Machine Learning with Police Homicide Data

Felix Chen, Anita Sun, Moe Wu, River Wang

Introduction

Motivation

- Police homicides associated with location and with measures of racial threat, social disorganization, and availability of firearms (e.g., Maksuta et al., 2024)
- Scholars have suggested the necessity of more nuanced interracial comparisons in terms of police homicides (e.g., Wilkes and Karimi, 2023)
- **Conclusion:** We need sophisticated modeling.

Introduction

Research Question

- To what extent can demographic attributes (e.g., race, income, education level) of a census tract within and adjacent to the three largest cities in the US predict whether or not a police homicide will occur in that census tract in the next 10 years?

Introduction

The Machine Learning Task

- Extant literature: e.g., Prabakaran and Mitra's (2019) PCA + clustering
- **Our machine-learning task:** Supervised classification
- Attributes: Tract-level demographics
  - Income: the median income estimate based on the previous 12 months of earnings (1) household median income, (2) white household median, (3) Black household median
  - (4) estimate of the % of the adult population who hold a high-school diploma or above
  - (5) estimate of the percentage of population that is white, (6) % of population Black
- Target: binary variable of whether any police homicide would take place within the boundary of the tract in the period of 2013 to the recent past (approximately in January 2025)

Data Sets

3 Different Sources

- MappingPoliceViolence.org, Police homicides
  - (Campaign Zero, 2024)
  - Contains information from 2013-2025
  - For US cities with 100,000+ population
- Census tract shapefiles (from 2013)
  - GEO_ID and shapely.polygon information
  - Downloaded for California, Illinois, and New York
- American Community Survey demographic data (from 2013)
  - Household Income, Education, and Racial Demography 5-year estimate
  - Downloaded for California, Illinois, and New York

Data Sets

1) Police Homicide Dataset

```
pol_df
```

	name	age	gender	race	victim_image	date	street_address
0	Steven Espinoza	36.0	Male	Hispanic	https://i0.wp.com/iecn.com/wp-content/uploads/...	1/12/2025	N Mountain Ave and 11th St
1	Jose Evans	42.0	Male	Hispanic	https://wgntv.com/wp-content/uploads/sites/5/2...	1/12/2025	8500 block of Cermak Rd
2	Benjamin Prowell, Jr.	34.0	Male	Black	https://cache.legacy.net/legacy/images/cobrand...	1/11/2025	10000 block of Crystal Hill Rd
3	Brian Rolstad	43.0	Male	Unknown race	NaN	1/11/2025	900 block of W 23rd St
4	Devin Shields	23.0	Male	Unknown race	NaN	1/11/2025	2300 block of Waverly Dr
...	...	...	...	...	...	...	...

Data Sets

1) Police Homicide Dataset

```
pol_df.shape, pol_df.columns
```

```
((14021, 38),
 Index(['name', 'age', 'gender', 'race', 'date', 'street_address', 'city',
        'state', 'zip', 'county', 'agency_responsible', 'ori', 'cause_of_death',
        'circumstances', 'disposition_official', 'officer_charged', 'news_urls',
        'signs_of_mental_illness', 'allegedly_armed', 'wapo_armed',
        'wapo_threat_level', 'wapo_flee', 'geography', 'encounter_type',
        'initial_reason', 'call_for_service', 'tract',
        'hhincome_median_census_tract', 'latitude', 'longitude',
        'pop_total_census_tract', 'pop_white_census_tract',
        'pop_black_census_tract', 'pop_native_american_census_tract',
        'pop_asian_census_tract', 'pop_pacific_islander_census_tract',
        'pop_other_multiple_census_tract', 'pop_hispanic_census_tract'],
       dtype='object'))
```

Data Sets

2) Census Tract Shape Files

```
gdf.shape
✓ 0.0s

(3265, 14)
```

```
gdf.iloc[0]
✓ 0.0s

STATEFP                                                     17
COUNTYFP                                                   161
TRACTCE                                                 022800
GEOID                                              17161022800
GEOIDFQ                                   1400000US17161022800
NAME                                                       228
NAMELSAD                                     Census Tract 228
MTFCC                                                    G5020
FUNCSTAT                                                     S
ALAND                                                  2103943
AWATER                                                       0
INTPTLAT                                           +41.4991066
INTPTLON                                          -090.5472913
geometry    POLYGON ((-90.557243 41.494331, -90.55724 41.4...
Name: 0, dtype: object
```

```
gdf.dtypes
✓ 0.0s

STATEFP       object
COUNTYFP      object
TRACTCE       object
GEOID         object
GEOIDFQ       object
NAME          object
NAMELSAD      object
MTFCC         object
FUNCSTAT      object
ALAND          int64
AWATER         int64
INTPTLAT      object
INTPTLON      object
geometry    geometry
dtype: object
```

