Methodology

Crime Data Clustering

Processing Arrest Data from Long to Wide Format

Principal Component Analysis to Reduce Crime Columns

K-Means Clustering with K = 2, on PCA and Non PCA Crime Columns

Shooting Data Processing

Processing Shooting Data, adding time bucket, times square distance, etc.

Taking processed shooting data to same granularity as clustered crime data

Adding crime cluster (safe/ unsafe) to shooting data, convert it to supervised data

Shooting Data Prediction

Running ANN & Random Forest model and validate the generated supervised model with good testing model results

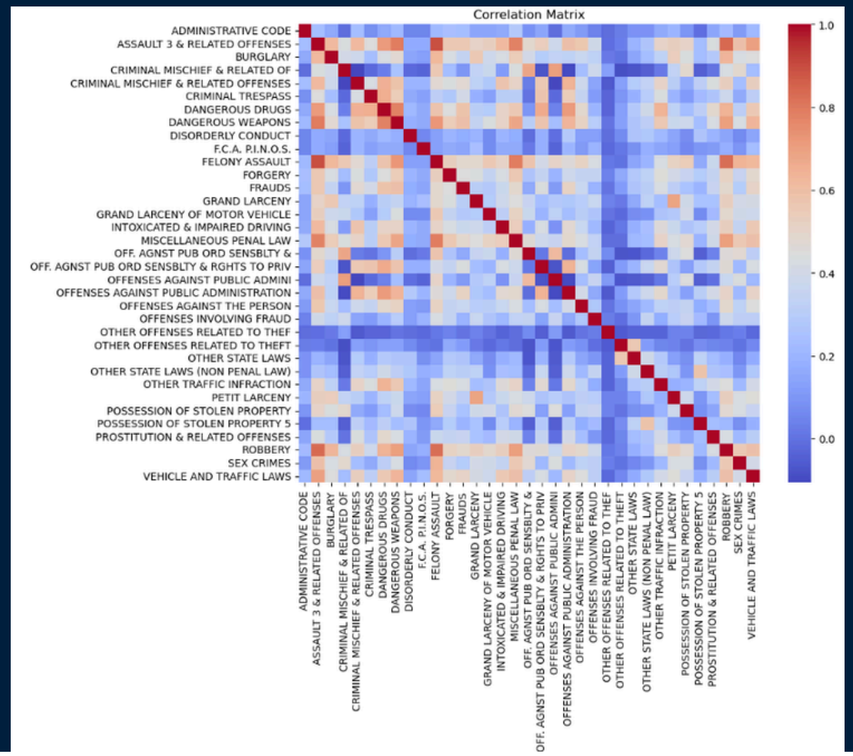
Scaling the dataset, with Minmax scaling,

Addition of more variables such as Unemployment flag, temperature, holidays, seasonality, facilities count, etc.

Data Pre-Processing

Crime Data Clustering

We performed PCA to reduce the crime columns from 35 to 26




```
Recommended variables based on PCA analysis:
PC1: ['ASSAULT 3 & RELATED OFFENSES', 'FELONY ASSAULT', 'ROBBERY']
PC2: ['CRIMINAL MISCHIEF & RELATED OF', 'OFFENSES AGAINST PUBLIC ADMINI', 'OFF. AGNST PUB ORD SENSBLTY &']
PC3: ['OTHER STATE LAWS (NON PENAL LAW)', 'OTHER STATE LAWS', 'POSSESSION OF STOLEN PROPERTY 5']
PC4: ['OTHER OFFENSES RELATED TO THEFT', 'OTHER STATE LAWS', 'F.C.A. P.I.N.O.S.']
PC5: ['ADMINISTRATIVE CODE', 'CRIMINAL TRESPASS', 'PETIT LARCENY']
```

Shooting Prediction

Taking shooting data to the granularity of crime clusters, i.e., Boro x Precinct x Jurisdiction x Time period

Adding the cluster column to shooting data, labels reflect “safe but vulnerable to shooting” or “unsafe and vulnerable to shooting”

Calculated distances from Times Square and Grand Central using latitude, longitude, and shooting location data.

