Bike and Pedestrian Traffic Incidents in Chicago

Nathan Rickert, Dharti Seagraves, Raman Hundal, Michael Goodman

Background

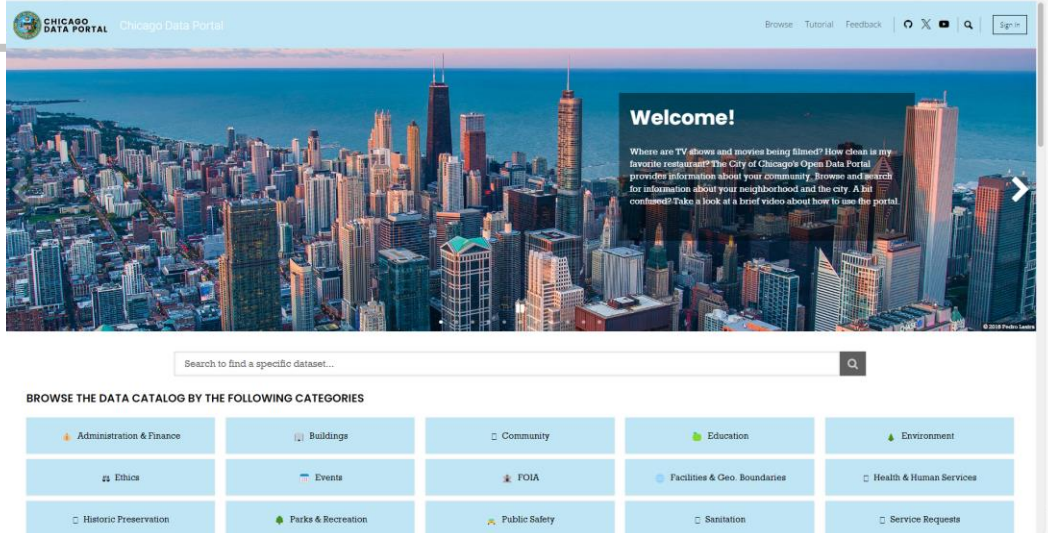
- City of Chicago faces challenges in mitigating traffic accidents and improving road safety
- Our goal is to understand and prevent traffic accidents by identifying patterns and key factors contributing to three categories of crashes
    - No Injury
    - Non-incapacitating Injuries
    - Incapacitating Injury/Fatalities

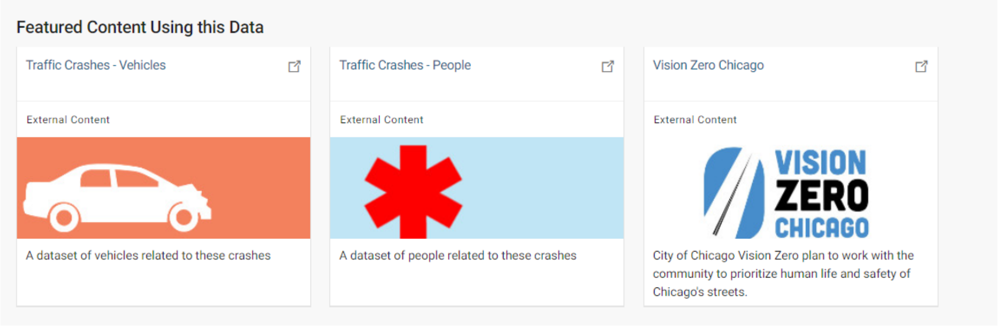
Goals of Analysis

- Classifying accident can help us:
  - Prioritize emergency response allocation
  - Identifying high-risk areas and times to increase/decrease surveillance teams
  - Informing City of Chicago residents of traffic safety issue and plan to alleviate stress

Data Acquisition

- Utilized the Chicago Data Portal SOCRATA API
- Queried all people affected by traffic accidents in the last 5 years that were classified as pedestrians or cyclists.
- Joined with crash event features and the associated vehicle information.
- This resulted in over 22,000 records.
- ~1,500 resulted in fatalities or incapacitating injuries



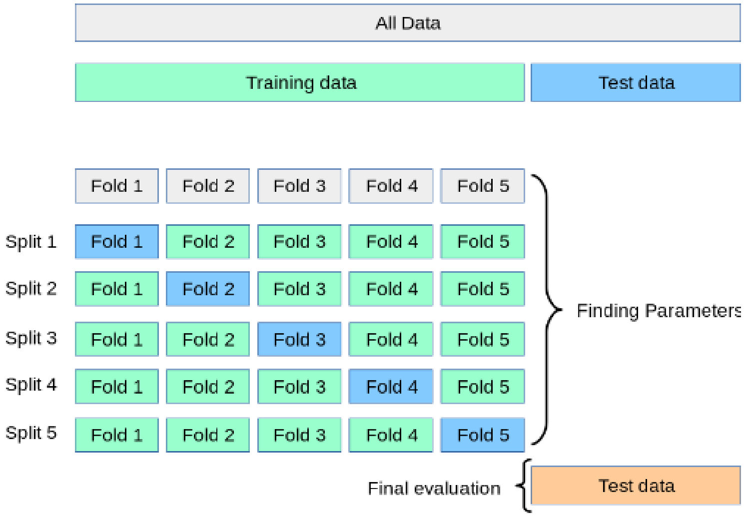


Model Preparation

- Remove unnecessary columns (Nulls to No)
- Engineer target variable
  - 0 – No injury
  - 1 – Non-incapacitating injury
  - 2 – Incapacitating injury/ Fatal
- Encode text categorical features into integers

