CIS 4930

PREDICTIVE MODELING OF CRIME IN CHICAGO (2012-2017)

Jaime Mejia, Wil Sowersby, Glair Bronola, Vivian Ta

OVERVIEW

PROJECT OBJECTIVE

To explore patterns in crime in Chicago from 2012–2017 and build predictive models to

- Predict the type of crime
- Predict whether a crime will occur at a specific location
- Analyze trends in crime over time across different districts

Original Fields:

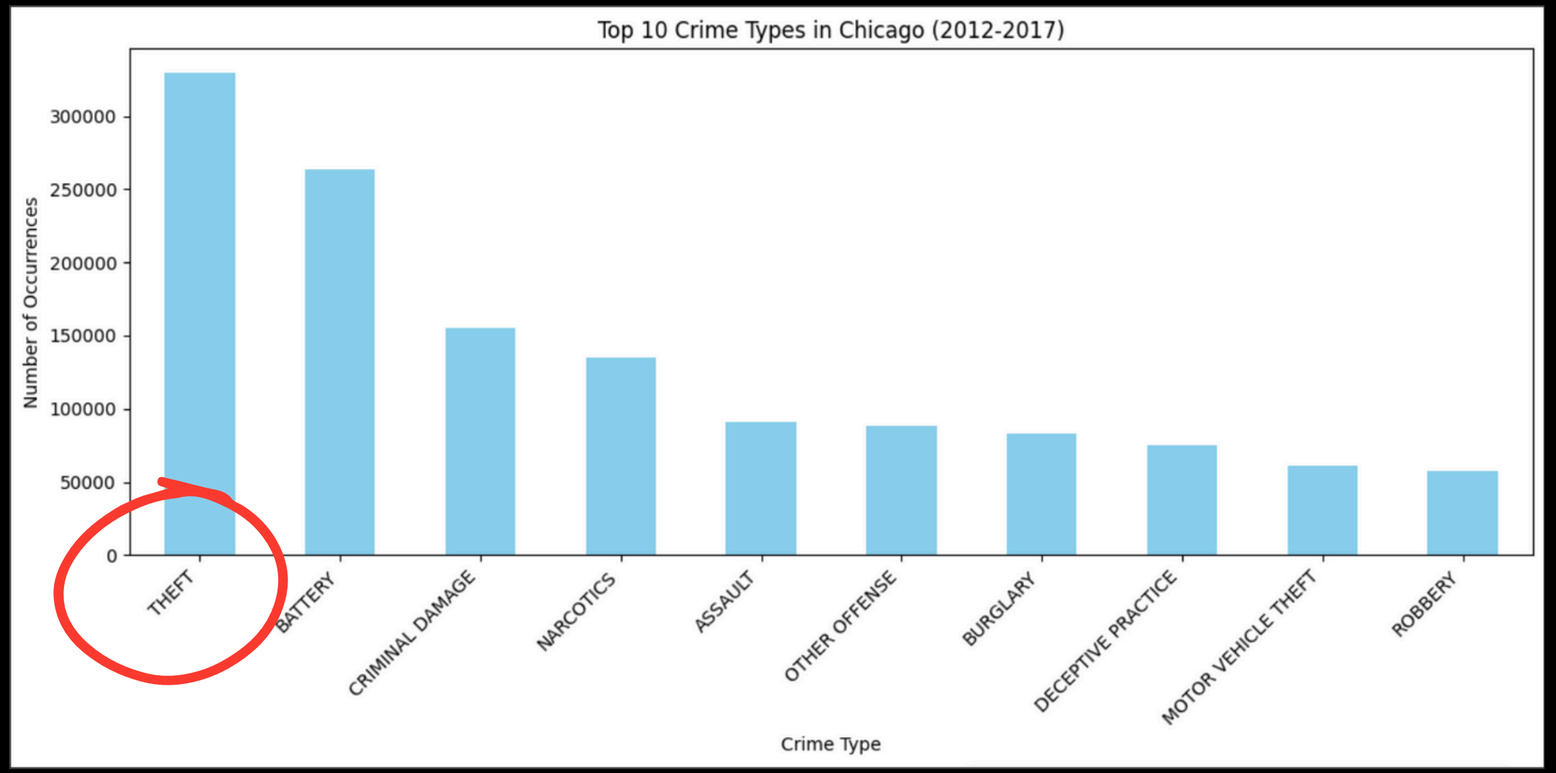
ID	0
Case_Number	1
Date	0
Block	0
IUCR	0
Primary_Type	0
Description	0
Location_Description	1658
Arrest	0
Domestic	0
Beat	0
District	1
Ward	14
Community_Area	40
FBI_Code	0
X_Coordinate	37083
Y_Coordinate	37083
Year	0
Updated_On	0
Latitude	37083
Longitude	37083
Location	37083