Enriched the dataset with holiday indicators, seasonality, unemployment status, and temperature-based flags

Variable Selection

```
# Filter on followng features

features = [
     'BORO'
    , 'PRECINCT'
    , 'JURISDICTION_CODE'
    , 'After_6PM_Flag'
   ,'Times Square Distance'
#    , 'Grand Central Distance'
    , 'Murder_Flag'
    ,'More_Than_25_Years'
#   , 'month'
   , 'seasonality'
   , 'New Year = 1'
   ,'Christmas = 12'
   , 'Thanksgivigiving = 11'
  , 'Indep. Day = 7'
    ,'Halloween = 10'
   , 'year'
    , 'Unemployment_Flag'
#    , 'Year_month_temperature'
   ,'Avg Temp (>=70 F) Flag'
    , 'Avg Temp (<40 F) Flag'
    # ,'ADMINISTRATION OF GOVERNMENT'
    ,'CORE INFRASTRUCTURE AND TRANSPORTATION'
    # ,'EDUCATION, CHILD WELFARE, AND YOUTH', 'HEALTH AND HUMAN SERVICES'
    # ,'LIBRARIES AND CULTURAL PROGRAMS', 'PARKS, GARDENS, AND HISTORICAL SITES'
   ,'PUBLIC SAFETY, EMERGENCY SERVICES, AND ADMINISTRATION OF JUSTICE'
     ]
```

- All variables which are uncommented were used to train the prediction model
- “Times Square Distance” outperformed “Grand Central Distance” as a predictor
- The “month” variable was excluded to avoid overlap with multiple holiday flags
- “Year_month_temperature” was too granular; replaced with summer (>70°F) and winter (<40°F) flags
- Facility types like “CORE INFRASTRUCTURE” and “PUBLIC SAFETY” showed stronger relevance to shooting patterns

Model Building

Modeling Approach Overview

- We implemented and compared two supervised classification models:

  Random Forest

  Artificial Neural Network (ANN) using MLPClassifier

- The target variable classified locations as:

  “Safe but Vulnerable to Shooting” (0)

  “Unsafe and Vulnerable to Shooting” (1)

Random Forest

```
rft = RandomForestClassifier(
    n_estimators=n_estimators_val,
    max_depth=max_depth_val,
    min_samples_split=min_samples_split_val,
    bootstrap=True,
    random_state=42
)
```

```
# Hyperparameter lists
n_estimators_list = [100, 200, 300]
max_depth_list = [10, 15, 20]
min_samples_split_list = [2, 5, 10]
```

Neural Network

```
ann = MLPClassifier(
    max_iter=max_iter_val,
    batch_size=batch_size_val,
    activation='relu',
    solver=solver_val,
    random_state=42,
    learning_rate='adaptive',
)
```

```
# Hyperparameter lists
max_iters = [1000, 2000, 3000]
batch_sizes = [64, 128, 256]
solvers = ['adam', 'lbfgs', 'sgd']
```

Model Results – Clustering

Crime Stats across Clusters

- K-Means Clustering Achieves Strong Separation of High- and Low-Crime Areas

Cluster	Avg. Crime Index (PCA)	Avg. Crime Index (Non-PCA)	Type	Coverage
0	-0.6	21.2	Safe	25,377
1	7.7	201.2	Unsafe	8,116

- K-Means Remains Robust at Borough Granularity

Avg. Crime Index (PCA)		Brooklyn	Manhattan	Queens	Staten Island
0	-0.7	-0.7	-0.2	-1.0	-1.0
1	10.3	6.9	7.6	6.5	6.0

Avg. Crime Index (Non-PCA)		ooklyn	Manhattan	Queens	Staten Island
0	17.9	21.7	23.0	18.0	32.3
1	240.9	201.5	181.1	182.4	213.2

Model Results – Classification

Random Forest Performance Metrics

Top Performing Model	
N Estimators	300
Max Depth	15
Min Sample Split	10
Train Accuracy	96.3%
Test Accuracy	90.7%

