# PREDICTING SEVERITY OF CRASHES

## In Chicago

**By Tzega Abera**

# WHY?

- Chicago: the largest urban center in Illinois, the third-largest in the United States.
- Population: 2.7 million and covers an area of 237 square miles
- Average daily traffic volume in 2019 = 20,000 vehicles/ day
- Reported motor vehicle crashes in 2019 = 117,949
- Purpose:
  - Building a crash severity predicting model  to determine if a crash will result in a severe/fatal incident

# DATA

- Source:  electronic crash reporting system (E-Crash) at CPD
- Includes traffic crashes on city streets within the City of Chicago limits and under the jurisdiction of Chicago Police Department (CPD)
- Records are added when a crash report is finalized or when amendments are made to an existing report in E-Crash
- Includes only crashes with a property damage value of $1,500 or more or involving bodily injury to any person(s)
- Source: https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

# METHOD

- Sampling technique to modify the format of the data structure for analysis
- Selected a subset of the data from the larger dataset was for analysis
- The raw dataset: 600,000 rows and 49 features.
- reduced dataset size to 200,000 rows and 38 features
  - Save computation time
  - make the data more manageable.

# DATA CLEANING

**Challenges:**

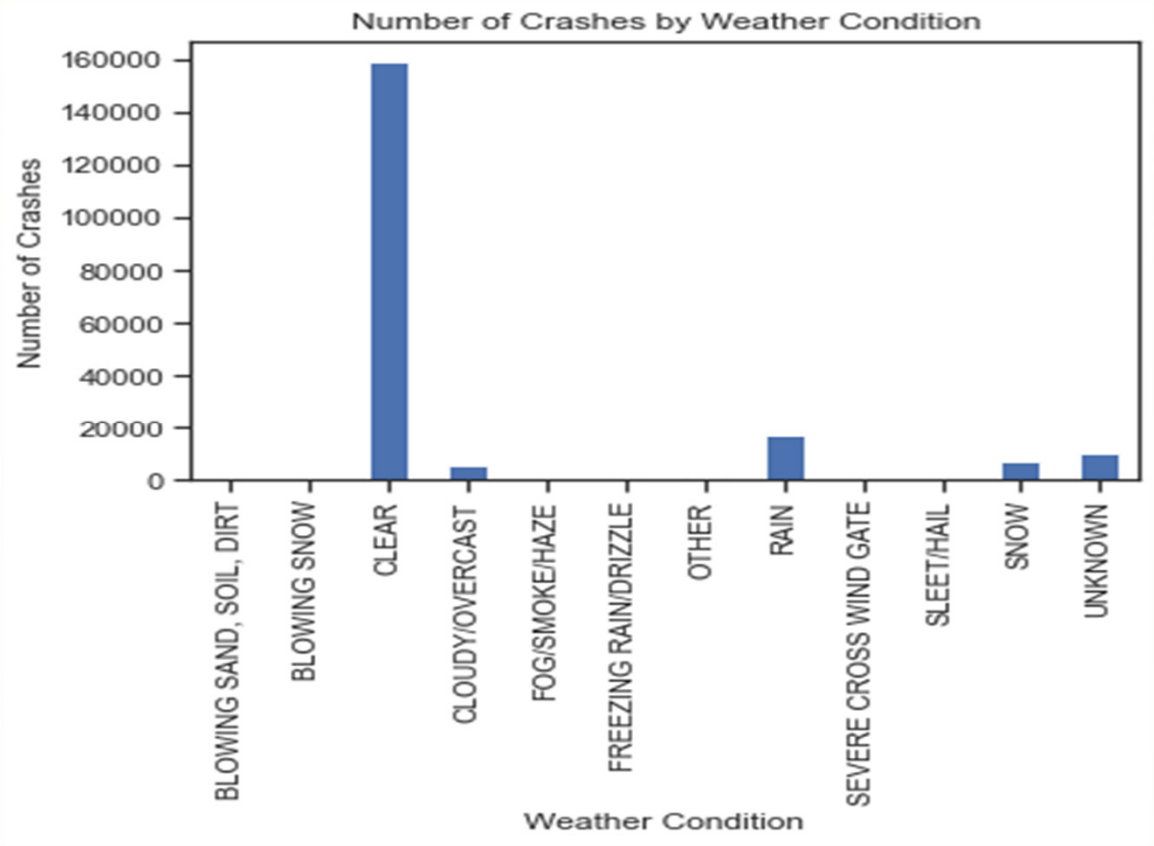
- Dataset was too large to work with
  - Reduced the size of the dataset by extracting data from 2013-2023
- Multiple columns with over 65% of missing values
  - Solution: dropped the columns
- The target variable was not available as binary datapoint
  - identified relevant criteria and used that to create a binary datapoint for the target variable

