Predicting Chicago Crime

by **Paul Schmidt**

Outlining the Presentation

- I. Introduction to Chicago Crime
    - A. Context on Chicago Crime
    - B. The Research Question
    - C. Method of Testing
- II. Preliminary Analysis of Datasets
    - A. Analyzing Crime Characteristics
    - B. Analyzing Geography of Crime
    - C. Analyzing Weather’s Effect on Crime
- III. Crafting a Solution
    - A. Model Creation and Tuning
    - B. Testing on Holdout
- IV. Model Relevance

I. Introduction to Chicago Crime

II. Analysis of Data

III. Crafting a Solution

IV. Model Relevance

I. Introduction to Chicago Crime

I. Introduction to Chicago Crime

Context on Chicago Crime

- The City of Chicago
  - Third largest city in the US behind New York and Los Angeles
  - Currently, the population is 2.679 million

- Crime in Chicago
  - Since the late 1800s, Chicago has been infamously known for its crime rates
  - Crime tends to ebb and flow based on certain conditions (most notable, weather and geography)

I. Introduction to Chicago Crime

What Is the Need?

- The Need
    - Crime affects various aspects of life for the citizen:
        - Peace of mind
        - Commuting
        - Location of living
        - Education
        - Life expectancy
    - Crime also affects various aspects of the police force:
        - Staffing size
        - Training emphasis
        - Geographic location

I. Introduction to Chicago Crime

What Is Its Solution?

- The Solution

While crime is not limited to a zip code, it does have recognizable and predictable patterns.

  - For the citizen, these patterns empower by:
    - Informing of dangerous times and places
    - Assisting in the decision of living location
  - For the police force, these patterns empower by:
    - Knowing how to staff their teams (time and location)
    - Knowing what training to emphasize for the task force

- The Goal: reliably predict crime *locations* and *times*
  - Predefined accuracy threshold: 85%

I. Introduction to Chicago Crime

II. Analysis of Data

III. Crafting a Solution

IV. Model Relevance

II. Analysis of Data

II. Analysis of Data

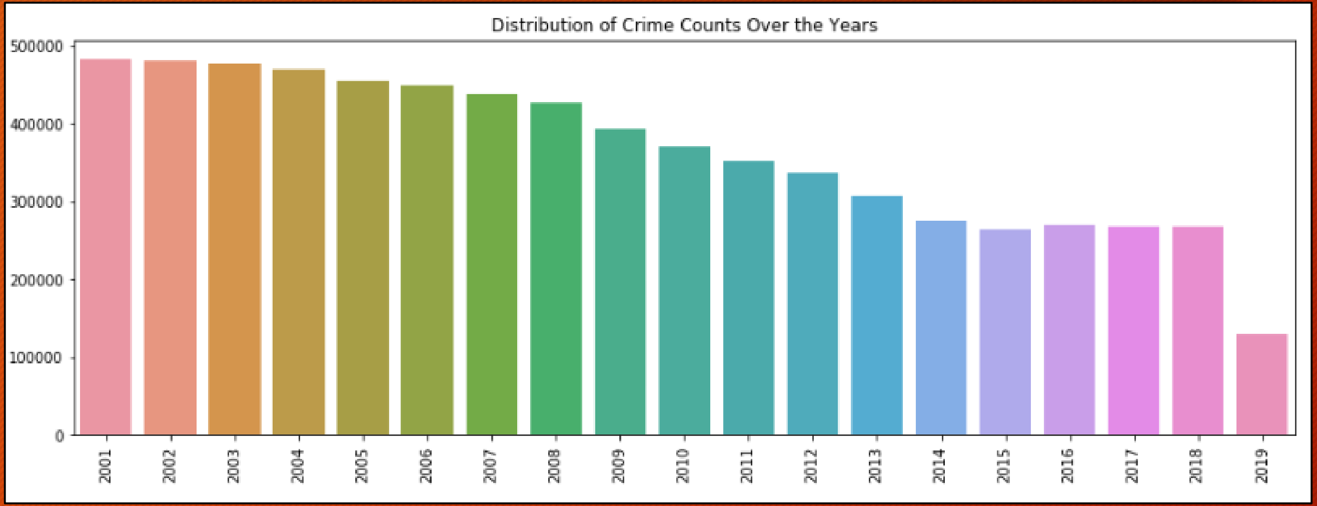
The Datasets

- Crime Dataset: *CLEAR* (Citizen Law Enforcement Analysis and Reporting)
  - Extracted directly from the Chicago Police Department (CPD)
  - Spans from 2001 to the present with over 6.5 million datapoints
  - Contains a variety of features like type of crime and location
- Weather Dataset: Chicago O'Hare Station within *Illinois ASOS Network*
  - Made available through Iowa State University
  - Spans from before 2001 to the present
  - Populated with one to two entries per hour
  - Contains variety of continuous and categorical features such as real-feel temperature and precipitation recorded

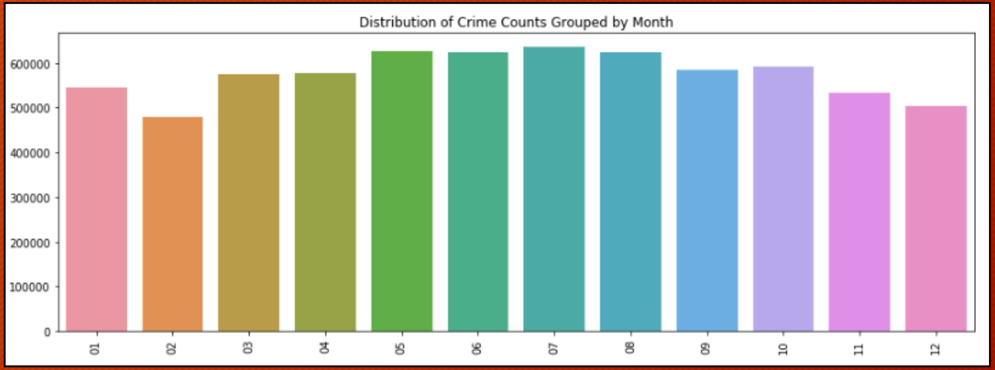
II. Analysis of Data

Characteristics of Crime (A)

