# Predicting Accident Outcomes in New York City

THAYER SCHOOL OFDF
ENGINEERINGG
AT DARTMOUTHH

# Abstract

## Purpose:

- Build a model that can accurately predict the outcome of a traffic accident and assists in the deployment of officers.

## Methods:

- Clean and Prepare Data for classification algorithms.
- Tested four different classification algorithms.
- Compare based on Recall of each algorithm and tuned parameters for best performance.

## Major Findings:

- Vast majority of accidents do not result in injury or death.
- There is a correlation between number of vehicles involved and injury or death.

## Conclusion:

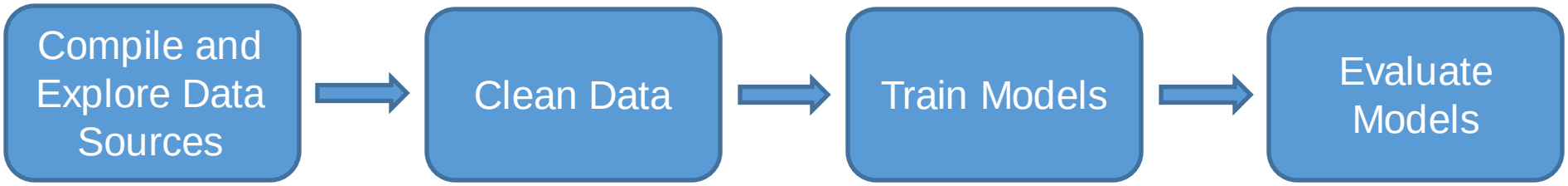
- The best performing algorithm, Naïve-Bayes Gaussian, still only managed a recall of 65%. This is not an accurate enough model to use with human lives.

# Introduction

Overview:

- The goal of this project is to accurately predict whether a traffic accident ends in injury or death using limited, observational data.

# Introduction - Workflow




- Used NYC Traffic incident data and NOAA Historical Weather Data.
- Found Traffic data to be quite messy

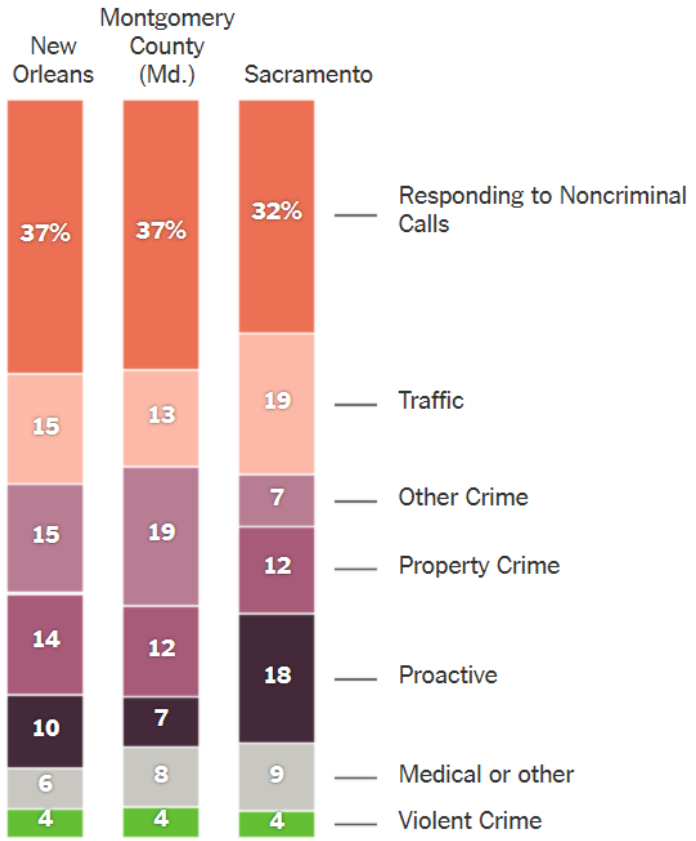
- Deleted columns with extraneous info
- Consolidated Vehicle types. i.e Pickup tRuck to pickup truck

- Trained four different models KNN, GaussianNB, Random Forest, Decision Tree

- Using grid search and cross validation to tune and evaluate models.