Data Sets

Merged Dataset for Model Training/Validation/Eval.

```python
pd.concat([df_ny[colname_dic.keys()], df_ny["target"]], axis=1)
```

✓ 0.0s

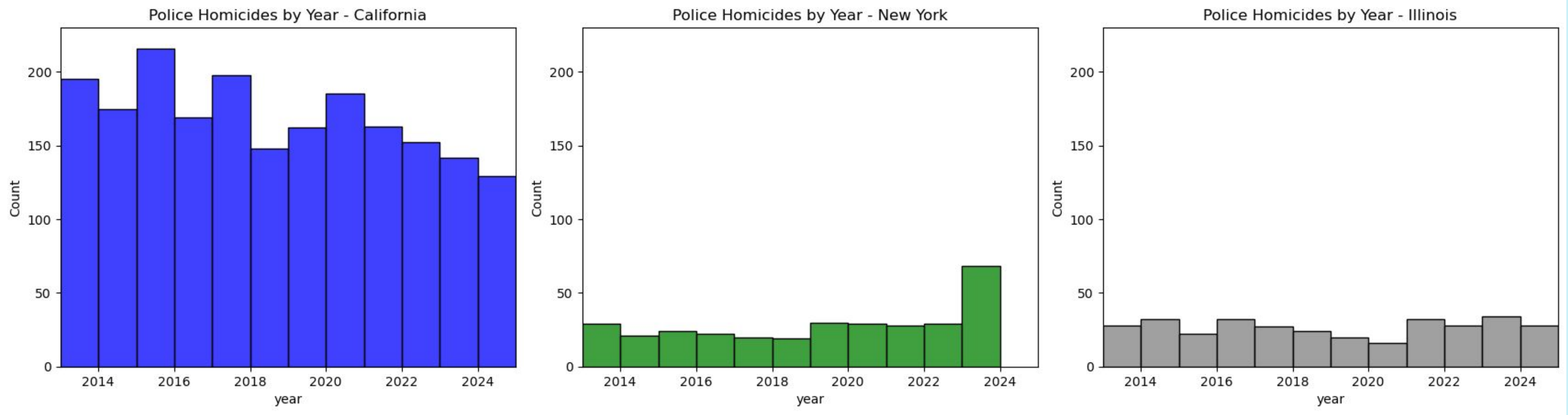
	S1501_C01_014E	S1903_C02_001E	S1903_C02_002E	S1903_C02_003E	DP05_0032PE	DP05_0033PE	target
89	34.0	NaN	NaN	NaN	14.8	58.6	0.0
90	77.9	69514.0	60795.0	86964.0	27.8	31.0	0.0
91	83.3	73036.0	58036.0	76346.0	25.1	29.3	0.0
92						32.9	0.0
93						25.5	0.0
...						...	...
3950						24.2	0.0
3951						50.6	1.0
3952						66.4	0.0
3953						26.8	0.0
3954	NaN	NaN	NaN	NaN	NaN	NaN	0.0

2167 rows × 7 columns

```
{"S1501_C01_014E": "highschool degree over higher rate",
 "S1903_C02_001E": "household median income",
 "S1903_C02_002E": "white household median income",
 "S1903_C02_003E": "black household median income",
 "DP05_0032PE": "white race rate",
 "DP05_0033PE": "black race rate"
 }
```

EDA for Police Homicide

Police Homicide Raw Count in Each State (Pre- ACS merge)



EDA for ACS

Descriptive Stat for Census Tracts

6 features:

