
# US Crash Accidents

## Capstone Sprint 3

By : John Narvaez

Date: Sept 5, 2023

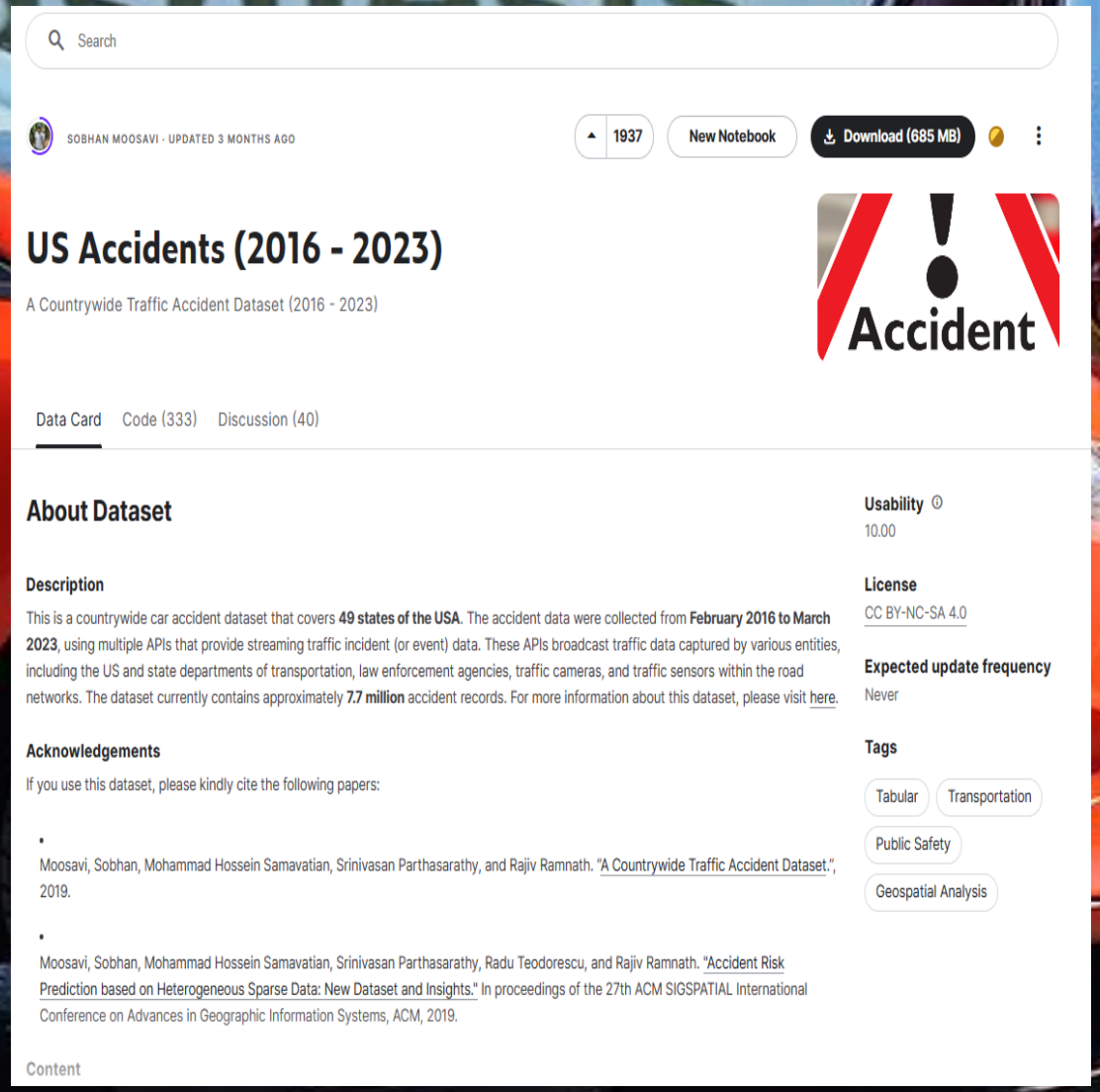
# Prelude


It was estimated of one year that it **killed 36,5000, injured 4.5 million and vehicle damage at 23 million.**

