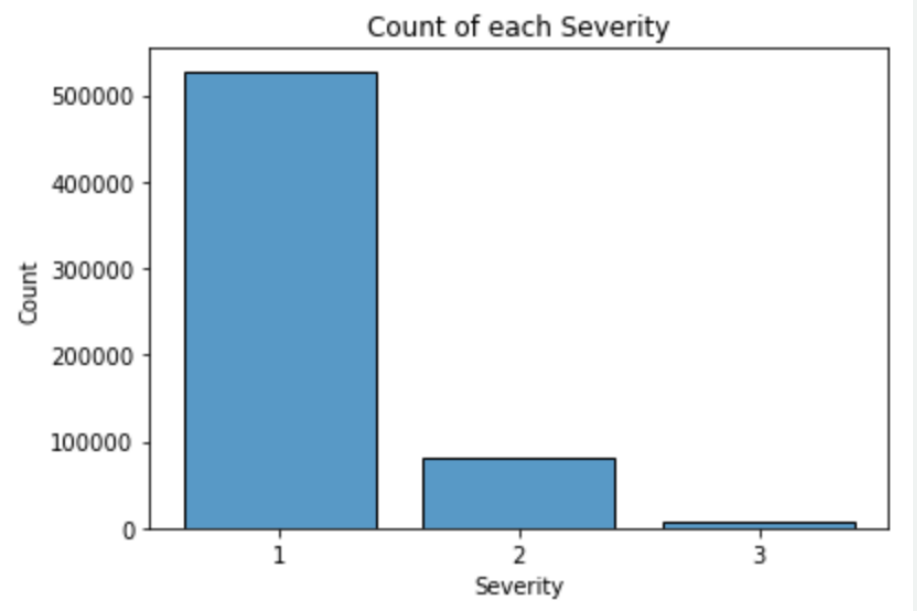
- Removing irrelevant features
- Handling missing value
  1. Dropping features with a proportion of missing values > 50%
  2. Dropping observations with missing values
- Creating the target variable Severity
  - 1: No injuries
  - 2: 1-2 injuries
  - 3: ≥ 3 injuries or ≥ 1 fatality

Data preprocessing for GLMs

- Removing irrelevant features
- Handling missing value
- Creating features
  - Severity, Community area
- Creating count data data-frame
  - Each row is # of crashes for each type of Severity in each community area in 2 hours interval.

Methodology – Modeling (1)

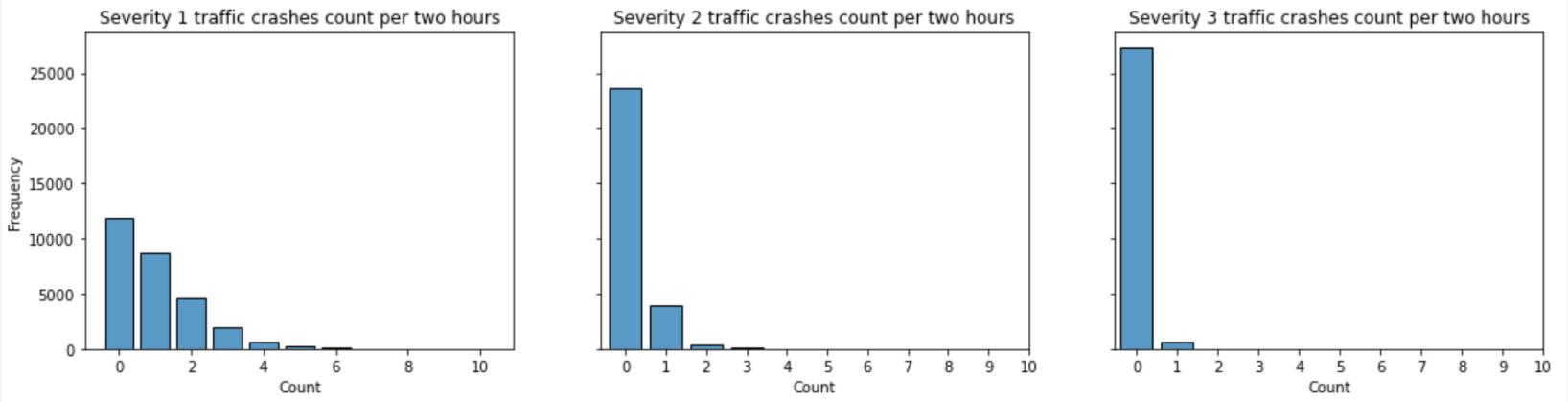
- **Classification models**
  - Building ML models
    - Decision Tree, Random Forest, Boosting, Neural network
    - Handling class imbalance
  - Evaluating models
    - Accuracy, Precision, Recall, F1-score
  - Interpret models
    - Visualizing the contribution of each features using **SHAP (Shapely Additive exPlanations)**



