# Unlocking the Secrets: Forecasting Crime Patterns in High-Risk Los Angeles Neighborhoods

Presented by : Dhanashree Badhe

Fall 2023 | BANA 273: Machine-Learning Analytics

# TOPIC OVERVIEW

- Expansive urban landscape of Los Angeles → diverse range of criminal activities
- Navigates a complex interplay of: socioeconomic factors, cultural diversity, & geography
- Potential unknown threatening dangers in urban landscape of Los Angeles

## ANALYZE

- Past crime data

## PREDICT

- Determine whether the situation is a violent crime

# THE DATA

Crime Data in Los Angeles (2010 to 2020)

**Kaggle**

Crime Data in Los Angeles (2020 to Present)

Crime Incidents in LA between years 2010 and 2022

276584 rows x 28 columns

# DATA CLEANING

- **Missing values**
  - Dropped (i.e. Premis Cd)
  - Average (i.e. Age)
- **Outlier Management**
  - Age < 0 → replacement with average age
- **Dropping Irrelevant Columns**
  - (i.e. Area Desc.)

Missing Values

Click here to ask Blackbox to help you code faster |

```python
#understanding missing values
miss_val = df.isnull().sum().sort_values(ascending=False)
miss_val.head(10)
```

[613]

```
Crm Cd 4           276559
Crm Cd 3           275808
Crm Cd 2           253843
Cross Street       226771
Weapon Used Cd     175499
Weapon Desc        175499
Mocodes             37993
Vict Descent        36362
Vict Sex            36357
Premis Desc            97
dtype: int64
```

Click here to ask Blackbox to help you code faster |

```python
#understanding percentage missing values
percent_miss = (df.isnull().sum() * 100)/df.isnull().count()
percent_miss = percent_miss.sort_values(ascending=False)
percent_miss.head(10)
```

[614]

```
Crm Cd 4           99.990961
Crm Cd 3           99.719434
Crm Cd 2           91.777905
Cross Street       81.989920
Weapon Used Cd     63.452333
Weapon Desc        63.452333
Mocodes            13.736514
Vict Descent       13.146820
Vict Sex           13.145012
Premis Desc         0.035071
dtype: float64
```

# DATA PREPARATION

- **Binning**
  - Age → Equal Frequency & Equal binning
- **Conversion**
  - Date → extracted day, weekday, month, & hour
- **Attribute Consolidation**
  - 132 crime types – converted to 6 & later 2 buckets

```
# Victim age into different bins
#Victim_Age1 - bins defined by us

df['Vict_Age1'] = pd.cut(df['Vict Age'], bins=[-10, 18, 30, 50, 70, 100], labels=[1,2,3,4,5])

df['Vict_Age2'] = pd.qcut(df['Vict Age'], 5 , labels=[1,2,3,4,5])
```

| Crime_Category | DayName | MonthName | Year | Date | Month | Hour | Vict_Age1 | Vict_Age2 |
|---|---|---|---|---|---|---|---|---|
| 1 | Wednesday | January | 2020 | 8 | 1 | 6 | 3 | 3 |
| 1 | Wednesday | January | 2020 | 1 | 1 | 1 | 2 | 1 |
| 0 | Wednesday | September | 2020 | 16 | 9 | 3 | 4 | 5 |
| 0 | Wednesday | January | 2020 | 1 | 1 | 5 | 5 | 5 |
| 0 | Wednesday | January | 2020 | 1 | 1 | 1 | 3 | 2 |
| 2 | Wednesday | January | 2020 | 1 | 1 | 1 | 2 | 1 |
| 0 | Thursday | January | 2020 | 2 | 1 | 4 | 2 | 1 |
| 0 | Saturday | January | 2020 | 4 | 1 | 1 | 3 | 3 |
| 3 | Saturday | January | 2020 | 4 | 1 | 1 | 2 | 1 |
| 0 | Saturday | September | 2020 | 12 | 9 | 1 | 2 | 1 |

# DATA ANALYSIS: Models

## Tree-Based Models

1. Decision Trees
2. Random Forests
3. Extra Trees

Gini and Entropy Criteria

## Naive Bayes

1. Multinomial Naive Bayes
2. Gaussian Naive Bayes
3. Complement Naive Bayes
4. Bernoulli Naive Bayes

# DATA ANALYSIS: Workflow

**Naive Bayes Models**

Stratified Sampling

**Decision Tree Models**

Pre-Pruning + Post-Pruning

Stratified Sampling

Ensemble Learning
Soft Voting
Hard Voting
Bagging
Boosting

| Model Name | Train Acc | Test Acc | Expected Value |
|---|---|---|---|
| Multinomial NB | 0.2537 | 0.25159... | -78316 |
| **Gaussian NB** | **0.47968...** | **0.47597...** | **215590** |
| Complement NB | 0.42244... | 0.42187... | 119674 |
| Bernoulli NB | 0.45537... | 0.45416... | 214102 |
| Decision Tree Classifier (Gini) | 0.99895... | 0.46449... | 162753 |
| Decision Tree Classifier (Entropy) | 0.99895... | 0.46716... | 165937 |
| **RandomForest** | **0.99894...** | **0.56410...** | **313554** |
| Extremely Randomized Trees (Gini) | 0.99895... | 0.54431... | 288804 |
| Extremely Randomized Trees (Entropy) | 0.99895... | 0.54048... | 284553 |
| Ensemble Learning (Soft) | 0.9215 | 0.5364 | 293610 |
| Ensemble Learning (Hard) | 0.9990 | 0.5245 | 261589 |