DATASET OVERVIEW

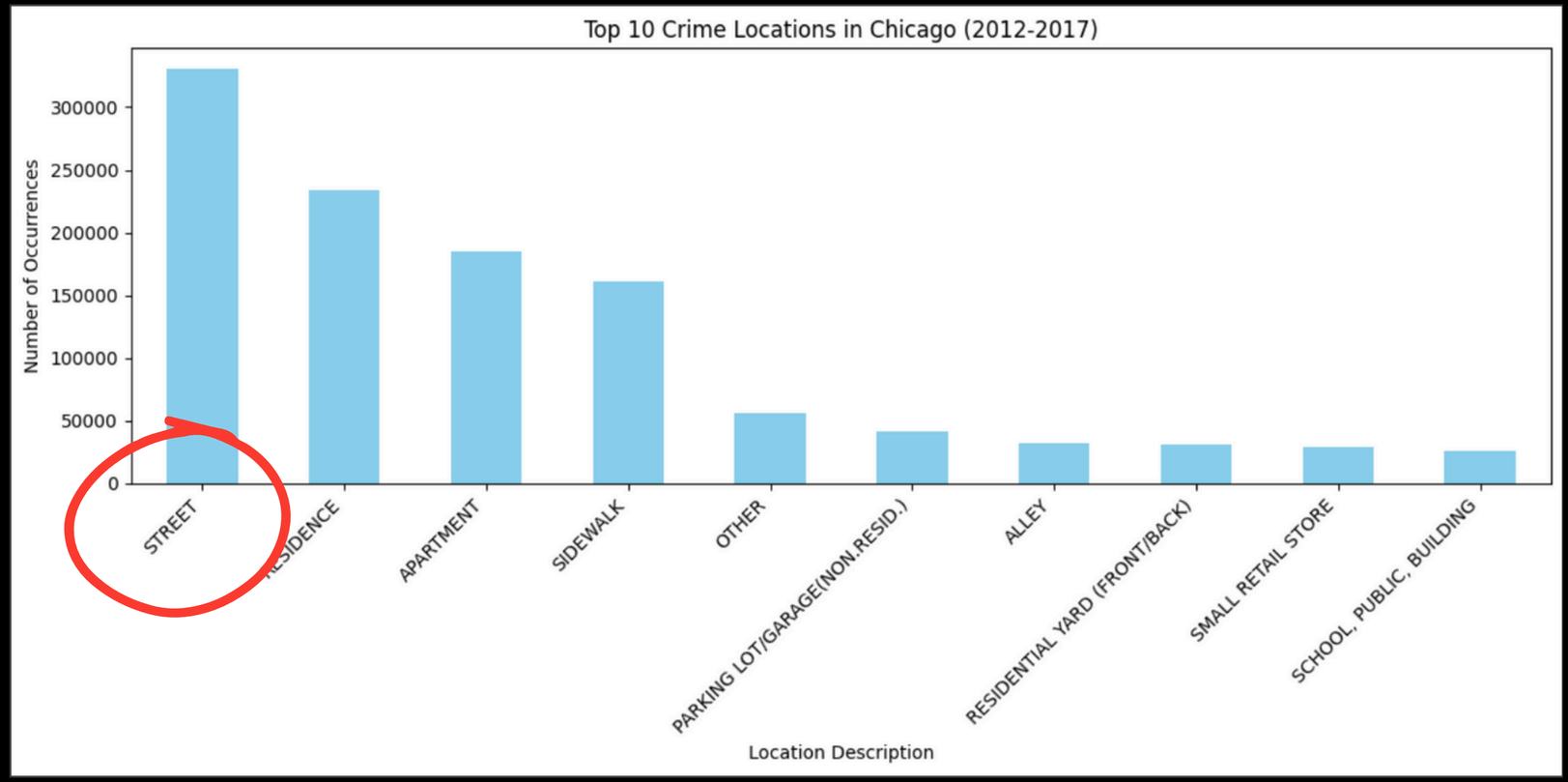
- **Dataset Name:** Crimes in Chicago
- **Source:** Kaggle Dataset by currie32
- **Original:** City of Chicago Police Department's CLEAR system
- **Timeframe:** 2001 - 2017
- **Data Volume:** Over 6 million records