# EDA

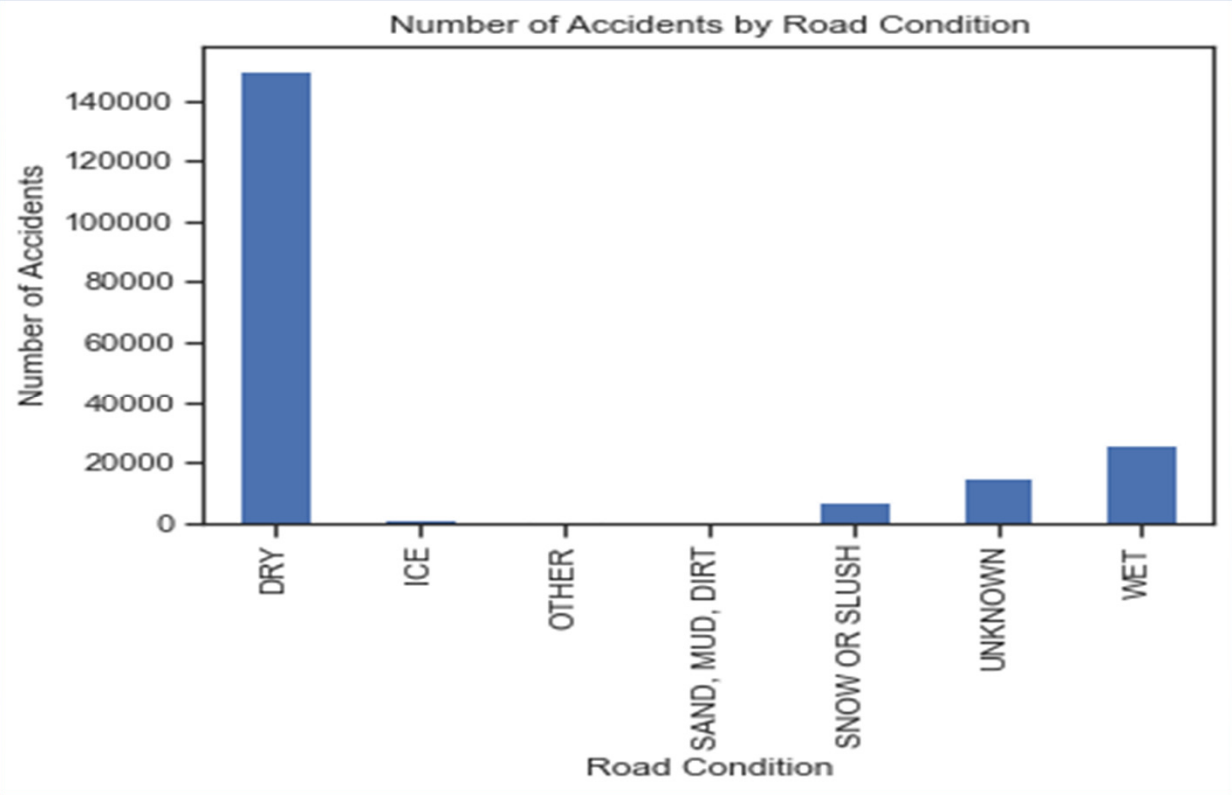
**Identified:**

- most common causes of traffic crashes in Chicago.
- relationship between weather conditions and the crashes
- relationship between road conditions and the number of accidents
- relationship between the time of day and the severity of crashes
- specific areas of the city that are more prone to crashes
- Correlation between speed and accidents