# STEPS TO IMPROVE ACCURACY FURTHER

## PART 2

### FEATURE SELECTION

1. FEATURE IMPORTANCES
   - ❑ Naive Bayes
   - ❑ Decision Trees
2. WRAPPER METHOD
   - ❑ Forward
   - ❑ Backward

### REDUCTION OF OUTPUT CATEGORIES

- ❑ From 6 to 2

## PART 3

### CATEGORY REDUCTION

**OVERSAMPLING 1 USING SMOTE SAMPLING METHOD**

**Crime Category % distribution from 95,5 to 50,50**

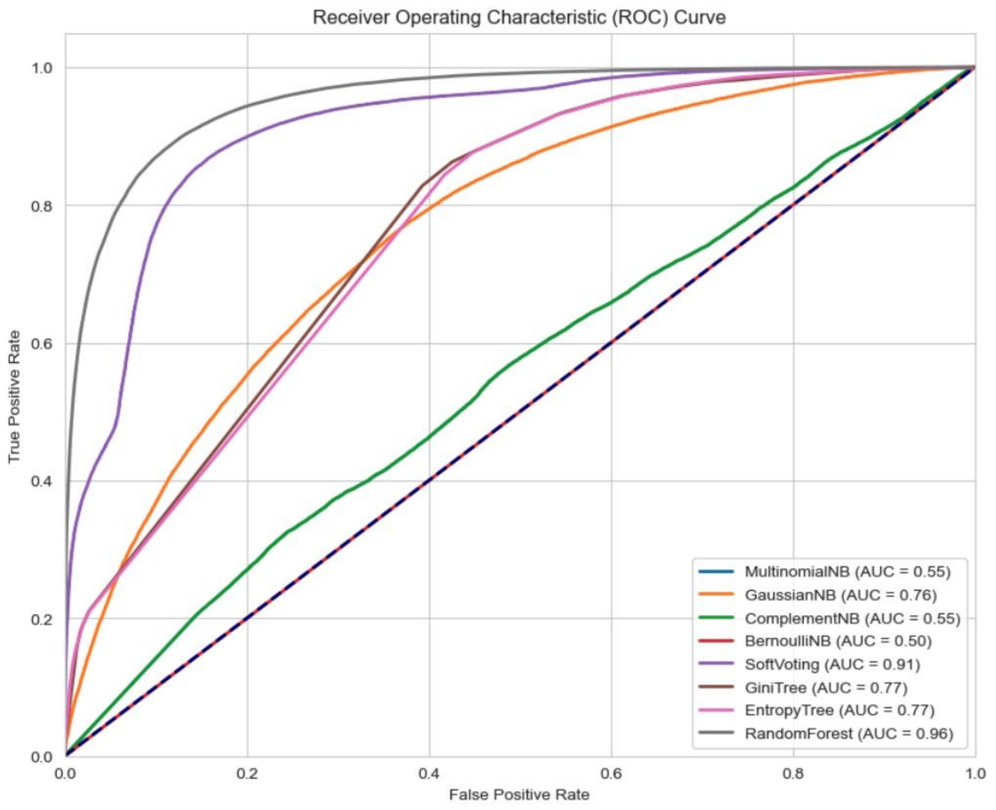
| | Part 2: Reducing Categories + Features | | | Part 3: Use SMOTE to create even distribution | | |
|---|---|---|---|---|---|---|
| | Train Accuracy | Test Accuracy | Expected Values | Train Accuracy | Test Accuracy | Expected Values |
| Str. Acc 0 | 0.95 | 0.95 | | 0.50 | 0.5 | |
| Multinomia NB | 0.6086 | 0.6087 | 741165 | 0.5331 | 0.5326 | 576546 |
| Gaussian NB | 0.9589 | 0.9572 | 2054035 | 0.6953 | 0.6967 | 962666 |
| Bernoulli NB | 0.9589 | 0.9572 | 2054035 | 0.5005 | 0.4987 | 617526 |
| Decision Tree Classifier (Gini) | 0.9999 | 0.9216 | 1904565 | 0.9999 | 0.8506 | 1346866 |
| pre-pruing | 0.9998 | 0.9216 | 1904565 | 0.9999 | 0.8506 | 1346866 |
| post-pruing | 0.9602 | 0.9583 | 2050325 | 0.5342 | 0.5289 | 282219 |
| Boosting | 0.9999 | 0.9486 | 2007415 | 0.9999 | 0.9095 | 1494306 |
| Decision Tree Classifier (Entropy) | 0.9999 | 0.9244 | 1916275 | 0.9999 | 0.8566 | 1362438 |
| Ensemble (Soft) | 0.9789 | 0.9583 | 2052955 | 0.9158 | 0.8182 | 1298066 |
| Ensemble(Hard) | 0.9998 | 0.9558 | 2038455 | 0.9999 | 0.8526 | 1352574 |
| Random Forest | 0.9998 | 0.9594 | 2051765 | 0.9999 | 0.8824 | 1425798 |

# MODEL ACCURACY EVALUATION

## Random Forest is BEST

Following...

Soft Voting
Gini Tree & Entropy Tree



# CONCLUSION

- If you give a new data row containing the following information, we can predict with Random Forest, if any violent/non-violent crime will occur with you.
    - Date
    - Month
    - Hour
    - LatLon
    - Vict Age
    - Vict Descent