**Cost taxpayers $30 Billion in 2019** roughly 9% of all motor vehicle crash cost.

The losses include **medical, lost of productivity legal and court cost, emergency, insurance, congestion and property damages.**

# Data Source



Search

SOBHAN MOOSAVI · UPDATED 3 MONTHS AGO

1937 | New Notebook | Download (685 MB)

## US Accidents (2016 - 2023)

A Countrywide Traffic Accident Dataset (2016 - 2023)

Accident

Data Card | Code (333) | Discussion (40)

### About Dataset

**Description**

This is a countrywide car accident dataset that covers **49 states of the USA**. The accident data were collected from **February 2016 to March 2023**, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately **7.7 million** accident records. For more information about this dataset, please visit here.

**Acknowledgements**

If you use this dataset, please kindly cite the following papers:

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

Content

**Usability** 10.00

**License** CC BY-NC-SA 4.0

**Expected update frequency** Never

**Tags** Tabular | Transportation | Public Safety | Geospatial Analysis

The dataset was uploaded by **Sobhan Moosavi** who is a scientist at Zoox (as of this writing).
He saw an application of studying hotspot location and extracting studies and rules to predict accidents.

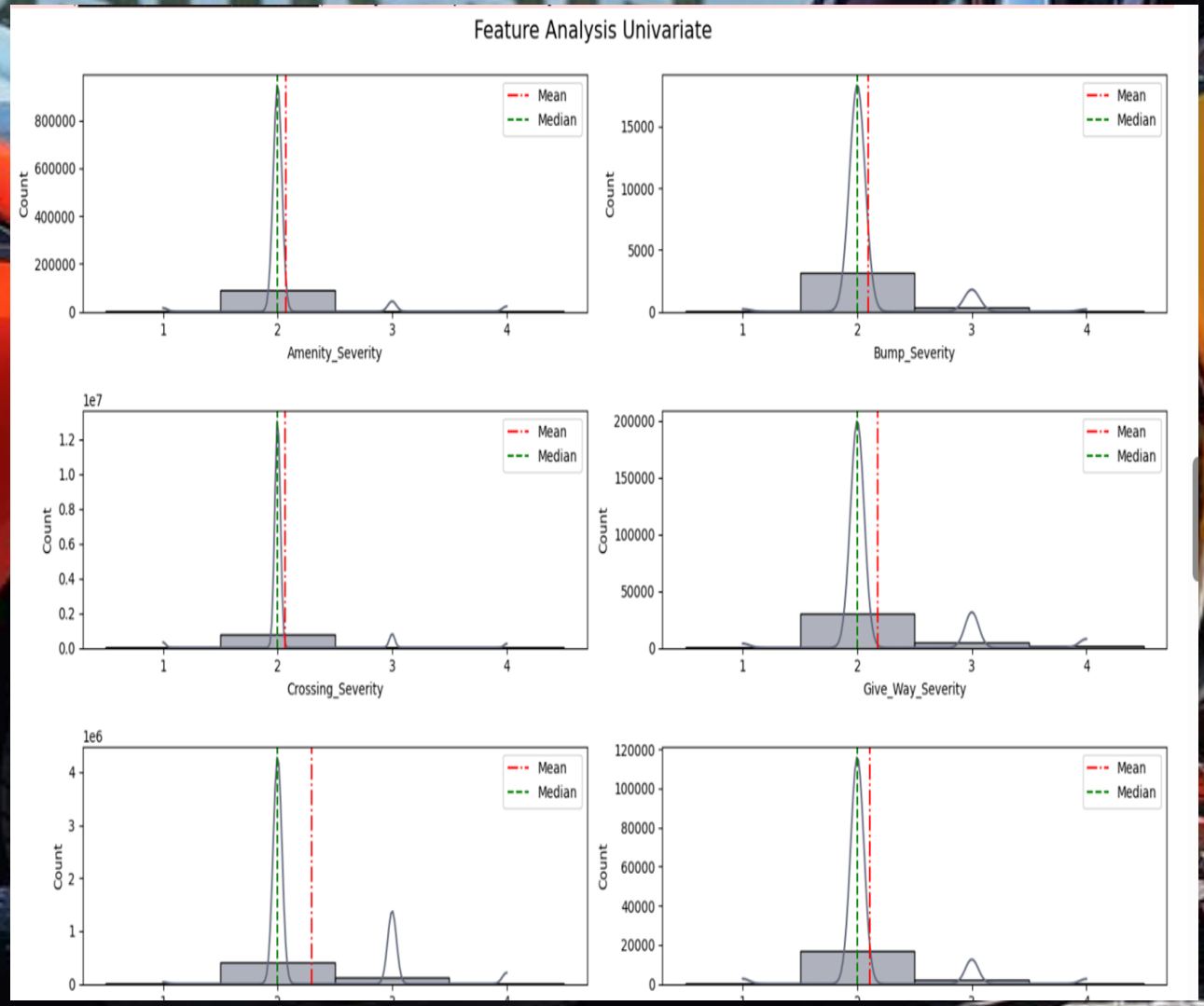
- **Covers 49 states.**
- **Collected from February 2016 to March 2023.**
- **By using Application Programming Interface (API) on telemetry, cameras and different sensors.**
- **Has 7.7 million recorded accidents.**
- **Raw data has 46 columns about weather data, road features and location.**