Methodology

- Feature Selection
  - Principal Component Analysis (PCA
  - K-mean clustering
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)
  - Latent Class Analysis (LCA)
- Model
  - Cross-validation: Repeated k-Folds
- Algorithms
  - Naïve Bayes (baseline)
  - Logistic Regression
  - Support vector machines
  - Random Forest
  - Gradient Boosting
  - Poisson Regression

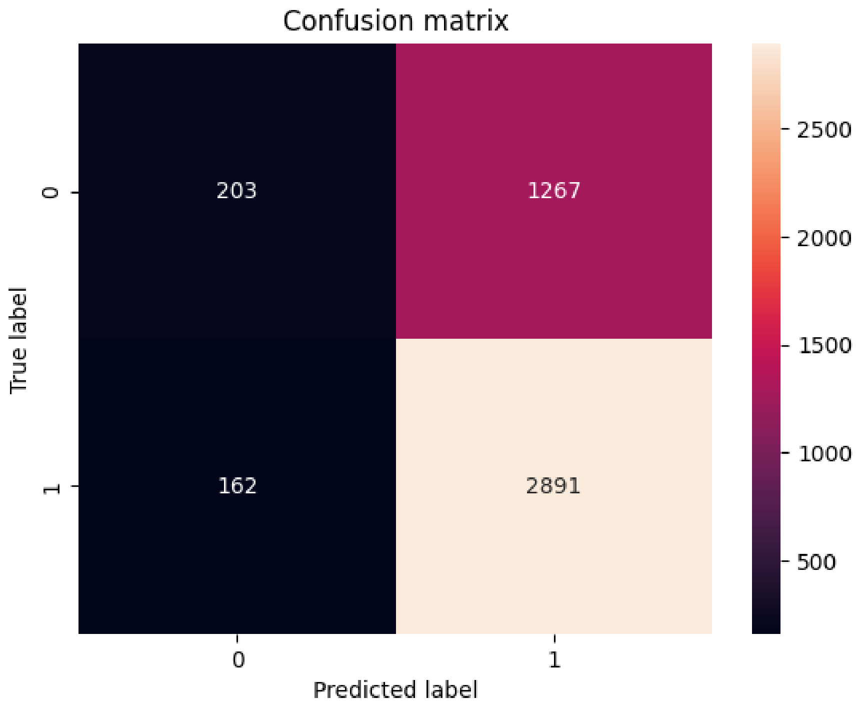


Class	Naïve Bayes	Logistic	SVM	Random Forest	Gradient Boosting
No Injury	0.182162	0.001079	0	0.297126	0.020105
Non Incapacitating	0.775907	0.999246	1	0.645137	0.984083
Fatal/Incapacitating	0.091884	0	0	0.085145	0.005513