		Predicted Label	
		Safe but Vulnerable to Shooting (0)	Unsafe and Vulnerable to Shooting (1)
True Label	Safe but Vulnerable to Shooting (0)	939	156
	Unsafe and Vulnerable to Shooting (1)	77	1332

Class 0: Safe but Vulnerable to Shooting

- Precision: 92.4%
- Recall: 85.7%
- F1 Score: 88.9%

Class 1: Unsafe and Vulnerable to Shooting

- Precision: 89.5%
- Recall: 94.5%

Neural Network Performance Metrics

Top Performing Model	
Max iter	2000
Batch Size	64
Solver	lbfgs
Train Accuracy	85.1%
Test Accuracy	83.5%

		Predicted Label	
		Safe but Vulnerable to Shooting (0)	Unsafe and Vulnerable to Shooting (1)
True Label	Safe but Vulnerable to Shooting (0)	799	296
	Unsafe and Vulnerable to Shooting (1)	118	1291

Class 0: Safe but Vulnerable to Shooting

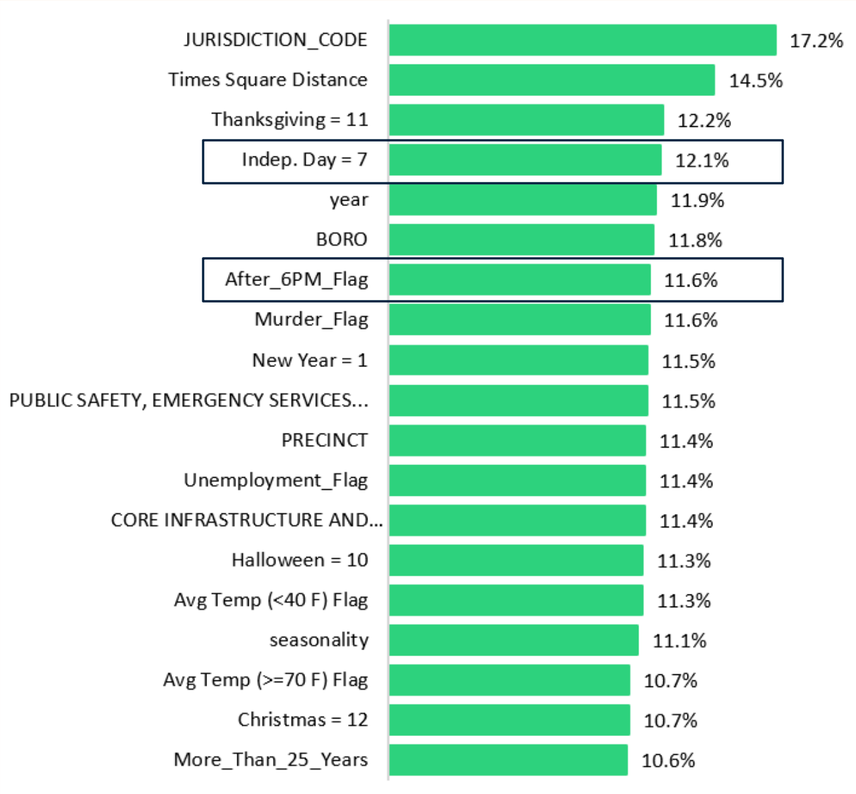
- Precision: 87.1%
- Recall: 72.9%
- F1 Score: 79.0%