- Crime rates over the years
  - Declined from 2001 to 2013
  - Plateaued from 2014 to present
  - Incomplete data in 2019



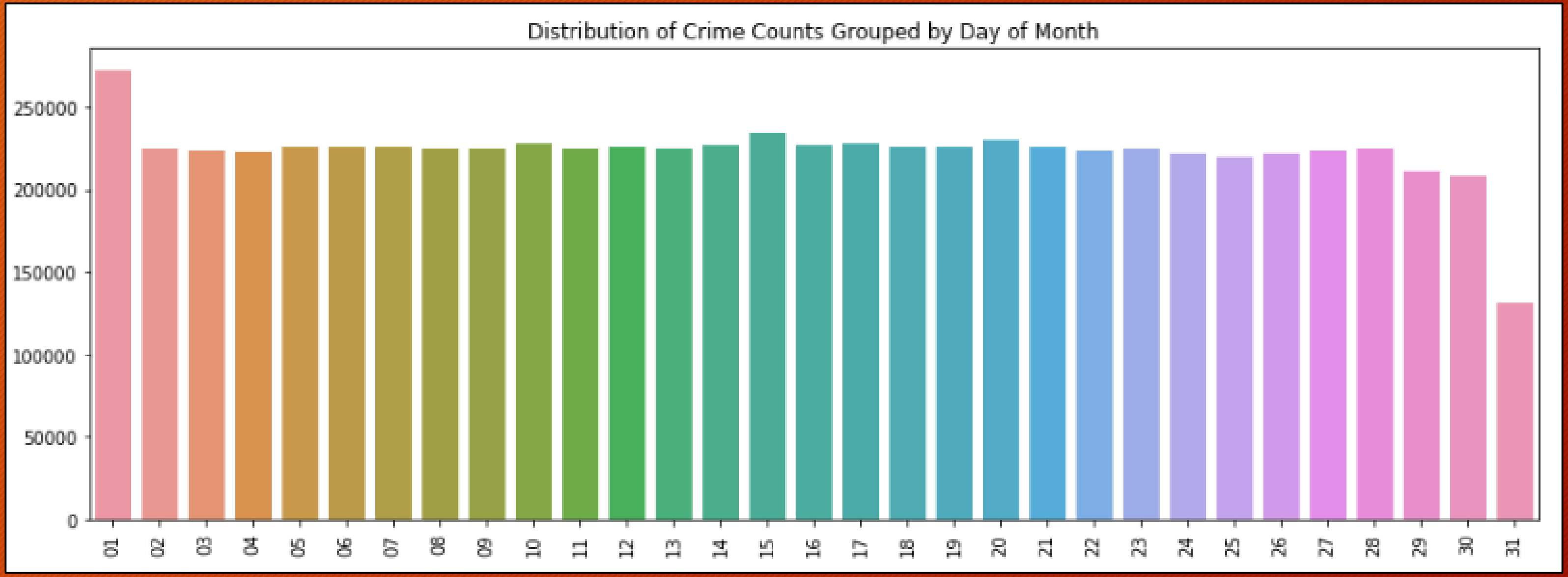
- Crime distribution across months
  - Spikes in crime during summer months
  - Decrease is strong winter months like February



II. Analysis of Data

Characteristics of Crime (B)

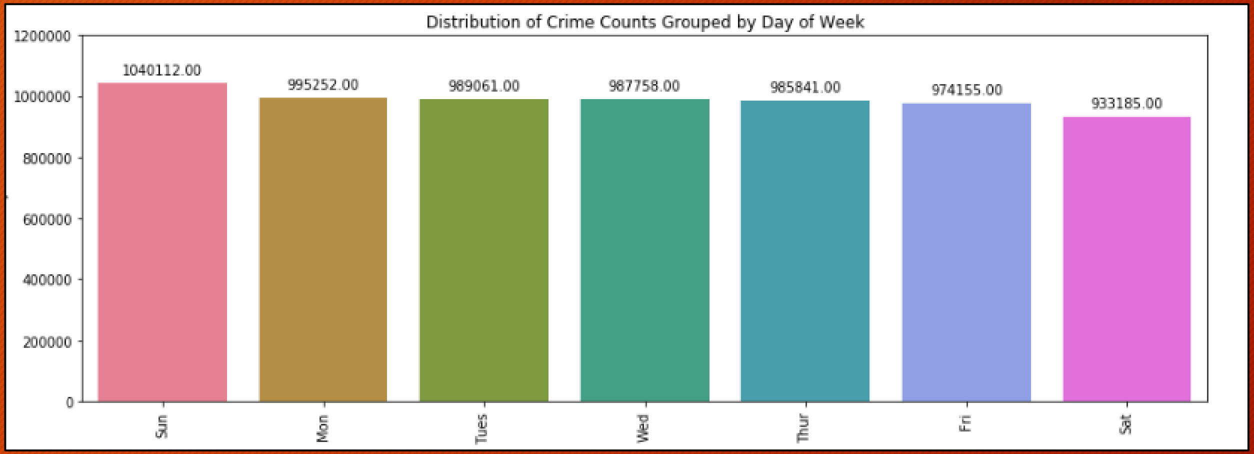
- Crime rates across day of month
  - Decrease in activity on 29th, 30th, and 31st day of month
  - Increase on 1st day of month
  - Slight increase on 15th day of month
    - After assessing the normal distributions of the 15th day against the distributions of all other days, the increase in the 15th day is indeed statistically significant (containing a t-value of 2.527 and p-value of 1.15%)



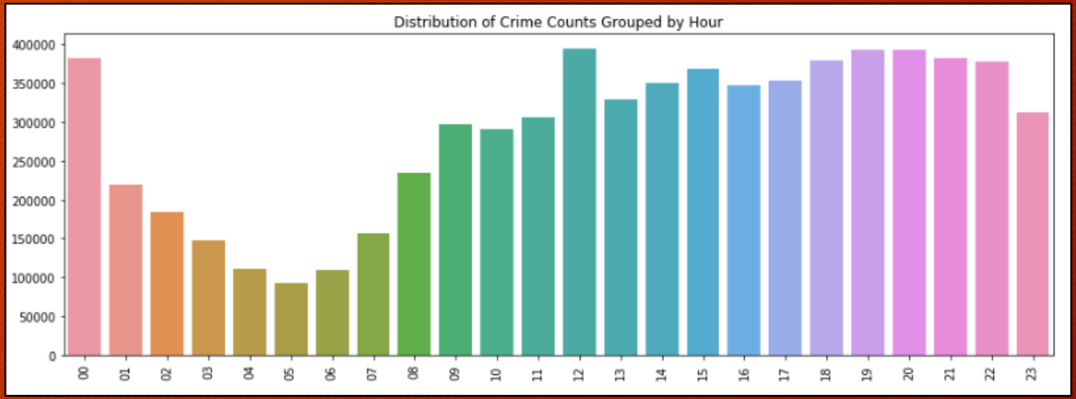
II. Analysis of Data

Characteristics of Crime (C)