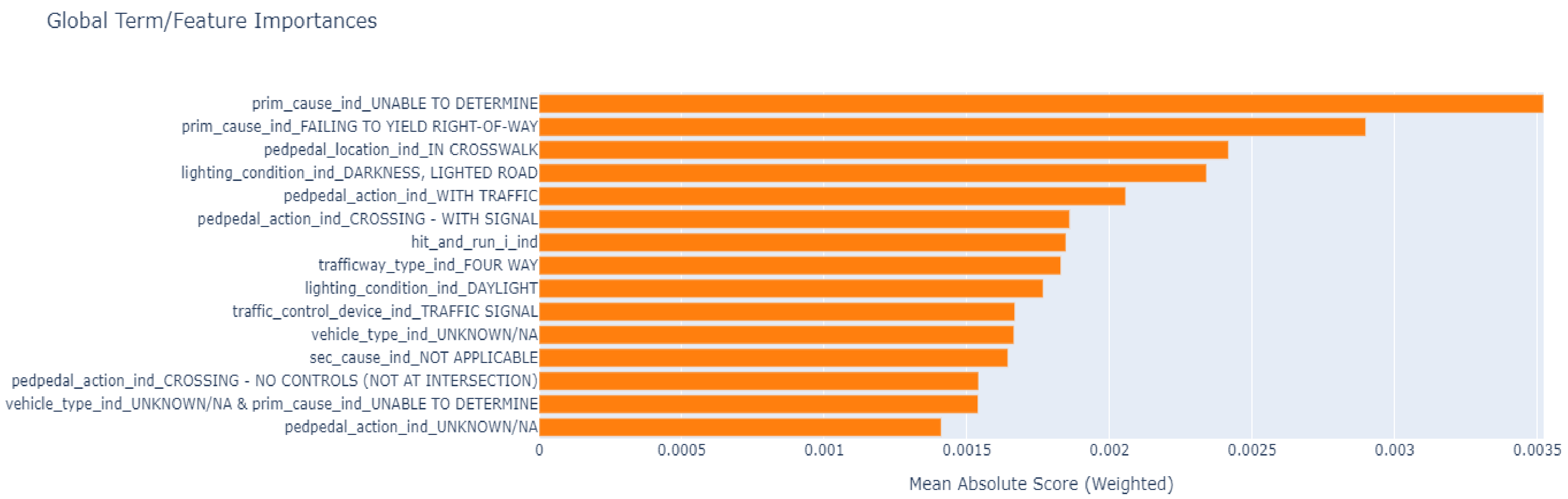
Model Interpretation

- Logistic regression was best for predicting non incapacitating features
- When combining incapacitating and non incapacitating into one variable (1), Random Forest accuracy was still strong and we identified the following key features

```
Accuracy: 68.41%
              precision    recall  f1-score   support

           0       0.56      0.14      0.22      1470
           1       0.70      0.95      0.80      3053

    accuracy                           0.68      4523
   macro avg       0.63      0.54      0.51      4523
weighted avg       0.65      0.68      0.61      4523

[[ 203 1267]
 [ 162 2891]]
```

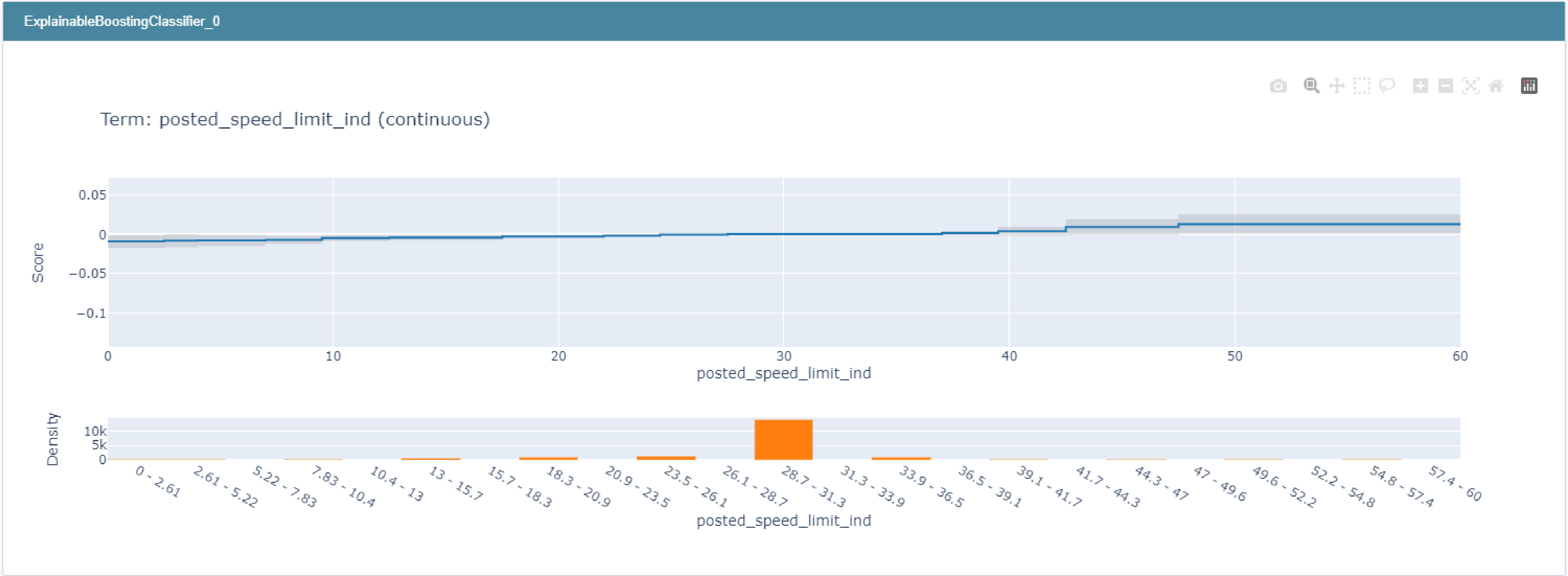


InterpretML- Explainable Boost Classifer



THE UNIVERSITY OF CHICAGO

InterpretML- Explainable Boost Classifer



Improvements

- Data collection – Police Report
- More sensors and studies for road safety
- More data audit and standardization
- More accountability