STATISTICS AND VISUALS

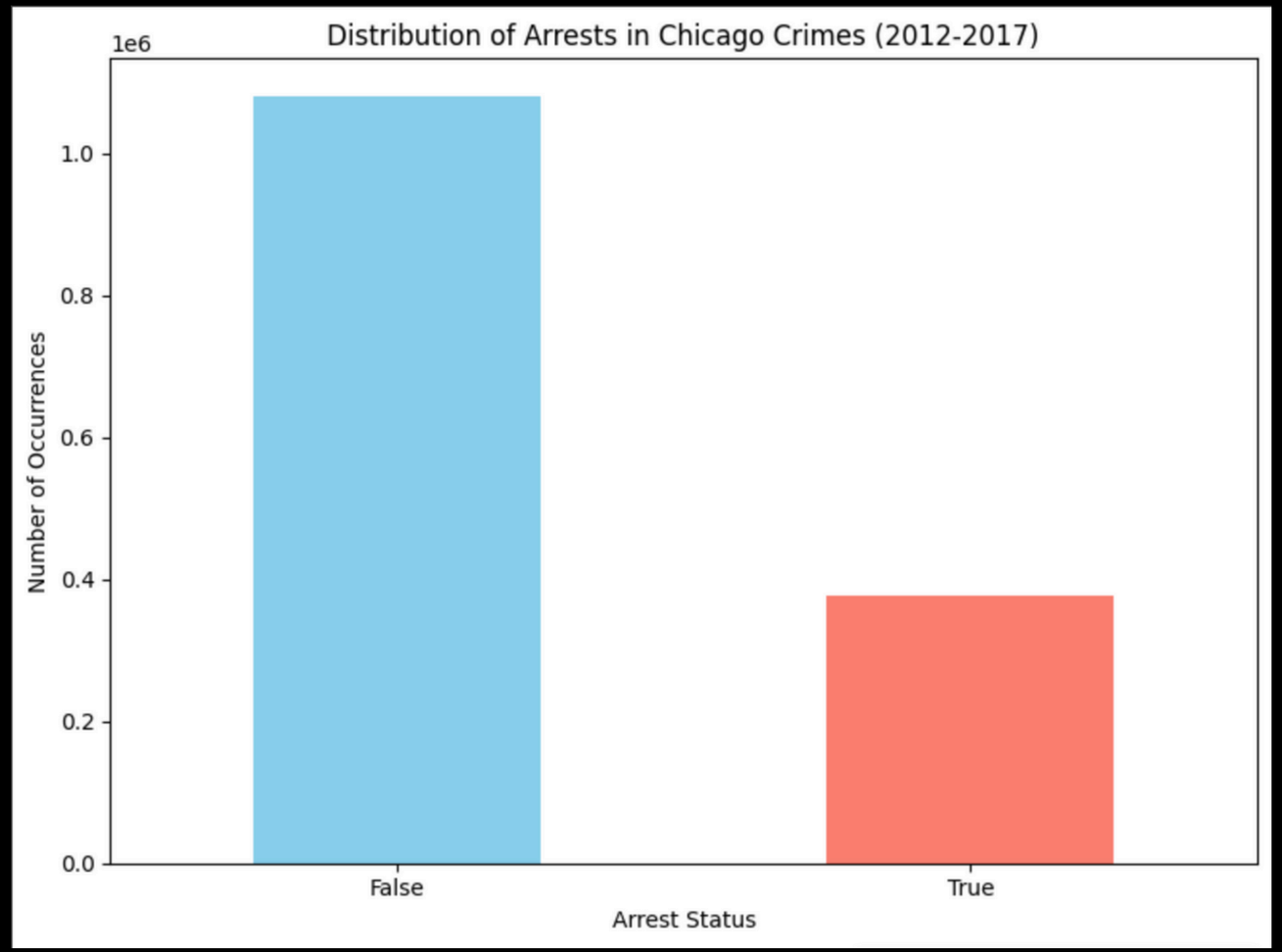
```
Code 0: ARSON
Code 1: ASSAULT
Code 2: BATTERY
Code 3: BURGLARY
Code 4: CONCEALED CARRY LICENSE VIOLATION
Code 5: CRIM SEXUAL ASSAULT
Code 6: CRIMINAL DAMAGE
Code 7: CRIMINAL TRESPASS
Code 8: DECEPTIVE PRACTICE
Code 9: GAMBLING
Code 10: HOMICIDE
Code 11: HUMAN TRAFFICKING
Code 12: INTERFERENCE WITH PUBLIC OFFICER
Code 13: INTIMIDATION
Code 14: KIDNAPPING
Code 15: LIQUOR LAW VIOLATION
Code 16: MOTOR VEHICLE THEFT
Code 17: NARCOTICS
Code 18: NON - CRIMINAL
Code 19: NON-CRIMINAL
Code 20: NON-CRIMINAL (SUBJECT SPECIFIED)
Code 21: OBSCENITY
Code 22: OFFENSE INVOLVING CHILDREN
Code 23: OTHER NARCOTIC VIOLATION
Code 24: OTHER OFFENSE
Code 25: PROSTITUTION
Code 26: PUBLIC INDECENCY
Code 27: PUBLIC PEACE VIOLATION
Code 28: ROBBERY
Code 29: SEX OFFENSE
Code 30: STALKING
Code 31: THEFT
Code 32: WEAPONS VIOLATION
```