- Crime rates across day of week
  - Polarity in weekends
    - Decrease in activity on Saturdays
    - Increase on Sundays
  - Steady throughout the weekdays



- Crime rates across hours of day
  - Greater activity in hours after noon
    - Peaks are at noon and 8:00pm
  - Low activity during the early mornings
    - Lowest rate at 5:00am

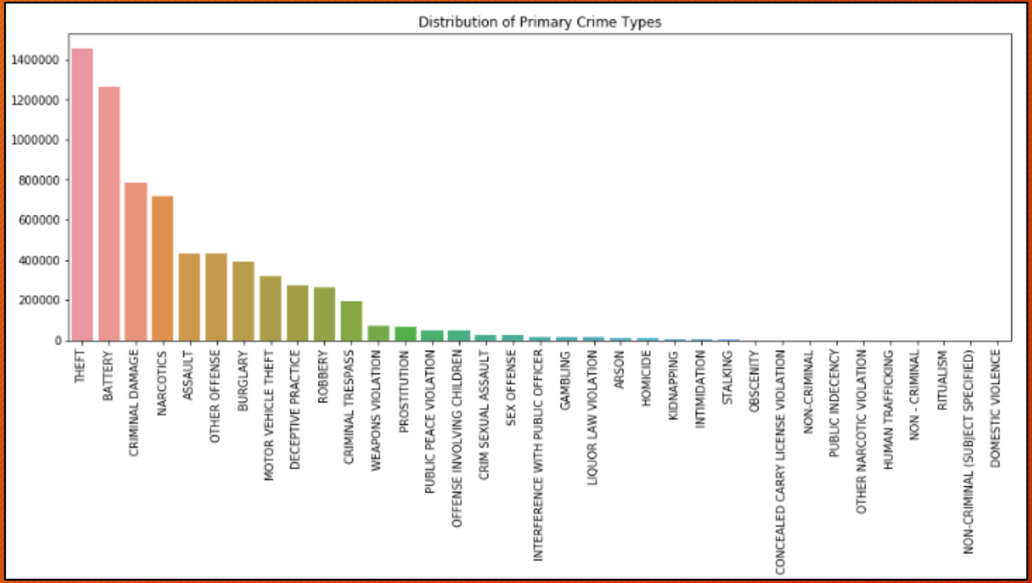


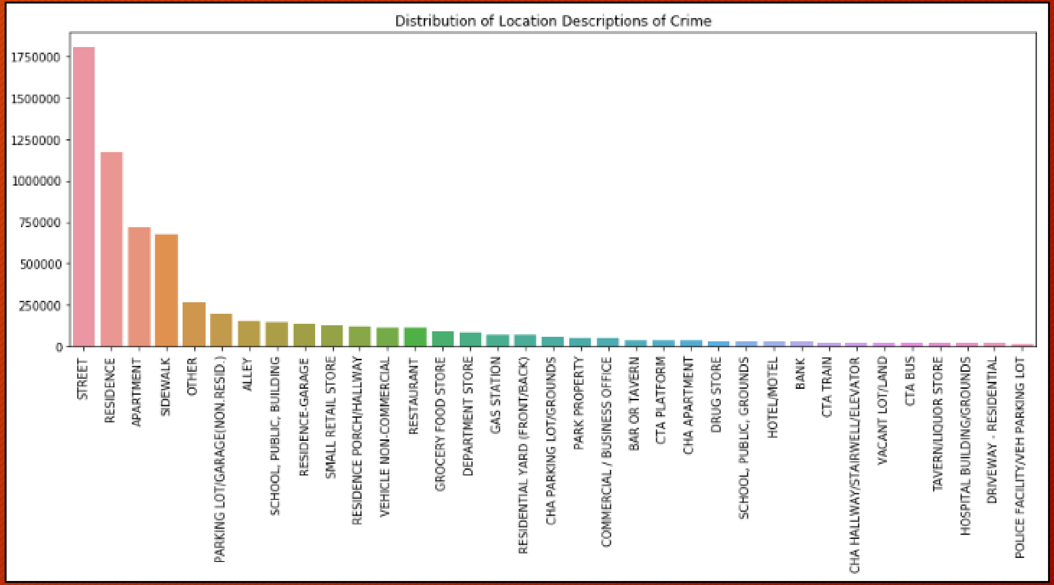
II. Analysis of Data

Characteristics of Crime (D)

- Top 5 Primary Crime Types:
  1. Theft
  2. Battery
  3. Criminal Damage
  4. Narcotics
  5. Assault

- Top 5 Crime Activity Areas:
  1. Street
  2. Residence
  3. Apartment
  4. Sidewalk
  5. Other