# Purpose

Can we predict the crash Severity knowing the road features?

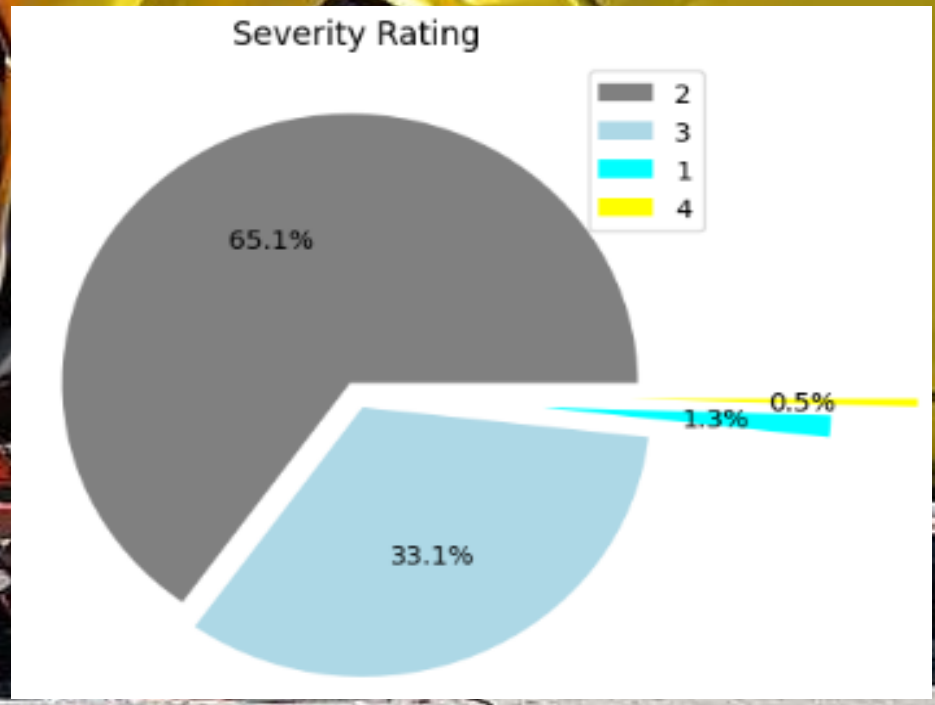
- **The target is the Severity rating (1, 2, 3, 4) on traffic condition. One (1) being the least impact.**
- **Independent variables are the road features.**


# Road Feature Data




The target is a **multiclass and discrete.**

The majority class is rating 2 and 3. And there is a huge class imbalance.