STATISTICS AND VISUALS

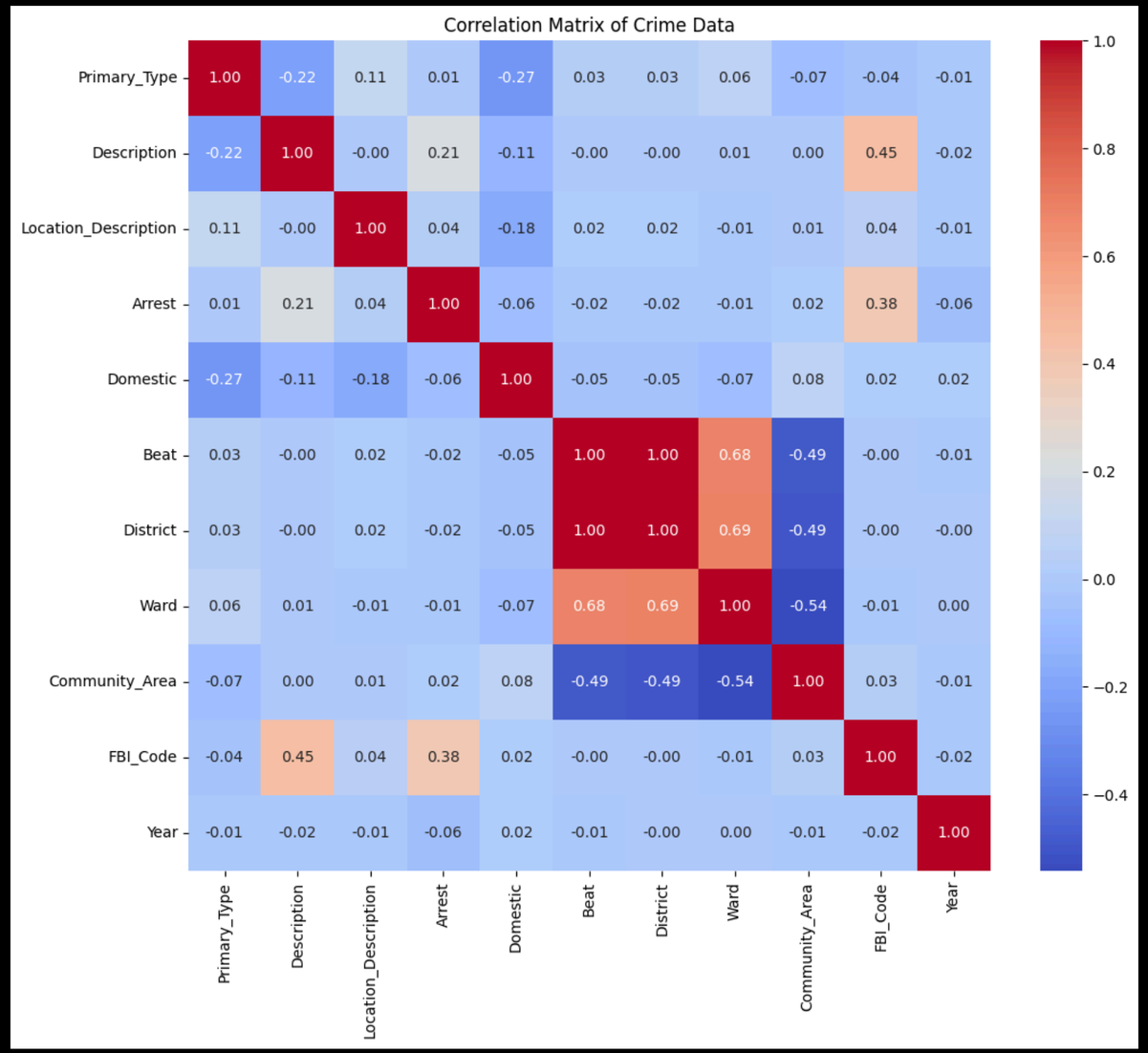


STATISTICS AND VISUALS



- **Most crimes did not result in an arrest**

CRIME DATA CORRELATION MATRIX



CLEANING AND PREPROCESSING

```
Correlation between Beat and District: 0.9973748525745405
```

almost identical

```
Correlation between District and Ward: 0.6869362672215436
```

moderate to high correlation

```
Correlation between Ward and Community_Area: -0.5420984002863685
```