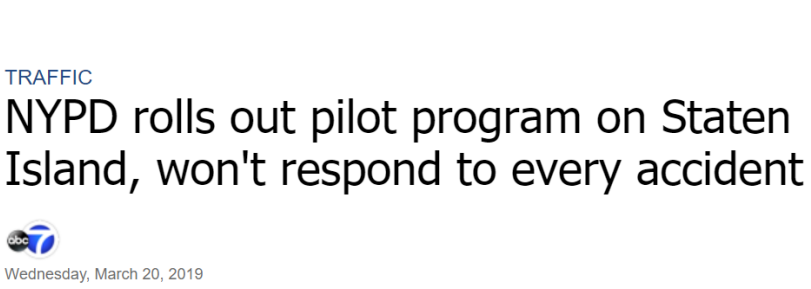
# The Problem

- Police spend a large portion of their time working traffic incidents.
- While budgets are tight, creative solutions are needed to help create more effective distribution of officers.
- Prioritizing accident response can allow departments to cut back on staff or utilize staff more effectively.




Source: https://www.nytimes.com/2020/06/19/upshot/unrest-police-time-violent-crime.html

# Purpose of the Study



TRAFFIC

NYPD rolls out pilot program on Staten Island, won't respond to every accident

Wednesday, March 20, 2019

METRO

**NYPD will stop responding to 'non-injury' car crashes amid coronavirus crisis**

By Olivia Bensimon and David Meyer

April 3, 2020 | 1:53pm | Updated

- Develop a model that can accurately predict if an accident has resulted in harm.
- Use only observational data that would be available to a camera.
- Help 911 dispatchers and police departments effectively distribute resources and time

# Hypothesis

Machine Learning can be used to predict traffic accident outcomes with an accuracy and recall above 90%. Using publicly available data.

# Project Description

## Data Cleaning

- Vehicle data contained thousands of misspellings, incomprehensible categories.
- Extended beyond the date range of the weather data.
- Descriptors of vehicles must be encoded so machine learning algorithms can understand

Sport Utility Vehicle → Value: 2

# Project Description

## Data Cleaning

- Weather data two columns with over 60% of values missing.
- Converted Vehicle Code values to feature columns using One-Hot Encoding
  - No strong correlation found between vehicle type and injury or death.

| Vehicle Type and Location | Correlation w/ Harm |
|---|---|
| Bicycle | 0.47 |
| Single Vehicle accident | 0.201 |
| Motorcycle | 0.102 |
| Sedan | 0.086 |