Median Household income

Median White Household income

Median Black Household income

% of Adults with High School or Higher

% of Population that is White

% of Population that is Black

Sample output: Chicago tracts descriptive stat

```
Desc for Chicago

highschool degree over higher rate
 count    1315.000000
mean        83.218403
std         13.001100
min         29.100000
25%         75.850000
50%         86.400000
75%         93.500000
max        100.000000
Name: S1501_C01_014E, dtype: float64


household median income
 count      1315.000000
mean      56258.679087
std       28395.446593
min        5725.000000
25%       36348.500000
50%       51250.000000
75%       70592.500000
max      231875.000000
Name: S1903_C02_001E, dtype: float64


white household median income
 count      1134.000000
mean      65150.330688
std       30361.879649
min        2500.000000
25%       44707.000000
50%       58965.500000
75%       80274.250000
max      250000.000000
Name: S1903_C02_002E, dtype: float64
```

```
black household median income
 count       869.000000
mean       41633.201381
std        29142.975107
min         2500.000000
25%        22582.000000
50%        34207.000000
75%        51342.000000
max       250000.000000
Name: S1903_C02_003E, dtype: float64


white race rate
 count    1315.000000
mean       53.228669
std        32.695498
min         0.000000
25%        23.050000
50%        62.100000
75%        82.200000
max        99.400000
Name: DP05_0032PE, dtype: float64


black race rate
 count    1315.000000
mean       29.488137
std        37.566123
min         0.000000
25%         1.700000
50%         6.200000
75%        62.650000
max       100.000000
Name: DP05_0033PE, dtype: float64
```

Target Variable

Post-Merge Processing

df['target']

1: census tract has at least one police homicide from 2013-2025

0: census tract has no police homicides from 2013-2025

Data sparsity issue; leads to aggregation of our target variable across years and condensing into binary class

```
New York
target
0    2048
1     119
Name: count, dtype: int64


Los Angeles
target
0    1944
1     402
Name: count, dtype: int64


Chicago
target
0    1185
1     134
Name: count, dtype: int64
```

Note: after merge, we drop all census tracts except for the boroughs of New York City, Los Angeles County, and Cook County in which Chicago is located.

Models

Tree-based Model: Data processing

1. Remove the str expressions before changing data type
2. Convert the data type to numerical value
3. Checking missing values and fill them with mean value
4. using min-max scaler to standardize variables

```
df.isna().sum()
```

✓ 0.0s

```
target            0
S1501_C01_014E   18
S1903_C02_001E   34
S1903_C02_002E   40
S1903_C02_003E  941
DP05_0032PE      17
DP05_0033PE      17
dtype: int64
```

```
#convert the data type to numerical value
df[selected_columns] = df[selected_columns].astype('float')
df2 = df.copy()
```

✓ 0.0s

```
#replace missing value in x using mean
for col in x_variable:
    df[col].fillna(value = df[col].mean(), inplace=True)
```

✓ 0.0s

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(0,1))
df[x_variables] = scaler.fit_transform(df[x_variables])
```

✓ 0.0s

Models

Tree-based Model: Normal Tree

1. Explore the base model first. set criterion = "entropy", and keep all the other parameters with the default value.
2. Evaluate the DecisionTreeClassifier performance using kfolder cross validation
3. Through validation, the first model we explore should be improved on the precision and recall score for case 1. The low scores are likely due to the data imbalance.

```
# 2. Evaluate the DecisionTreeClassifier performance using kfolder
# Perform KFold splitting

kf = KFold(n_splits=5, shuffle=True, random_state=929)
scores = cross_validate(dt_clf1, X_train1, y_train1, cv=kf,
                        scoring=['accuracy', 'precision', 'recall'])


print("Accuracy scores: ", scores['test_accuracy'])
print("Precision scores: ", scores['test_precision'])
print("Recall scores: ", scores['test_recall'])

print("Mean Accuracy: ", scores['test_accuracy'].mean())
print("Mean Precision: ", scores['test_precision'].mean())
print("Mean Recall: ", scores['test_recall'].mean())
```

✓ 0.1s

```
Accuracy scores:  [0.69946809 0.70666667 0.74933333 0.728      0.688     ]
Precision scores:  [0.12903226 0.19736842 0.25       0.16       0.20512821]
Recall scores:  [0.11940299 0.234375   0.28333333 0.23529412 0.22535211]
Mean Accuracy:  0.7142936170212766
Mean Precision:  0.18830577684907057
Mean Recall:  0.21955150974621507
```

Models

Tree-based Model: Normal Tree

1. Under-sampling : extract 1/4 from the x value data in the original data for training
2. Based on the feature importance and multicollinearity, I further filtered out two features
3. Considering the model complexity, it may risk overfitting. so we then tune parameters using gridSearchCV

```
grid_cv = GridSearchCV(estimator=dt_clf2, param_grid=param_grid, cv=5,
                    scoring={
                     'precision': 'precision',
                     'recall': 'recall',
                     'accuracy': 'accuracy',
                     'f1-score': "f1"},
                     refit='recall') #choose recall here because the most important info
```