Keeping highly correlated variables can lead to overfitting and confusion, ‘District’ was kept as it is the most meaningful variable

CLEANING AND PREPROCESSING

Initial Steps

- Loaded data
  - Focusing on crimes from 2012 to 2017
- Dropped irrelevant/redundant columns

Categorical to Numerical

- Converted ‘Primary_Type’, ‘Description’, and ‘Location_Description’ into categorical codes for modeling

Handling Missing Values

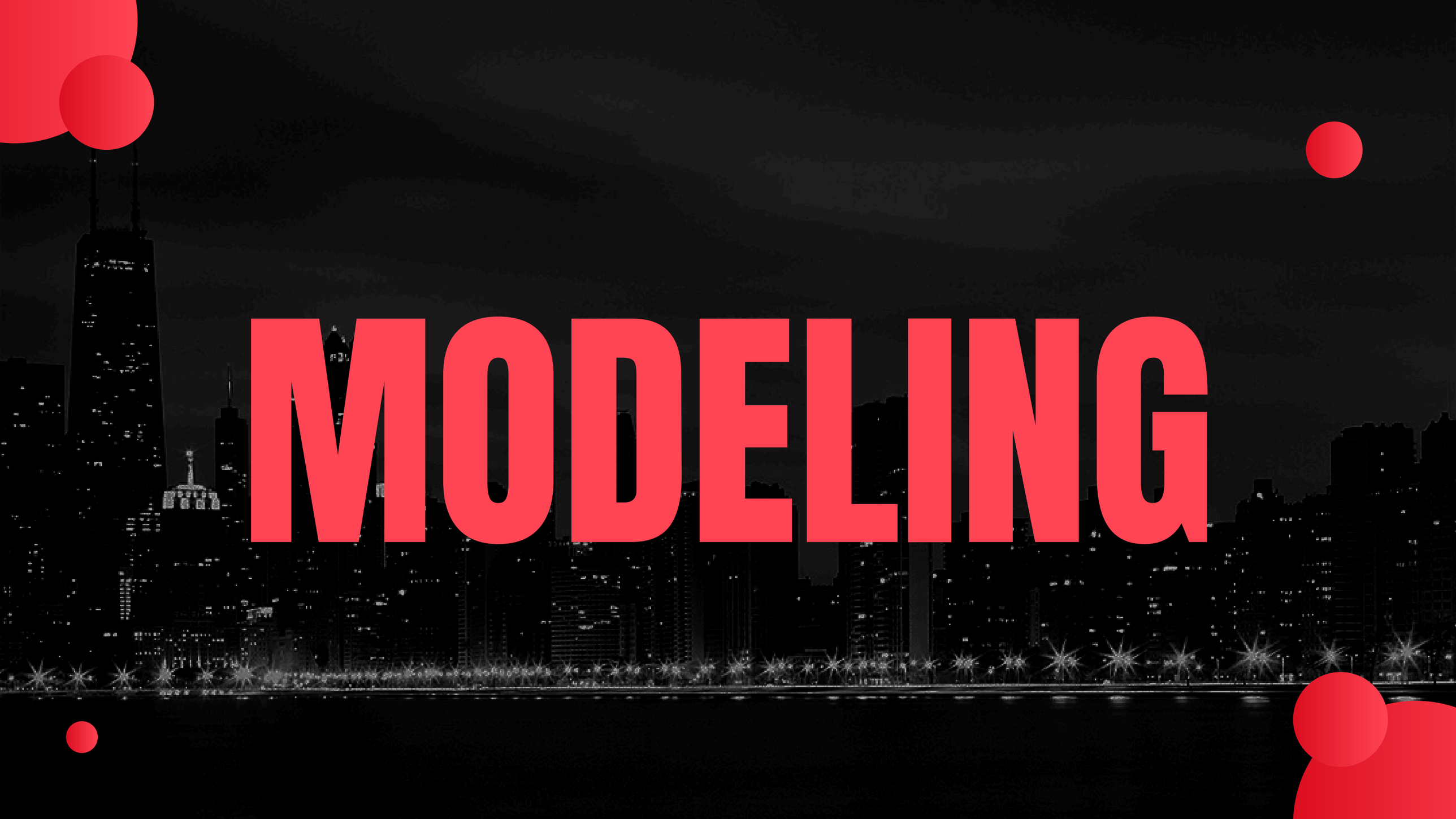
- Dropped rows with missing values

Correlation

- Identified highly correlated variables using correlation matrix
- Dropped: 'Beat', 'Ward', 'Community_Area' to reduce multicollinearity with ‘District’