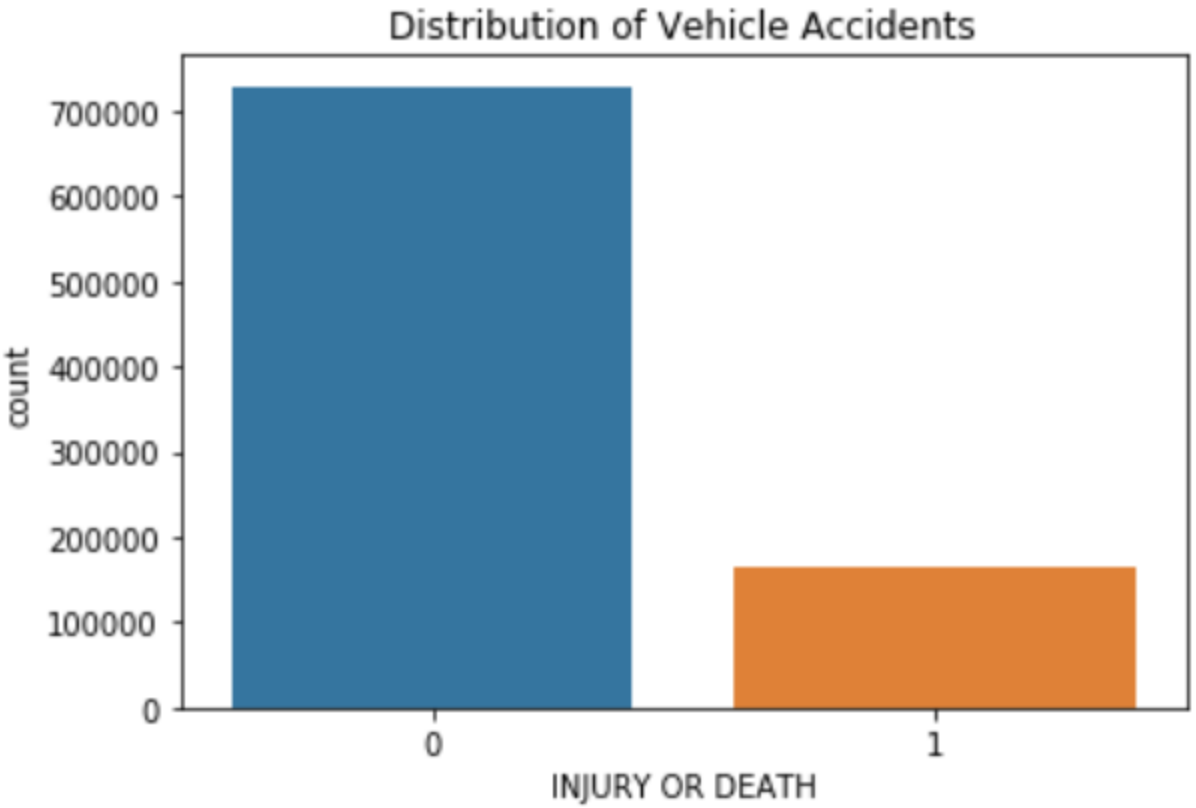
# Project Description

Train Models

- Created target variable for Injury or Death in an accident.
- Data has a heavy imbalance towards no-harm incidents.
- Used Random undersampling to balance the dataset. Data shrunk from 896,000 to 332,000.