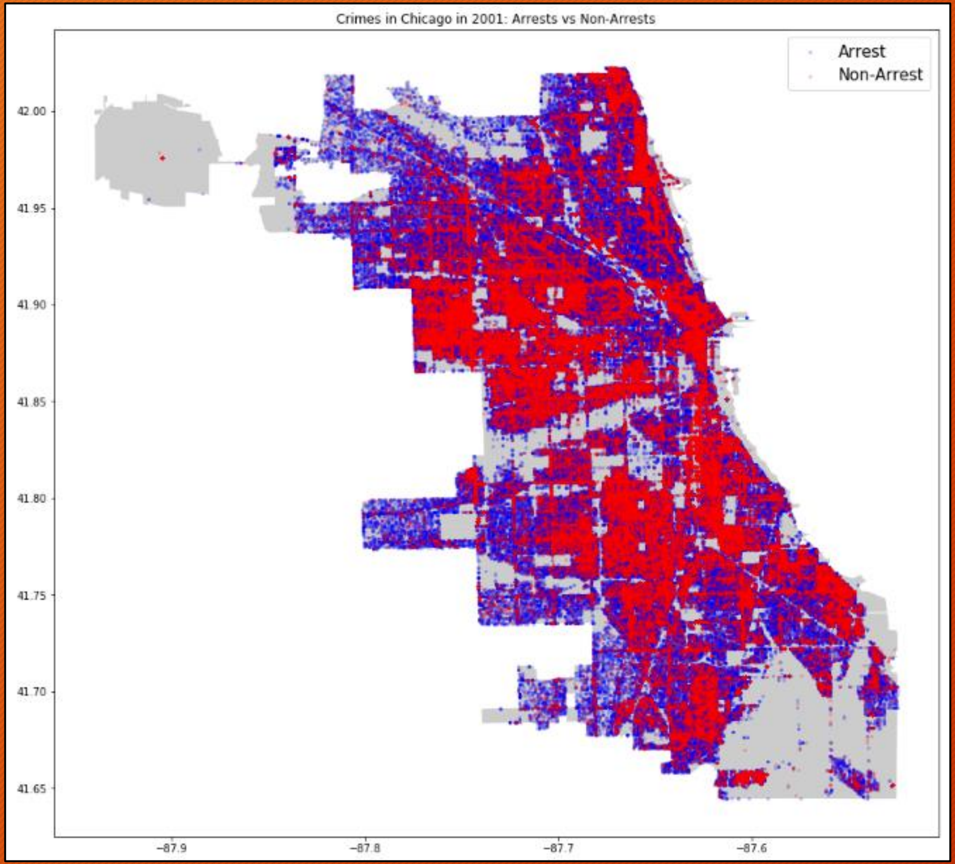




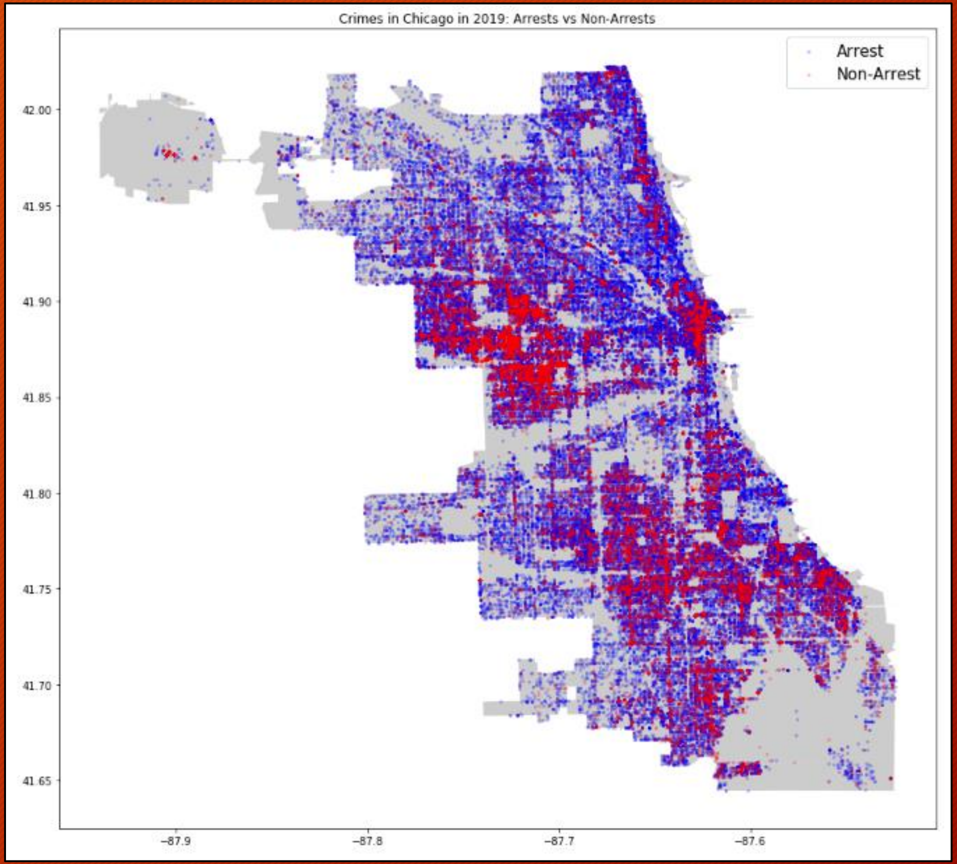
II. Analysis of Data

Geography of Crime (A)

- Visualizing the geographical distribution of crime between 2001 and 2019



Crimes in Chicago in 2001: Arrests vs Non-Arrests

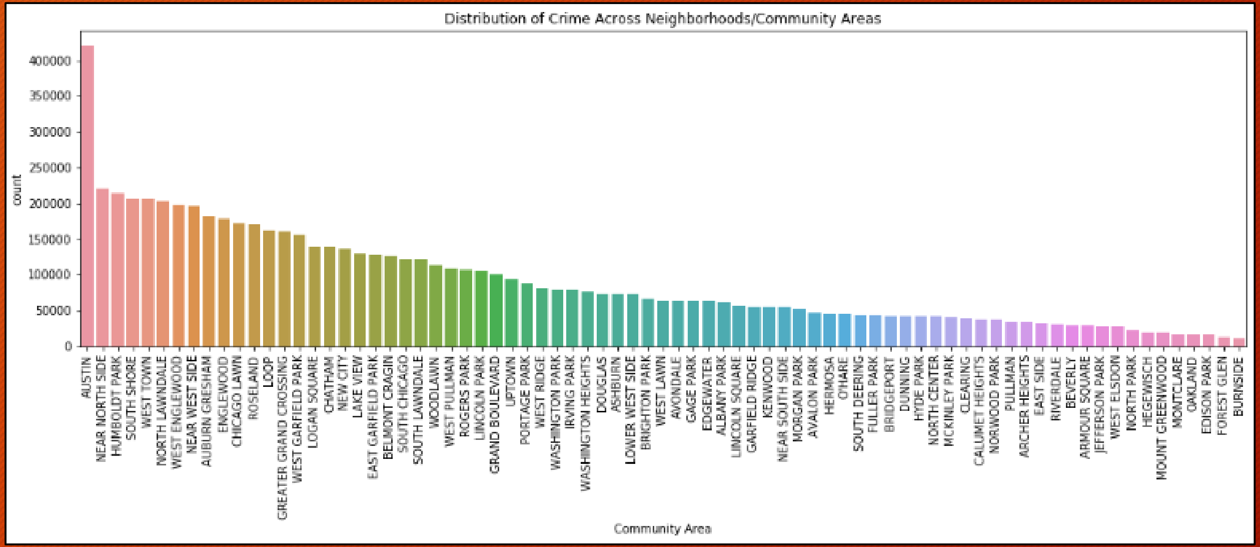


Crimes in Chicago in 2019: Arrests vs Non-Arrests

II. Analysis of Data

Geography of Crime (B)