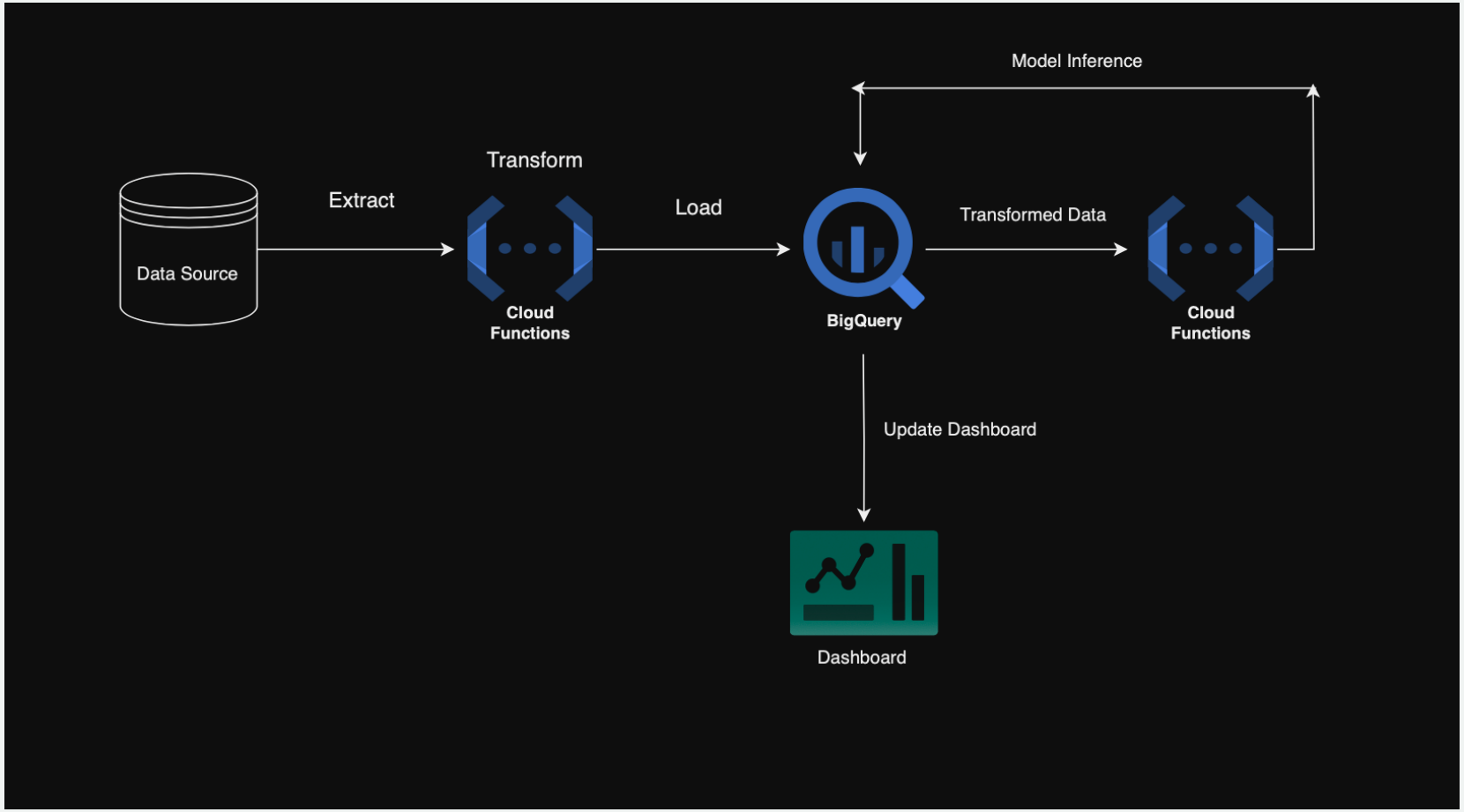
Methodology – Data Modeling (2)

- **GLMs**
  - Building GLMs
    - To find the expected value of count of crashes by using month , day of the week, severity and community area as features.
    - Assuming Poisson model, Negative Binomial model, or Zero-inflated model
  - Evaluating models
    - AIC, BIC

Distribution of traffic accident occurrences for each severity level in Austin



ETL pipeline and Dashboard



Purpose :

- To track changes in the risk map over time
- To detect if there is any change with contributors to the severity of crashes.

Reference

[1] WHO website (https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries)

[2] National Highway Traffic Safety Administration (NHTSA) report (https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813428)

[3] NHTSA website (https://www.nhtsa.gov/press-releases/traffic-crashes-cost-america-billions-2019)

[4] Bhuiyan, H., Ara, J., Hasib, K. M., Sourav, M. I. H., Karim, F. B., Sik-Lanyi, C., ... & Yasmin, S. (2022). Crash severity analysis and risk factors identification based on an alternate data source: a case study of developing country. *Scientific reports*, *12*(1), 21243.

[5] Ghandour, A. J., Hammoud, H., & Al-Hajj, S. (2020). Analyzing factors associated with fatal road crashes: a machine learning approach. International journal of environmental research and public health, 17(11), 4111.

[6] Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7), 61

[7] Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, *37*(1), 35-46.

[8] Lord, D., Guikema, S. D., & Geedipally, S. R. (2008). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*, *40*(3), 1123-1134.

[9] Chicago Data Portal (https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data)

Thank you for listening!