# Project Description

Train Models



# Project Description

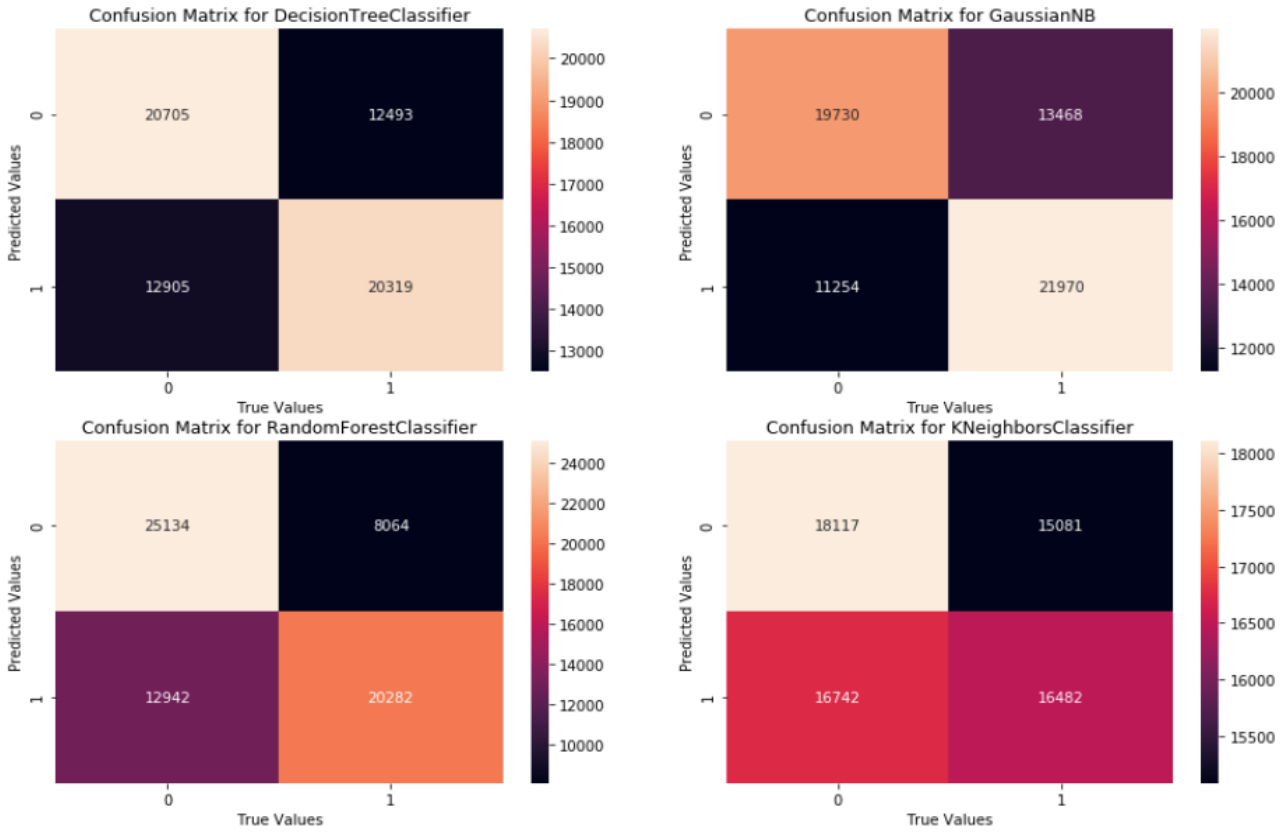
Evaluate Models

- Metrics for evaluation:
  - Accuracy
  - Recall
  - Speed

- A bias towards False Positives is preferred.
  - Sending help with no injury vs. not sending help with injury.

# Project Description

## Evaluate Models




Based on a first pass the Random Forest Classifier meets our criteria for preferring False Positives.

# Project Description

## Evaluate Models

- Cross Validation, scoring based on each models recall.
  - Used to evaluate the performance of each model on subsets of the data set.

| | Cross Validation Mean | Std |
|---|---|---|
| DecisionTreeClassifier | 0.613694 | 0.002187 |
| GaussianNB | 0.658194 | 0.010651 |
| RandomForestClassifier | 0.610062 | 0.003773 |
| KNeighborsClassifier | 0.497710 | 0.002036 |

The Gaussian Naïve-Bayes Algorithm scores the highest. Still below our threshold of 90%.

# Project Description: Methods

- Algorithms selected:
  - Decision Tree
  - Gaussian Naïve-Bayes
  - K-Nearest Neighbors
  - Random Forest

- Classification algorithms build networks of similar instances in the data.

- Car accidents have many similar factors with different outcomes

# Project Description: Results

- Best performing models:
  - GaussianNB
    - Recall: 66%
    - Accuracy: 63%
  - Random Forest Classifier:
    - Recall: 75%
    - Accuracy: 69%

Based on the outcomes of these models it is not possible to build a model with 90% recall and accuracy that can predict accident outcomes.

# Conclusion

- Goal:
  - Prove that machine learning can create an accurate model for accident harm prediction. This was not successful

- Possible Reasons:
  - Data was lacking pertinent information.
  - Weather data must be more localized.
  - Lack of standardization in reporting.

# Conclusion – Next Steps

- Gather more precise data about motor vehicle collisions.
- Use real time traffic and weather reports to aid in model training.
- Eventually combine this with a network of cameras.