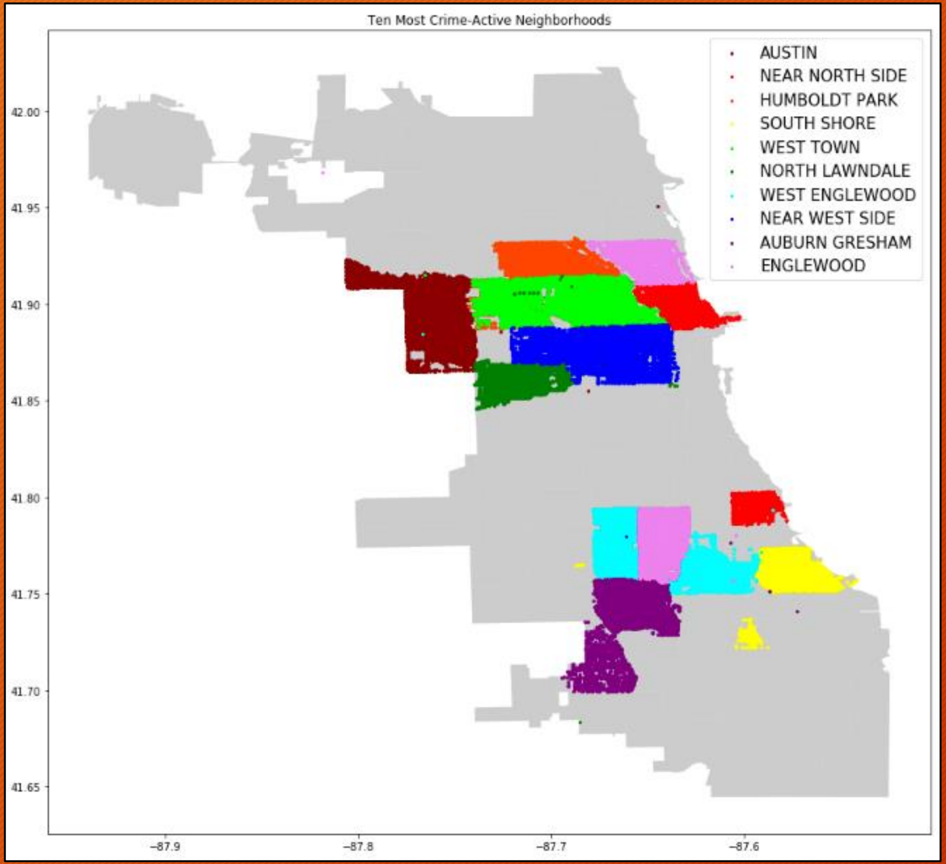
- Crime across Neighborhoods:
    - Austin is considerably more active than other neighborhoods
    - Steady decrease throughout remainder

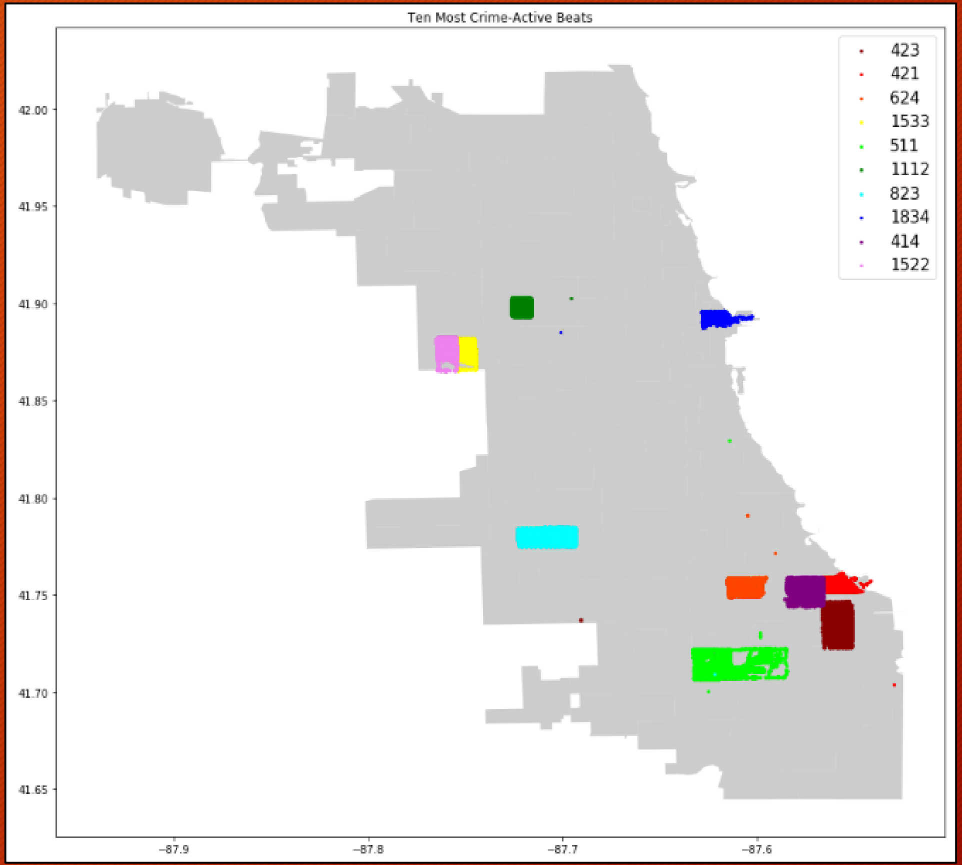


II. Analysis of Data

Geography of Crime (C)