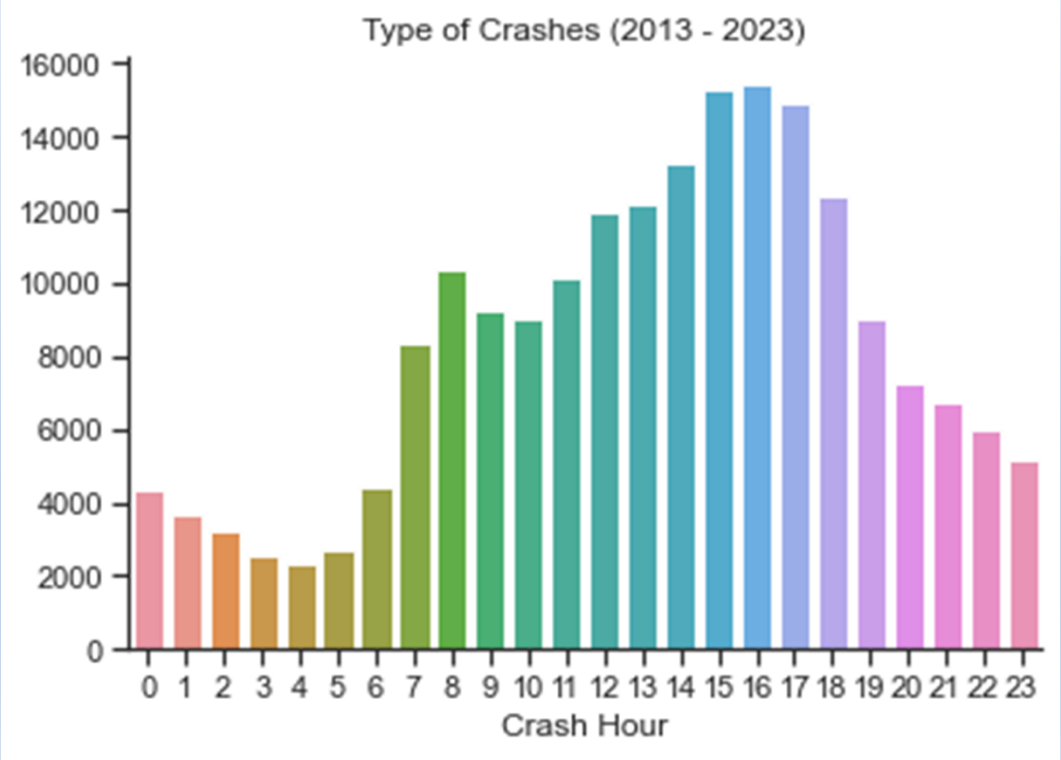
# EDA - Weather Condition and Number of Crashes