# Modelling

Explored five (5) machine learning model which are good in handling a multiclass classification problem.

- **Logistic Regression**
  - **Simple and can be quickly trained.**
  - **Less prone to overfitting.**

- **Decision Tree**
  - **Ability to handle multi-output**
  - **Not affected by feature scaling.**
  - **Requires little data preparation.**

# Modelling

- Naïve Bayes (BeronoulliNB)
  - Suitable for discrete and large data.
  - Designed for binary / Boolean features.
- XGBoost
  - Parallel tree boosting.
  - Gradient boosting .
- Tensor Flow
  - Can handle large complex data.
  - Generalization capability

# Modelling Results

- **Logistic Regression**
  - Train : 65.81 %
  - Test score : 65.77 %
- **Decision Tree**
  - Train : 65.90 %
  - Test score : 65.86 %
- **Bernoulli NB**
  - Train : 65.78 %
  - Test score : 65.73 %
- **XGBoost**
  - Train : 65.90 %
  - Test score : 65.87 %
- **Tensor Flow**
  - Train : 65.89 %
  - Test score : 65.87 %

# Hyperparameter

- Adjusting and controlling the model structure, function and performance. It allow data scientist to tweak a mode.

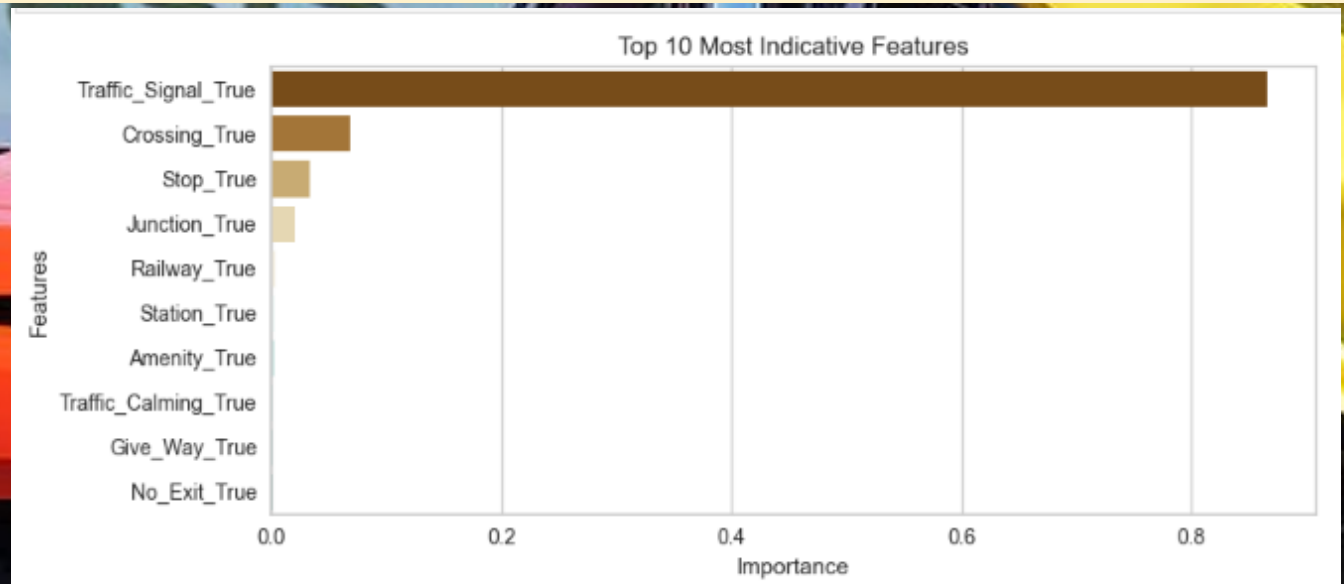
- Best estimator (Decision Tree)
  - max_depth : 10
  - min_samples_leaf : 2