```
Best parameters found:
{'max_depth': 5, 'max_features': 4,
'min_samples_leaf': 1,
'min_samples_split': 2}
Best score achieved:
0.5078341013824884
```

Models

Tree-based Model: Random Forest

```
rf_clf = RandomForestClassifier(random_state=929)
rf_param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [5, 10, 15, 20],
    'max_features': [3, 4],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}

rf_grid = GridSearchCV(
    estimator=rf_clf,
    param_grid=rf_param_grid,
    cv=5,
    scoring = {
    'precision': 'precision',
    'recall': 'recall',
    'accuracy': 'accuracy',
    'f1-score': "f1"
},
    refit='recall')
```

```
Best parameters found:
{'max_depth': 15,
'max_features': 4,
'min_samples_leaf': 1,
'min_samples_split': 5,
'n_estimators': 200}

Best score achieved:
0.4665642601126472З
```

Models

Tree-based Model : Model Evaluation

	precision	recall	f1-score	support
0.0	0.82	0.81	0.82	381
1.0	0.24	0.26	0.25	89
accuracy			0.71	470
macro avg	0.53	0.54	0.54	470
weighted avg	0.71	0.71	0.71	470

Default model

Fine tuned dt

	precision	recall	f1-score	support
0.0	0.81	0.84	0.82	381
1.0	0.19	0.17	0.18	89
accuracy			0.71	470
macro avg	0.50	0.50	0.50	470
weighted avg	0.69	0.71	0.70	470

Random forest

	precision	recall	f1-score	support
0.0	0.82	0.61	0.70	381
1.0	0.21	0.44	0.28	89
accuracy			0.58	470
macro avg	0.52	0.52	0.49	470
weighted avg	0.71	0.58	0.62	470

Models

Tree-based Model : Model Evaluation

LA

	precision	recall	f1-score	support
0.0	0.82	0.61	0.70	381
1.0	0.21	0.44	0.28	89
accuracy			0.58	470
macro avg	0.52	0.52	0.49	470
weighted avg	0.71	0.58	0.62	470

NY

	precision	recall	f1-score	support
0.0	0.97	0.81	0.88	419
1.0	0.07	0.40	0.12	15
accuracy			0.79	434
macro avg	0.52	0.60	0.50	434
weighted avg	0.94	0.79	0.86	434

Chicago

	precision	recall	f1-score	support
0.0	0.92	0.36	0.51	241
1.0	0.09	0.70	0.16	23
accuracy			0.39	264
macro avg	0.51	0.53	0.34	264
weighted avg	0.85	0.39	0.48	264

Models

Logistic Regression Model: Data processing

1. Remove the str expressions before changing data type
2. Convert the data type to numerical value
3. Checking missing values and fill them with mean value
4. Using standard scaler to standardize variables

Models

Logistic Regression Model: Training

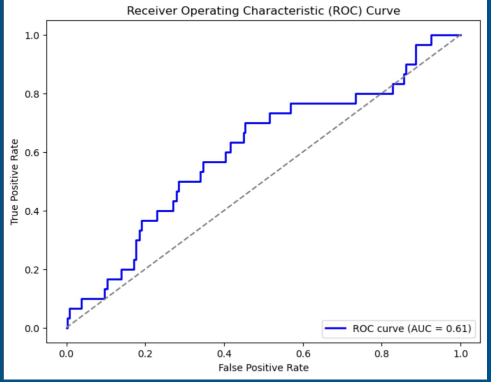
1. Undersampling the training data in favor of the minority class