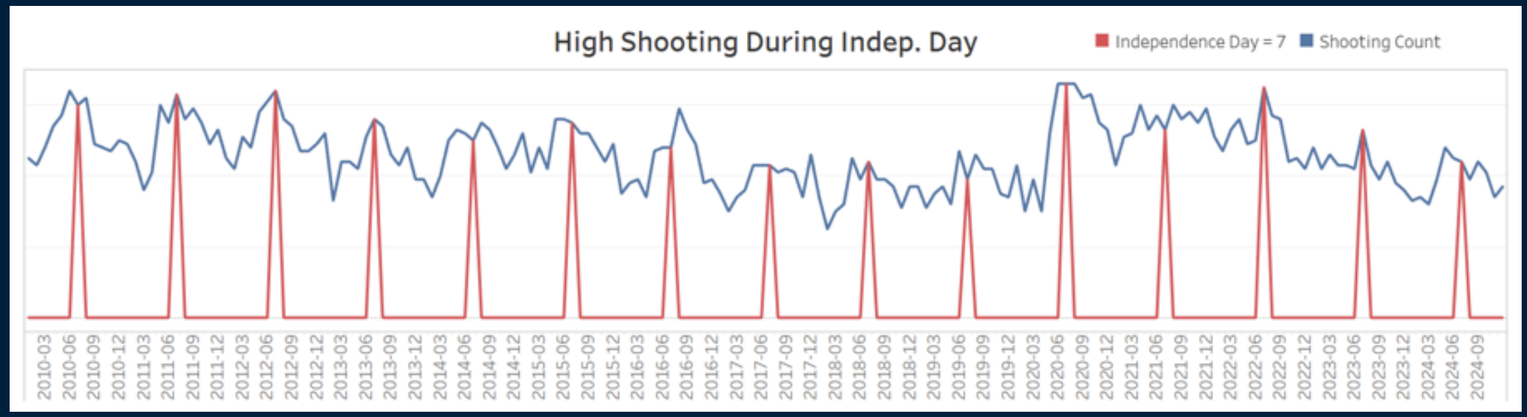
Class 1: Unsafe and Vulnerable to Shooting

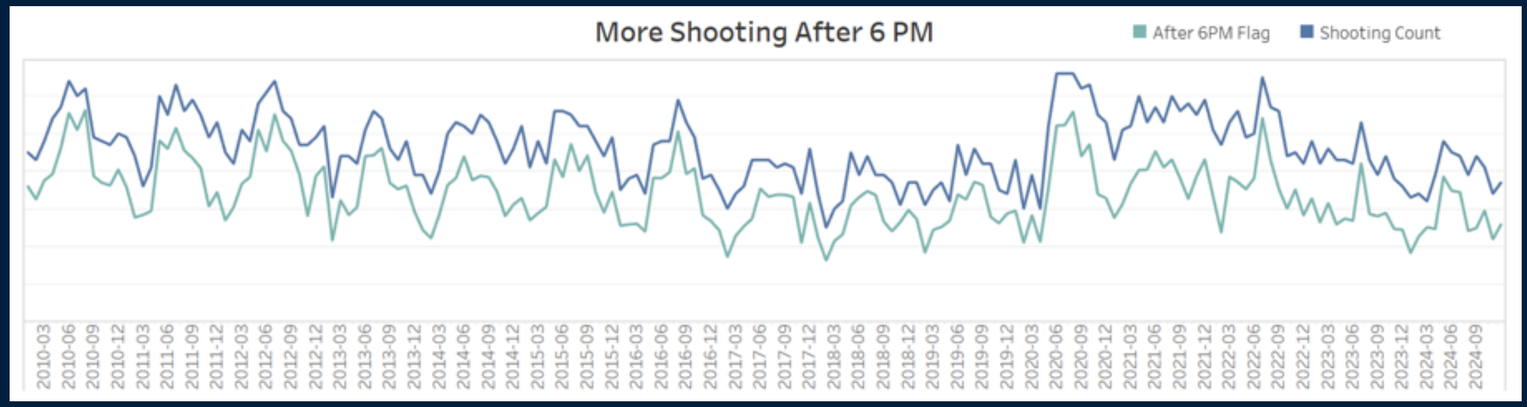
- Precision: 81.4%
- Recall: 91.6%

Top Features

Most features show similar importance, suggesting no single variable dominates the prediction.