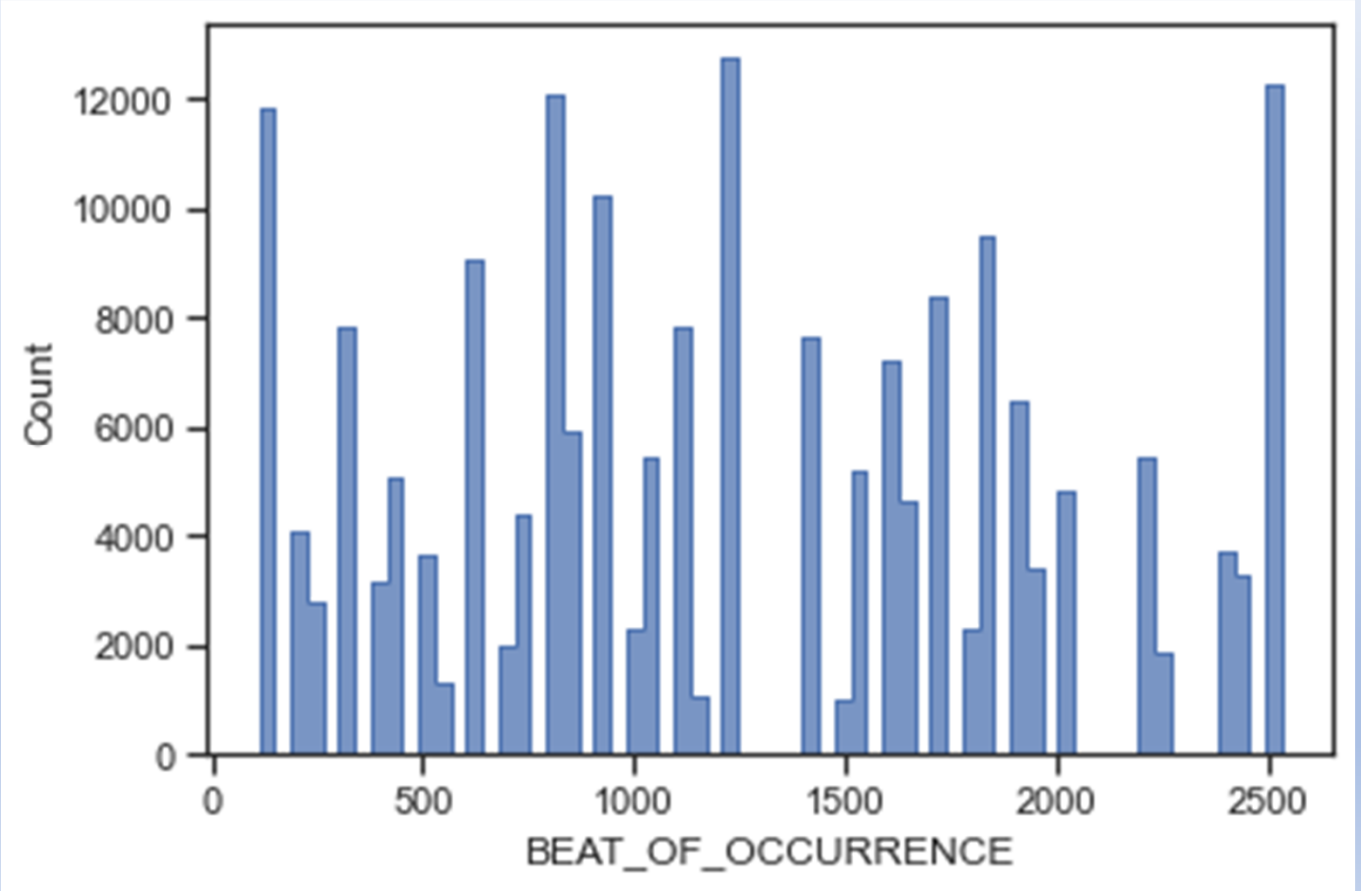
# EDA - ROAD CONDITION AND NUMBER OF ACCIDENTS



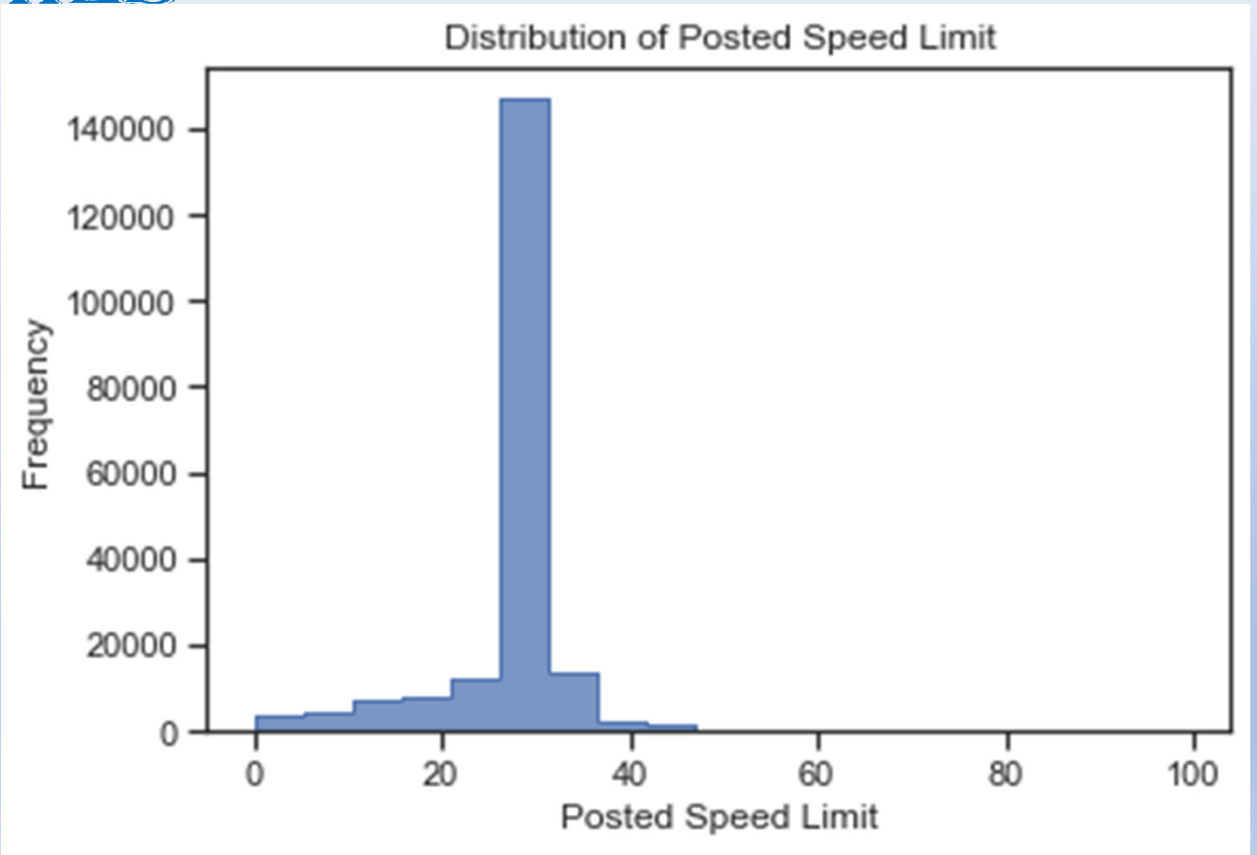
# EDA - CRASH HOUR AND NUMBER OF CRASHES