- Visualizing most active sub-locations within Chicago

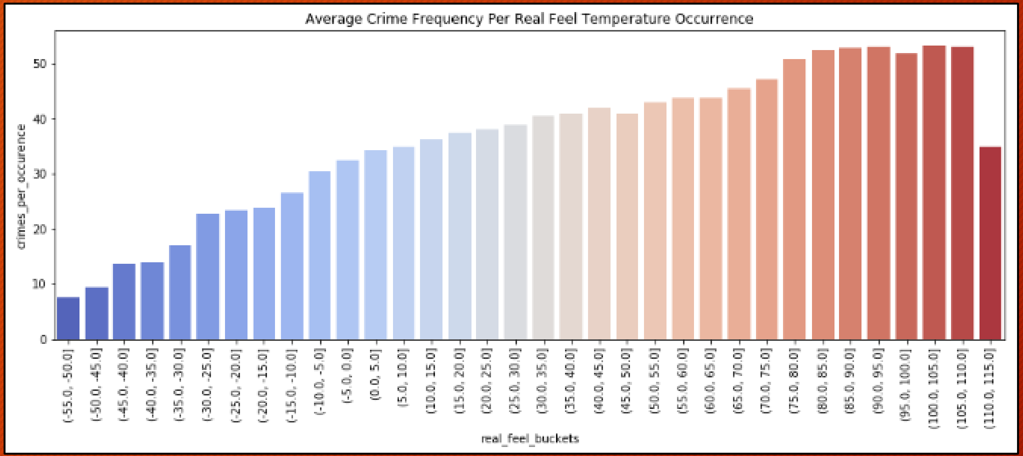


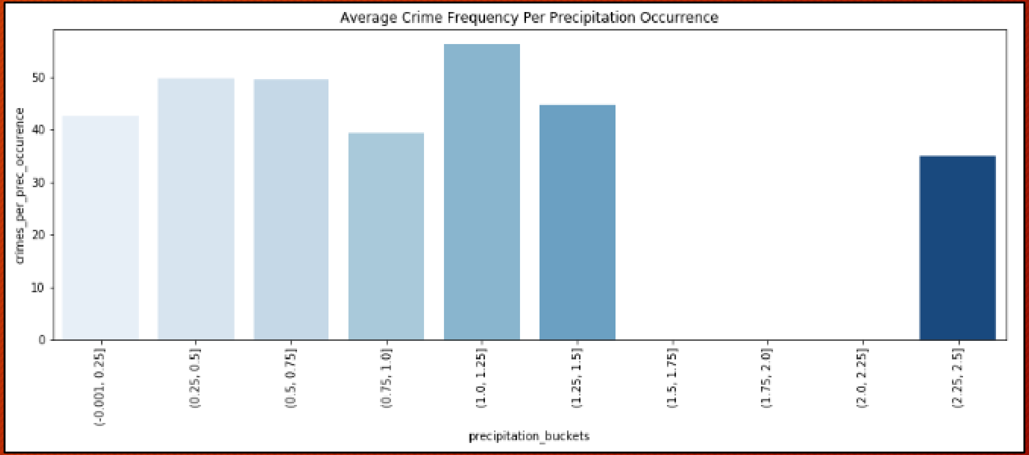


II. Analysis of Data

Weather’s Effect on Crime

- Temperature appears to have a strong causal relationship with frequency of crime
  - For a given hour between -55 and -50 degrees Fahrenheit: 7.5 crimes will occur
  - For a given hour between 100 and 105 degrees Fahrenheit: 53.3 crimes will occur
- The correlation is not nearly as evident when considering precipitation
  - Surprisingly, maximum crime frequency takes place during hours with 1” and 1.25” inches of precipitation
  - Unsurprisingly, minimum crime occurrence is when precipitation is heaviest


Average Crime Frequency Per Real Feel Temperature Occurrence


Average Crime Frequency Per Precipitation Occurrence

I. Introduction to Chicago Crime

II. Analysis of Data

III. Crafting a Solution

IV. Model Relevance

III. Crafting a Solution

III. Crafting a Solution

Overview of Models

The two model types consist of predicting daily crime for:

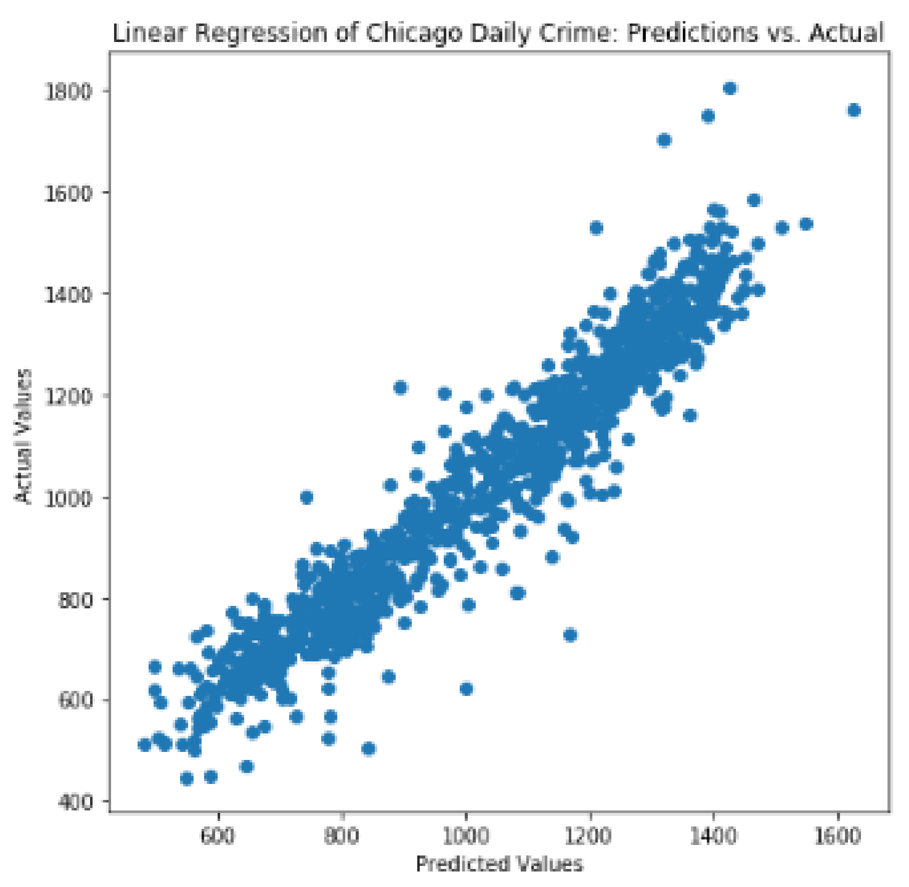
1. The **city** as a whole
2. The individual 77 **neighborhoods**