FINAL VARIABLES FOR MODELING

	Primary_Type	Description	Location_Description	Arrest	Domestic	District	FBI_Code	Year
0	2	120	17	1	1	10.0	10	2016
1	2	120	111	0	1	3.0	10	2016
2	27	265	127	0	0	15.0	24	2016
3	2	283	123	0	0	15.0	10	2016
4	31	0	111	0	1	15.0	7	2016


MODELING

RANDOM FOREST

Confusion Matrix

	Predicted Negative (Non-Arrests)	Predicted Positive (Arrests)
Actual Negative (Non-Arrests)	210423	5156
Actual Positive (Arrests)	27520	48233

88.78%
accuracy

Correctly predicted

- 210,423 non-arrests
- 48,233 arrests

Errors

- 27,520 false negatives
- 5,156 false positives

DECISION TREE

Confusion Matrix

	Predicted Negative (Non-Arrests)	Predicted Positive (Arrests)
Actual Negative (Non-Arrests)	208894	6685
Actual Positive (Arrests)	27380	48373

88.31%
accuracy

Correctly predicted

- 208,894 non-arrests
- 48,373 arrests

Errors

- 27,380 false negatives
- 6,685 false positives

LOGISTIC REGRESSION

Confusion Matrix

	Predicted Negative (Non-Arrests)	Predicted Positive (Arrests)
Actual Negative (Non-Arrests)	198700	16879
Actual Positive (Arrests)	58412	17341

74.16%

accuracy

Correctly predicted

- 198,700 non-arrests
- 17,341 arrests

Errors

- 58,412 false negatives
- 16,879 false positives

NEURAL NETWORK

Confusion Matrix

	Predicted Negative (Non-Arrests)	Predicted Positive (Arrests)
Actual Negative (Non-Arrests)	210916	4663
Actual Positive (Arrests)	40477	35276

84.51%
accuracy

Correctly predicted

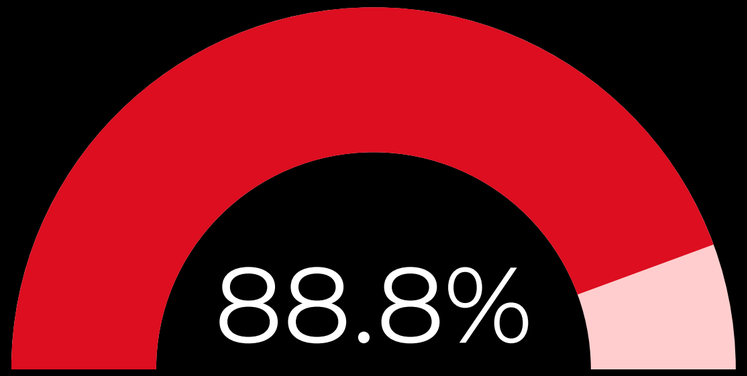
- 210,916 non-arrests
- 35,276 arrests

Errors

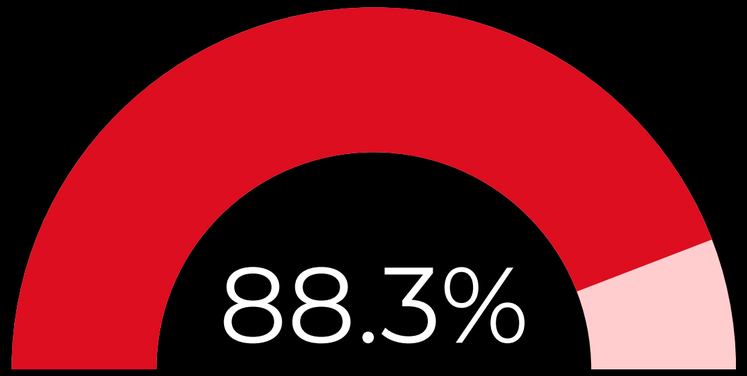
- 40,477 false negatives
- 4,663 false positives