2. Use grid search to find the best model for each of the three cities
3. Solver is ‘liblinear’ because it works well with small datasets
4. Use ‘recall’ as the metric to ensure good performance in predicting true homicides
5. Find the best parameters, generate classification reports, analyze the predictive value of coefficients, and use roc-auc to check for validity of the model

Models

Logistic Regression Model: New York

1. Percentage of black population is the most predictive of future police homicides
2. Percentage of high school graduates and white household income are also informative
3. Race matters, while income and education also have some correlation to future homicides
4. The model is somewhat useful in separating the two classes

	precision	recall	f1-score	support
0	0.96	0.50	0.65	404
1	0.09	0.70	0.17	30
accuracy			0.51	434
macro avg	0.53	0.60	0.41	434
weighted avg	0.90	0.51	0.62	434

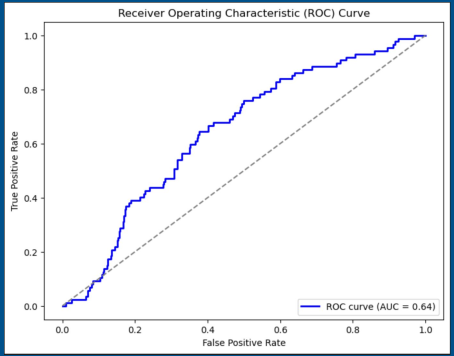


Models

Logistic Regression Model: Los Angeles

1. Percentage of black population is again the most predictive of homicides
2. Percentage of individuals with high school degree or above is a close second
3. Percentage of white population has some predictive value.
4. Similar to New York City, race is informative of future homicides, and education also matters

	precision	recall	f1-score	support
0	0.88	0.60	0.71	383
1	0.27	0.66	0.38	87
accuracy			0.61	470
macro avg	0.58	0.63	0.55	470
weighted avg	0.77	0.61	0.65	470

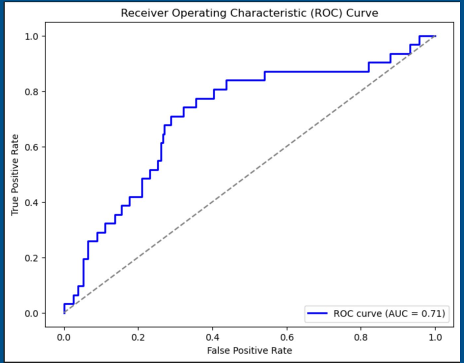


Models

Logistic Regression Model: Chicago

1. Percentage of black population is still the most predictive of homicides
2. Percentage of white population is a close second
3. Household median income has some correlation to future homicides
4. Across all three cities, race is constantly informative for prediction of homicides, especially in Chicago

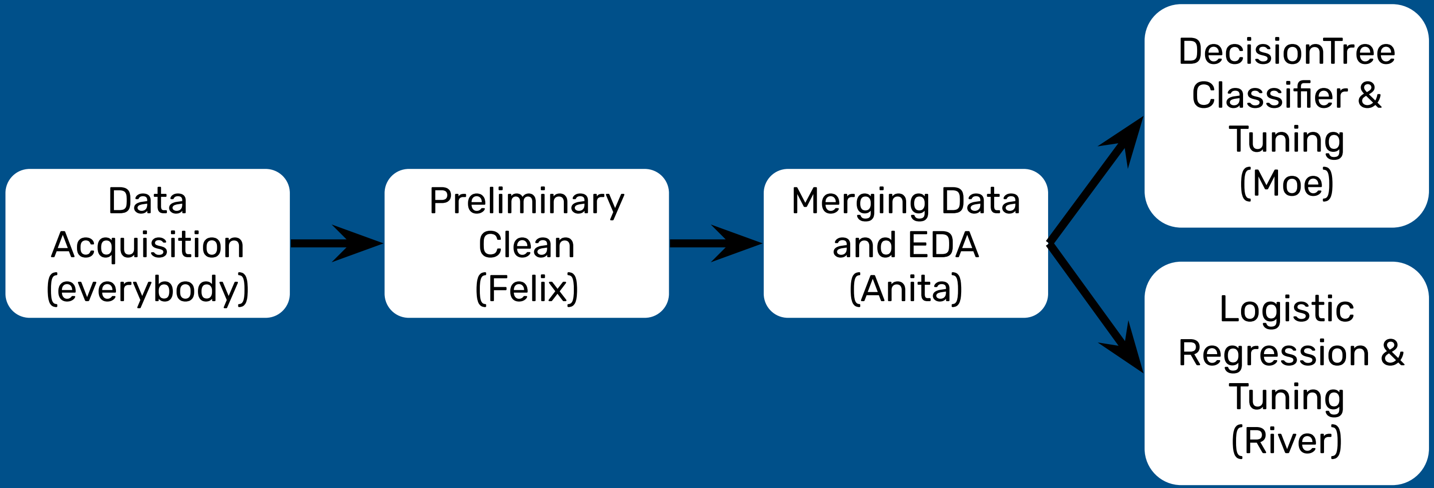
	precision	recall	f1-score	support
0	0.95	0.69	0.80	233
1	0.23	0.71	0.35	31
accuracy			0.69	264
macro avg	0.59	0.70	0.58	264
weighted avg	0.86	0.69	0.75	264



Thanks for Listening!

Data Pipeline

Our Process



Peer Feedback

- Unit of Analysis
  - Could use a bigger area for unit of analysis
  - Zip code or county
  - Voting districts - but gerrymandering concerns
- Population confounding
  - Census tracts might vary in population size
  - Density may be a factor in homicide