III. Crafting a Solution

Predicting Daily Crime in Chicago (A)

- Preliminary model scores for the **City**:

Linear:	91.55% ★
Lasso:	90.63%
Ridge:	91.55%
Random Forest:	90.23%
Neural Network:	67.82%



III. Crafting a Solution

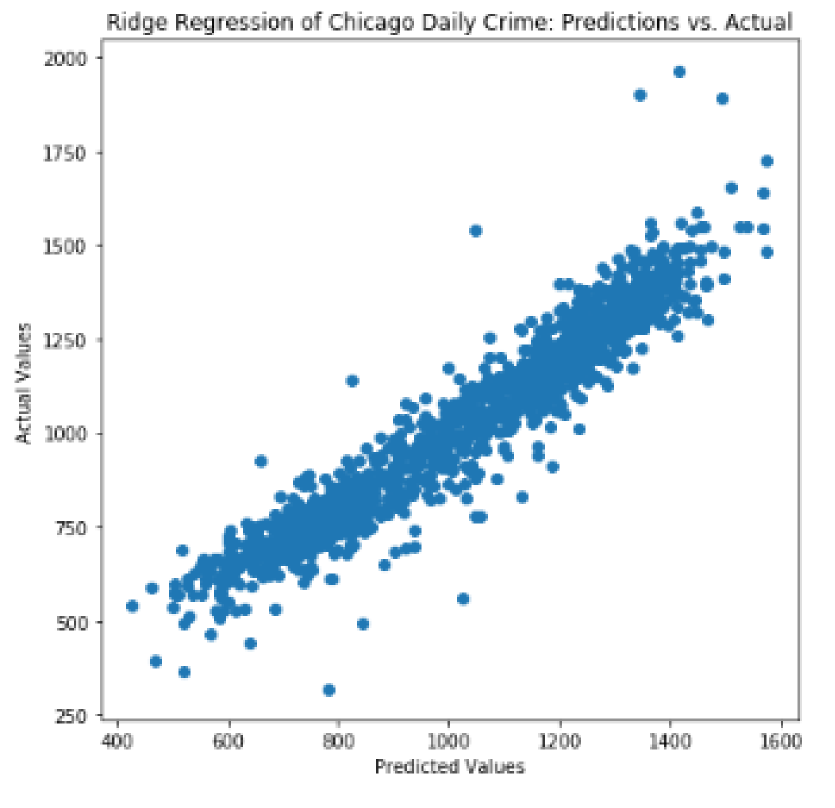
Predicting Daily Crime in Chicago (B)

- After assessing, linear models with corresponding regularizations were chosen for model tuning due to:
  - Computational efficiency
  - Reliable accuracy
  - Communicable reasoning
- After testing Linear, Lasso, and Ridge models, the tuned Ridge Regression model (on the training data) was the most accurate:
  - Original Ridge: 91.55%
    - Hyperparameter default of: “0.001”
  - Tuned Ridge: 94.16%
    - Hyperparameter alpha of: “0.01”

III. Crafting a Solution

Predicting Daily Crime in Chicago (C)

- Testing Ridge model on Holdout:
  - Accuracy: 91.47%
  - Mean Absolute Error: 54.63
- Polarizing Coefficients:
  - Positive correlations:
    - Years 2001, 2003, 2002, 2004
    - Dew_pt_temp
    - First day of month
  - Negative correlations:
    - Years 2019, 2015, 2017, and 2016
    - Precipitation_by_hour
    - Rel_humidity

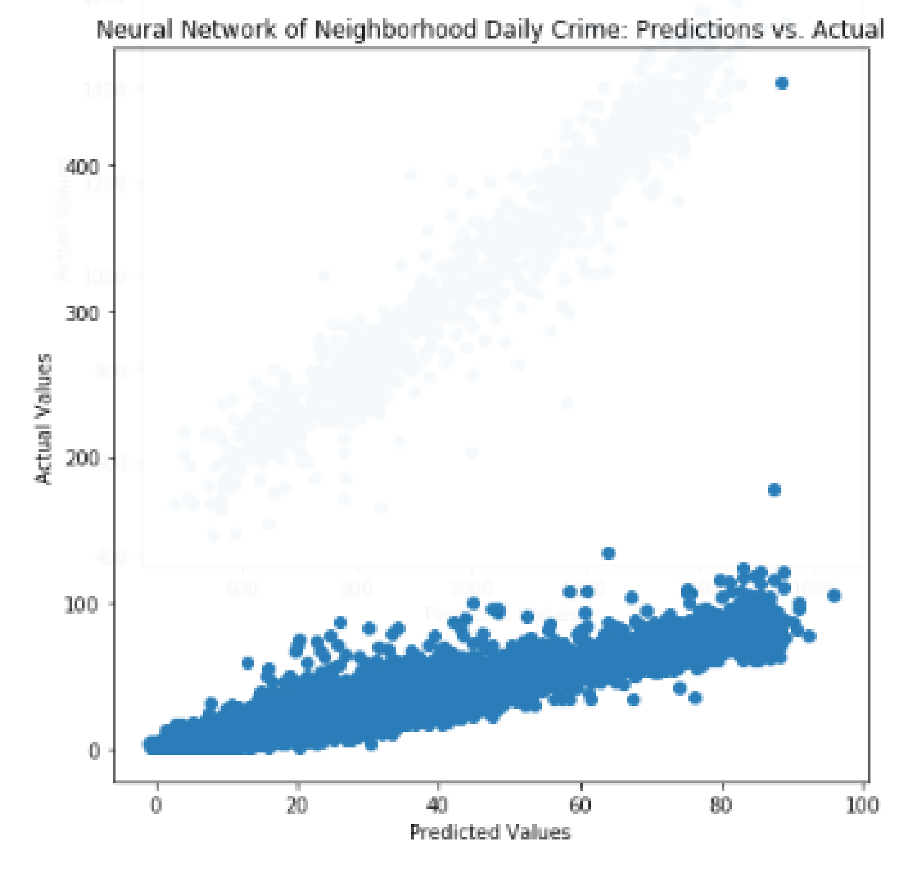


III. Crafting a Solution

Predicting Daily Crime in Neighborhoods (A)