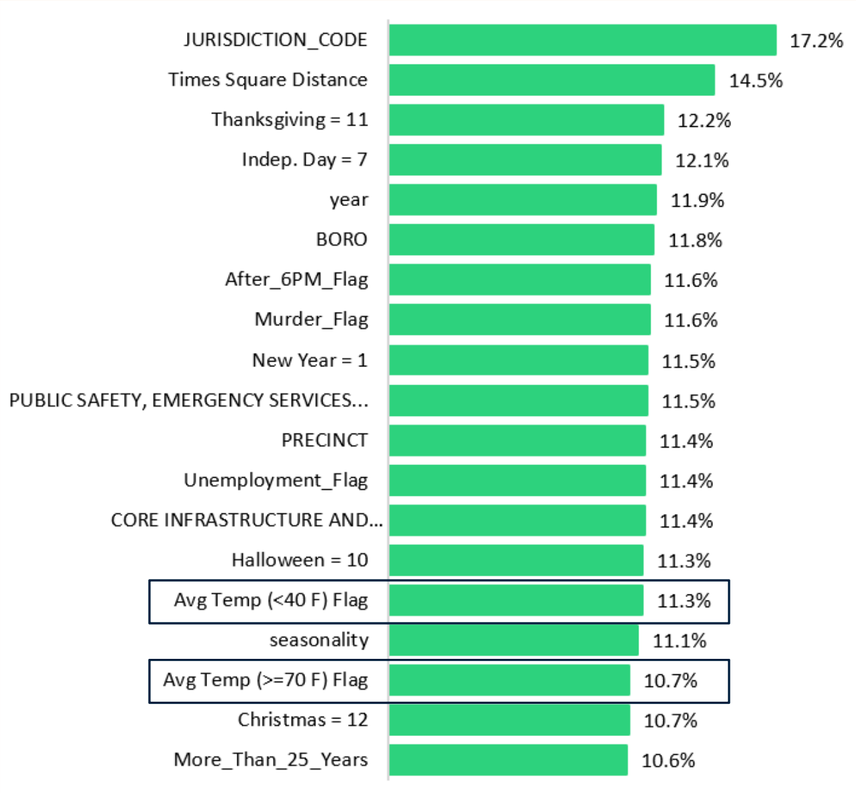


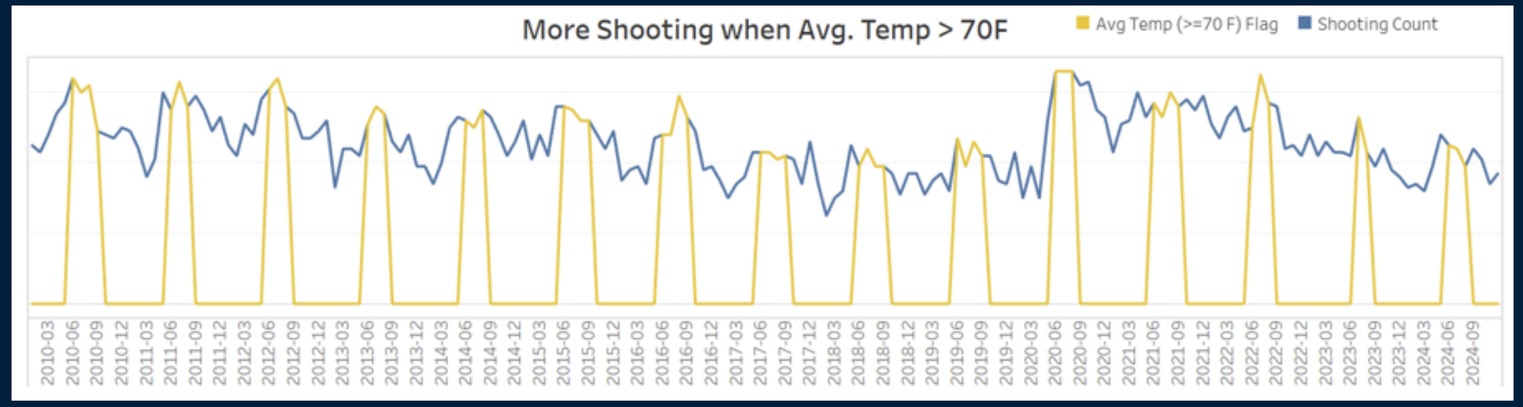


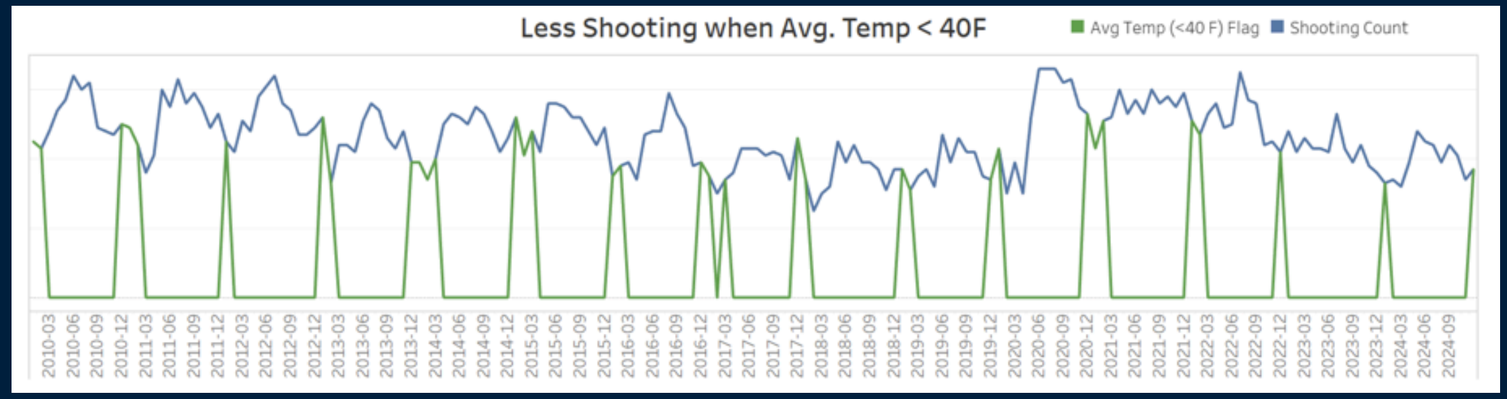


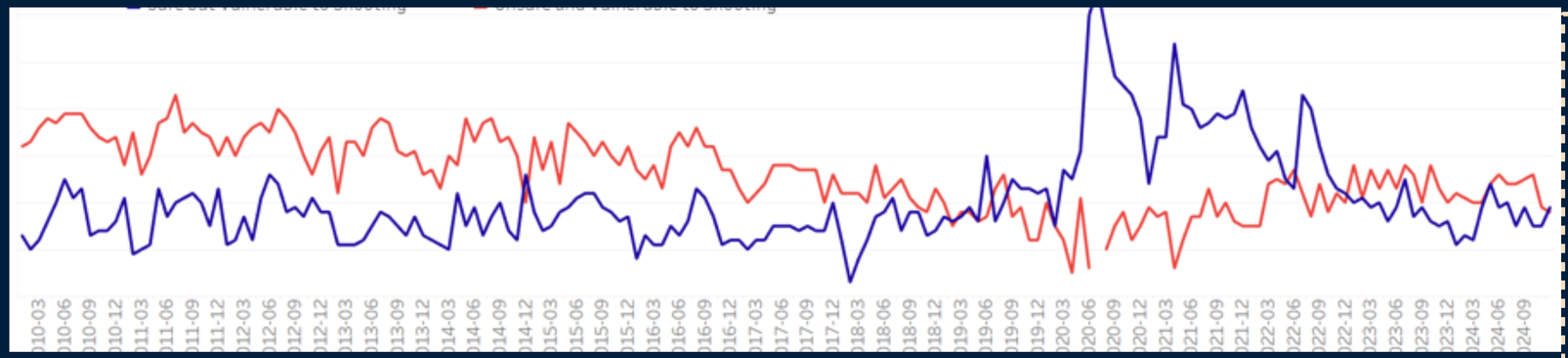
Top Features

Most features show similar importance, suggesting no single variable dominates the prediction.









Though different kinds of Crime declined over time in NYC (including shooting in these areas), but Shooting increased in historically safer area especially around 2019-2022