# EDA-BEATS AND NUMBER OF CRASHES

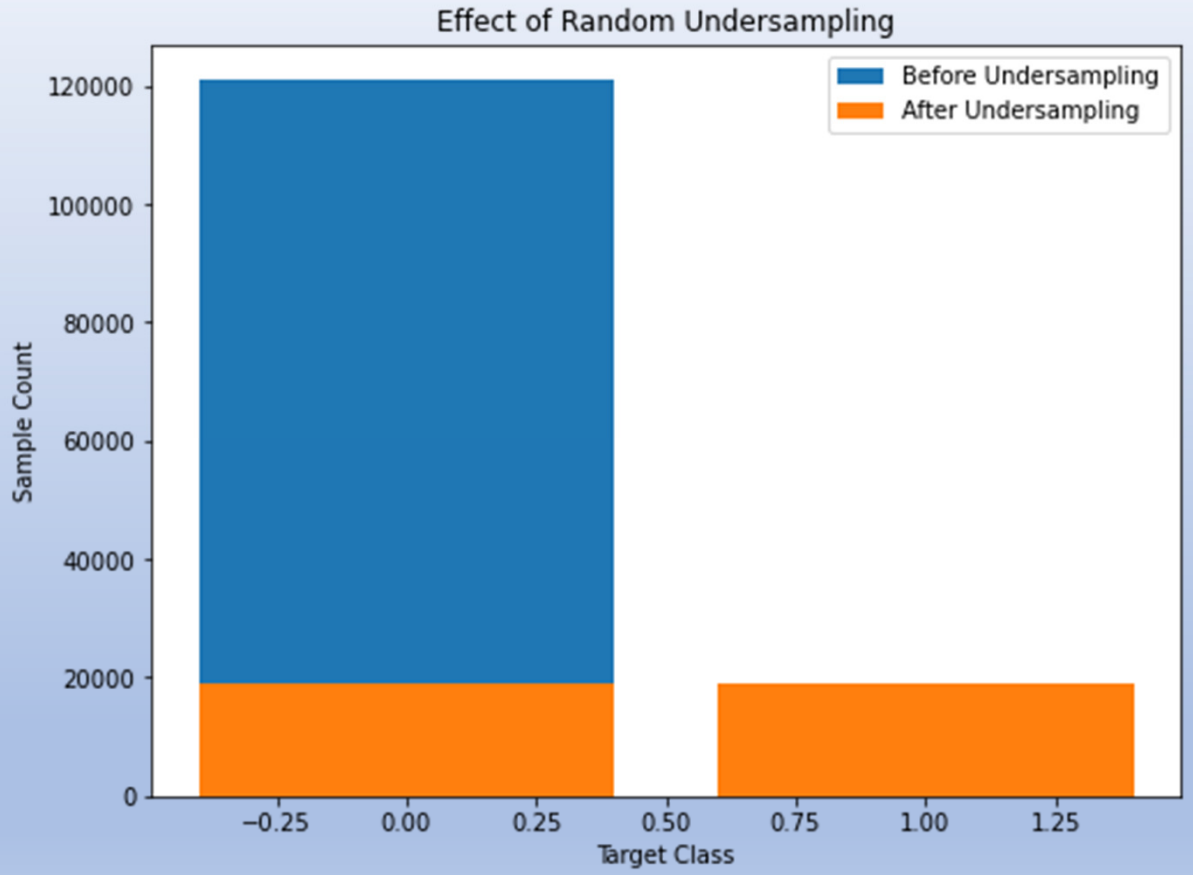


# EDA - SPEED LIMIT AND NUMBER OF CRASHES



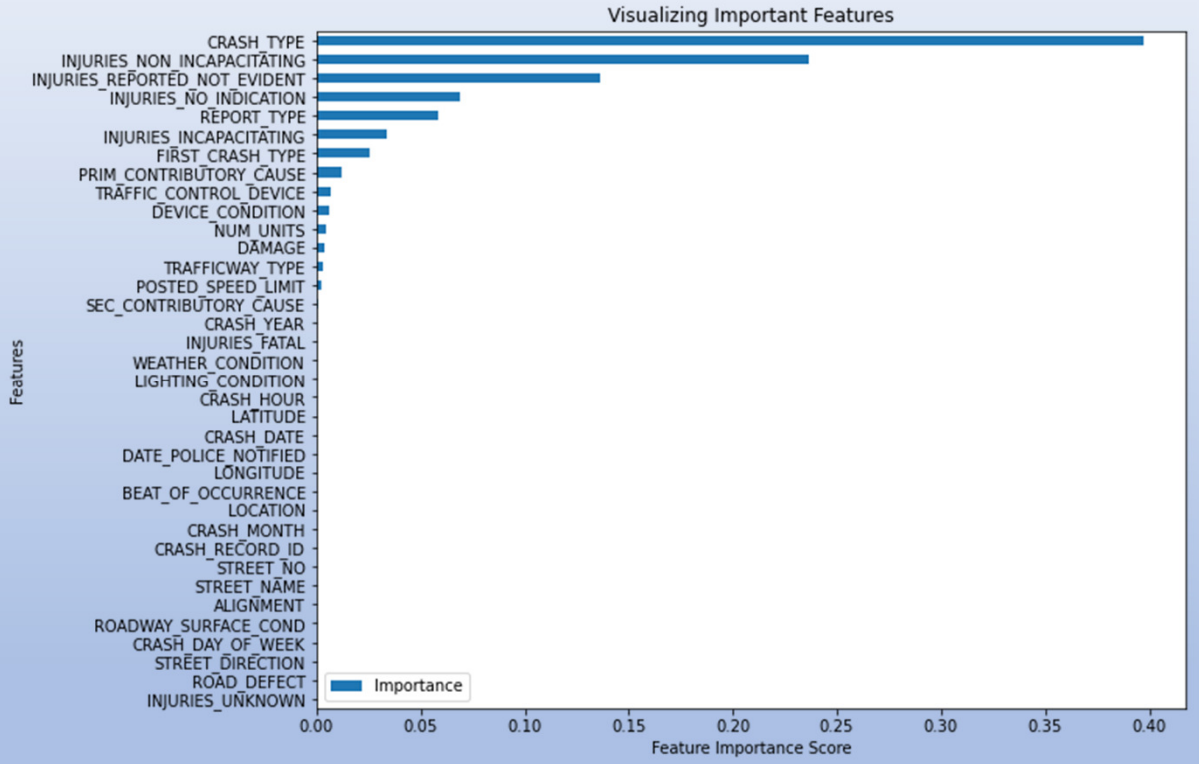
# MODEL SELECTION

- Under-sampling technique to balance the classes in our dataset.



# FEATURE SELECTION

feature selection analysis: dropped the features "INJURIES_TOTAL", and "MOST_SEVERE_INJURY" to prevent overfitting and bias



# COMPARE ML MODELS

- The logistic regression model achieved the highest accuracy score

| Algorithm | Model accuracy score |
|---|---|
| KNN | 0.999633 |
| Random Forest | 0.949183 |
| Logistic Regression | 1.000000 |
| Gradient Boost | 0.995367 |
| Naive Bayes | 0.999283 |

# COMPARE ML MODELS

all algorithms have good prediction ability, but the logistic regression model achieved the highest scores

| | Algorithm | ROC-AUC train score | ROC-AUC test score |
|---|---|---|---|
| 0 | KNN | 0.987377 | 0.964990 |
| 1 | Random Forest | 0.999945 | 1.000000 |
| 2 | Logistic Regression | 1.000000 | 1.000000 |
| 3 | Gradient Boost | 0.999873 | 0.999938 |
| 4 | Naive Bayes | 1.000000 | 0.999971 |

# Limitations and Future Improvements

- analysis only considers the incidence of crashes and does not account for factors such as driver behavior or road conditions
- subsampled dataset may not be representative of the larger dataset
- Used only underdamping technique to balance the data
  - Oversampling should also be used for comparison