ACCURACY

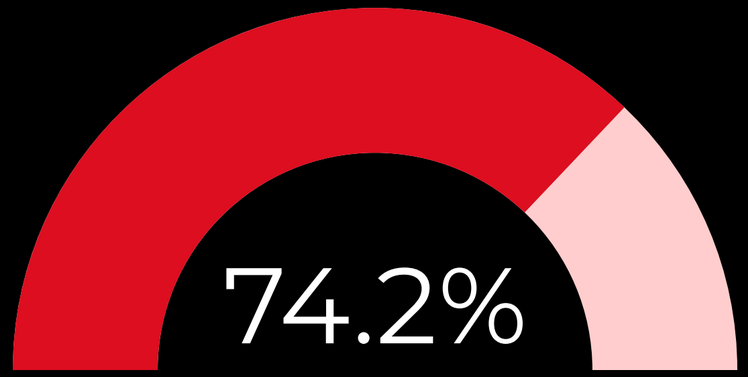
Random Forest



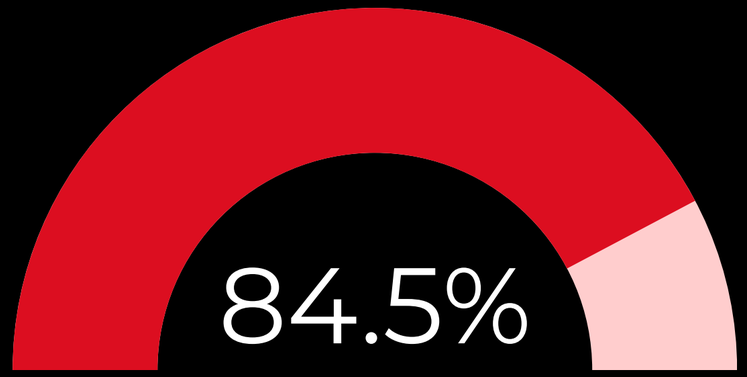
Decision Tree



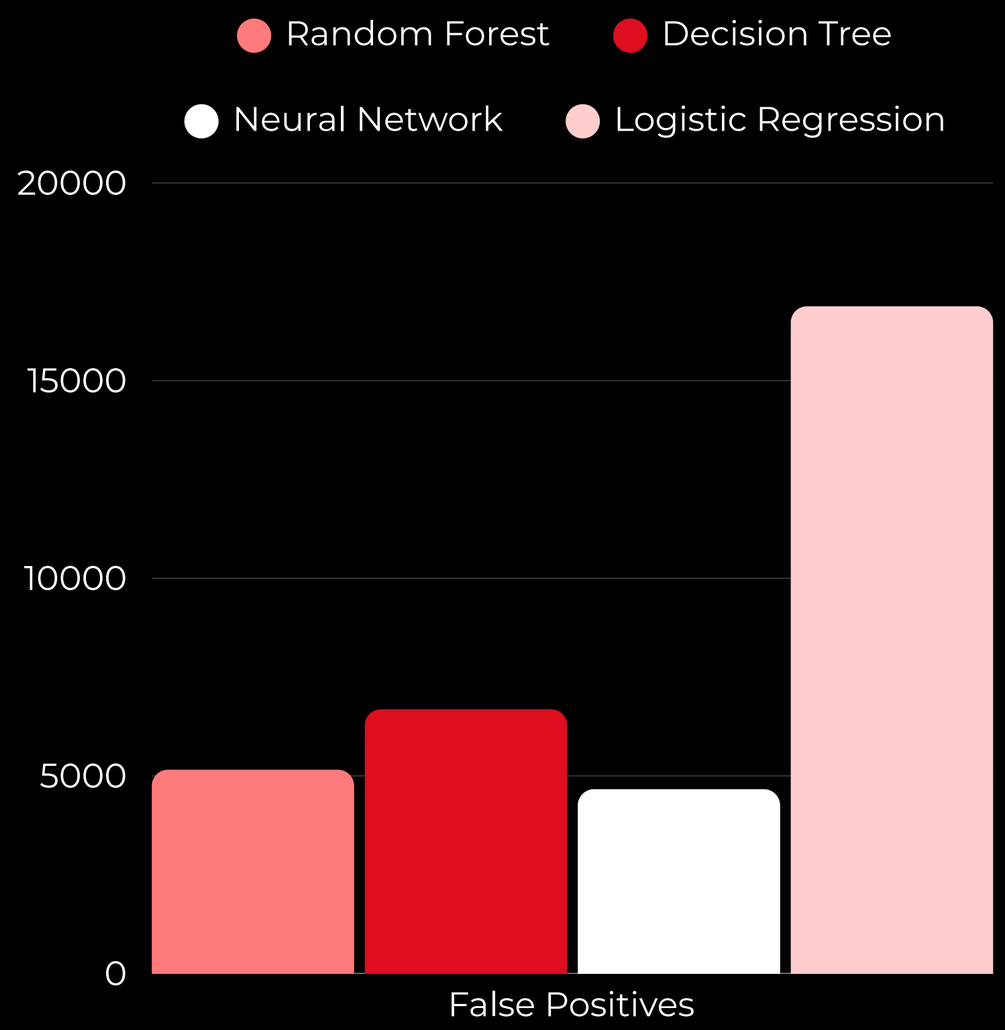
Logisitic Regression

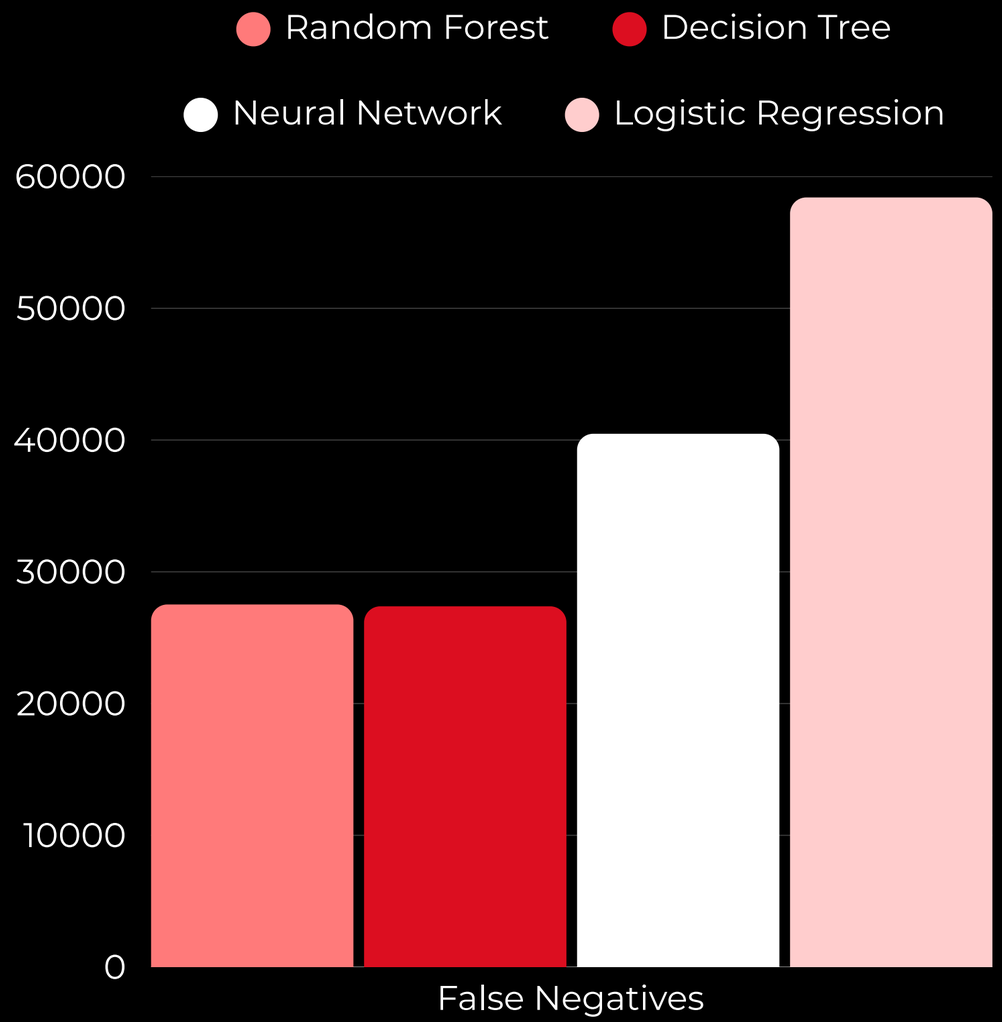


Neural Network



DATA VISUALIZATION





Best Model:

RANDOM FOREST MODEL

- Highest accuracy: 88.78%
- Strongest balance between false positives and false negatives
- Generalizes well
  - Consistent accuracy across various crime types, not just the most frequent ones

REFLECTIONS + LIMITATIONS

Reflections

- Preprocessing data is essential for optimal performance
- Evaluating different models is important for comparing accuracy and complexity
- Narrowing down variables helped reduce multicollinearity, making the model more efficient

Limitations

- Some crime types show up frequently, which can cause bias in the model
- Due to privacy, location precision is limited
- Data focuses only on 2012 - 2017, which patterns may have changed since

CONCLUSION

Key Takeaways

- Processed over 6 million rows of crime data (2012–2017)
- Encoded high-cardinality variables with .cat.codes; while this adds ordinality, Random Forest and Decision Tree models handled it well
- Dropped highly correlated features (Beat, Ward, Community_Area) to reduce redundancy
- Trained 4 models: Random Forest performed best (88.78% accuracy) and generalized well across crime types
- Preprocessing data had a major impact on results

CIS 4930

THANK YOU

2025 Machine Learning Presentation