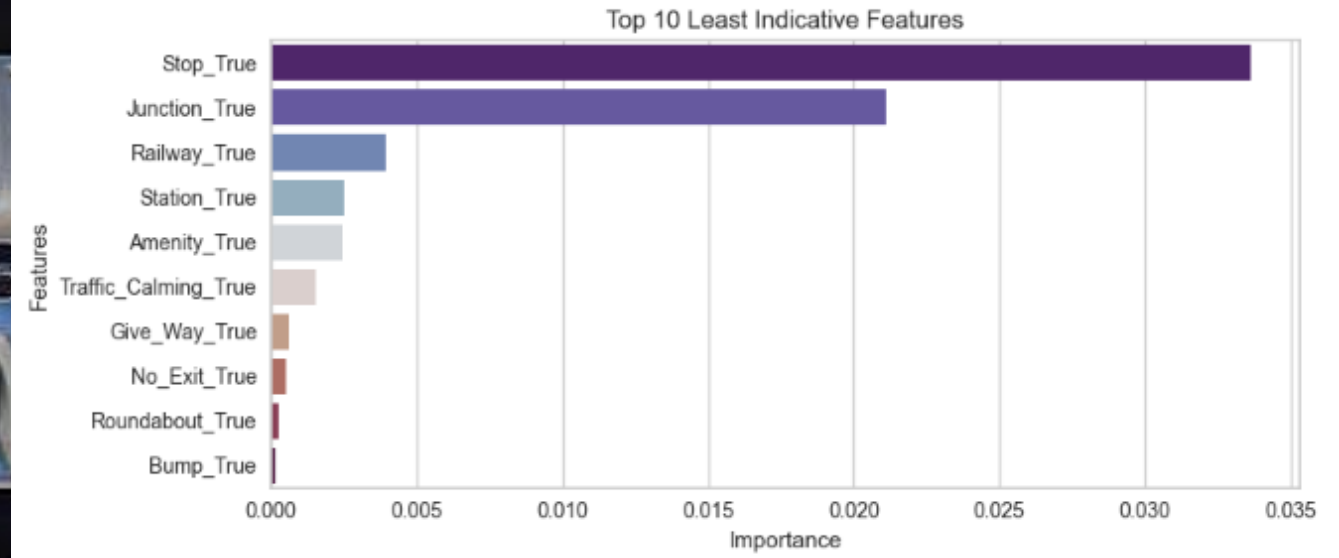
- Decision Tree
  - Train : 65.91 %
  - Test score : 65.87 %

(Base line)

- XGBoost
  - Train : 65.90 %
  - Test score : 65.87 %

# Feature importance



# Remarks

Here we tested four (4) supervised and one (1) deeplearning machine models.

The results for me is quite surprising that neither showed to be the best model.
All train-test score are closed between models.

During extensive research of each different models there is no set rules or guidelines for choosing the perfect parameters.

Result or prediction are mostly centered to Severity 2 and 3. Due to the fact they are the majority class and the data is highly imbalance. I was thinking of lumping together 1 and 2 as "moderate severity" and 3 and 4 as "severe".

In my opinion the precision would be good enough. The purpose of this is to predict which road features would cause a high severity. Because we can study and improve the road features. Unlike weather, temperature and alike; that is sometimes beyond human control.

We can also concentrate resources to highly severe accident road conditions. We could improve this prediction by adding more data like which state, county and city. As these would be different due to several factors.


# Thank You