- Preliminary model scores for Specified **Neighborhoods:**

Linear:	73.53%
Lasso:	03.10%
Ridge:	73.53%
SVR:	72.66%
Random Forest:	84.93%
Neural Network:	86.47% ★



III. Crafting a Solution

Predicting Daily Crime in Neighborhoods (B)

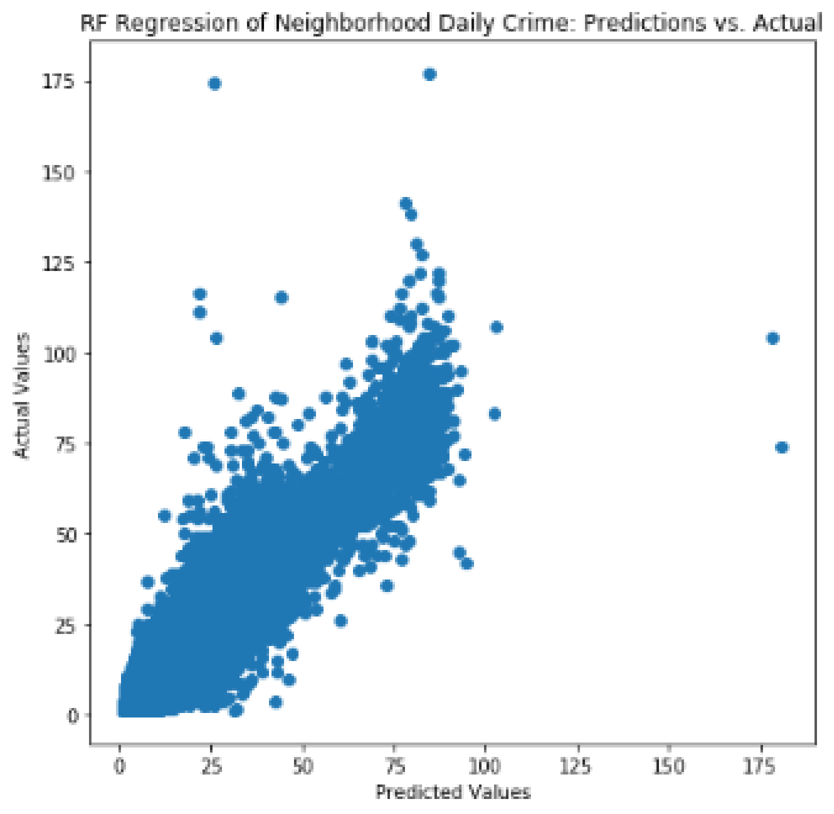
- After assessing, deep learning models were chosen for model tuning:
  - Pro: Reliable accuracy
  - Con: Computationally demanding

- After tuning RFR and MLP on samples of data, MLP seems to provide greater accuracy:
  - RFR accuracy improvement: 1.78% (from 84.93%)
    - Hyperparameters: *{'n_estimators': 800, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 100, 'bootstrap': True}*
  - MLP accuracy improvement: 7.78% (from 86.47%)
    - Hyperparameters: *{'alpha': 0.0001, 'hidden_layer_sizes': (100, 50)}*

III. Crafting a Solution

Predicting Daily Crime in Neighborhoods (C)

- Testing RFR model on Holdout:
  - Accuracy: 86.88%
- Observations:
  - When narrowing geographical influence, there is considerably greater disparity in predictions
  - As mentioned, this model is incredibly computationally demanding:
    - Tuning on sample
    - Implementing on holdout



I. Introduction to Chicago Crime

II. Analysis of Data

III. Crafting a Solution

IV. Model Relevance

IV. Model Relevance

IV. Model Relevance

Predicting Daily Crime in City (A)

Ridge Regression

- Why Linear?
  - Surprisingly, when observations are restricted to a consistent and large location, crime activity is very predictable
    - Follows normal distributions
    - Consistent patterns derived from relatively few features

- Why Ridge?
  - This model outperformed other models while providing interpretable reasonings for its results
    - Provides direct insights of feature importance on final prediction
    - L2 Regularization penalizes high correlating features in order to avoid overfitting (benefit over simple Multi-Variable Linear Regression)

IV. Model Relevance

Predicting Daily Crime in City (B)

Ridge Regression

- How does the model benefit?
  - The benefit of targeting daily crime activity assists the Chicago Police Department by helping:
    - Calculate daily staffing based on daily conditions
    - Prioritize training emphasis based on primary types of crime recorded
    - Forecast probable events throughout the day based on given daily conditions

- Thinking Ahead?
  - To run in a production environment, the model may be adjusted to automatically run on monthly, weekly, and daily bases in order to inform officers on the expected crime rates and types for the predefined time increment
  - Because variance in crime activity is low across the years, the script should not require frequent maintenance
    - However, in order to optimize results, accuracy can be tuned on an annual basis to ensure consistency and relevance

IV. Model Relevance

Predicting Daily Crime in Neighborhoods (A)

Random Forest Regressor

- Why Ensemble Modeling?
  - Unsurprisingly, when creating a single model to predict all activity within smaller geographical sections of Chicago, activity is no longer simple and linear.
    - Different ethnic/economic patterns of neighborhoods contribute in more nuanced ways than a simple coefficient
    - Factors that affect one neighborhood may not affect another neighborhood in the same way
- Why Random Forest?
  - When compared to support vector machines and neural networks models, this model was often more computationally demanding with the positive tradeoff of greater accuracy
    - Consistency across all folds of training and holdouts
    - Collective model performs better than single models of individual neighborhoods

IV. Model Relevance

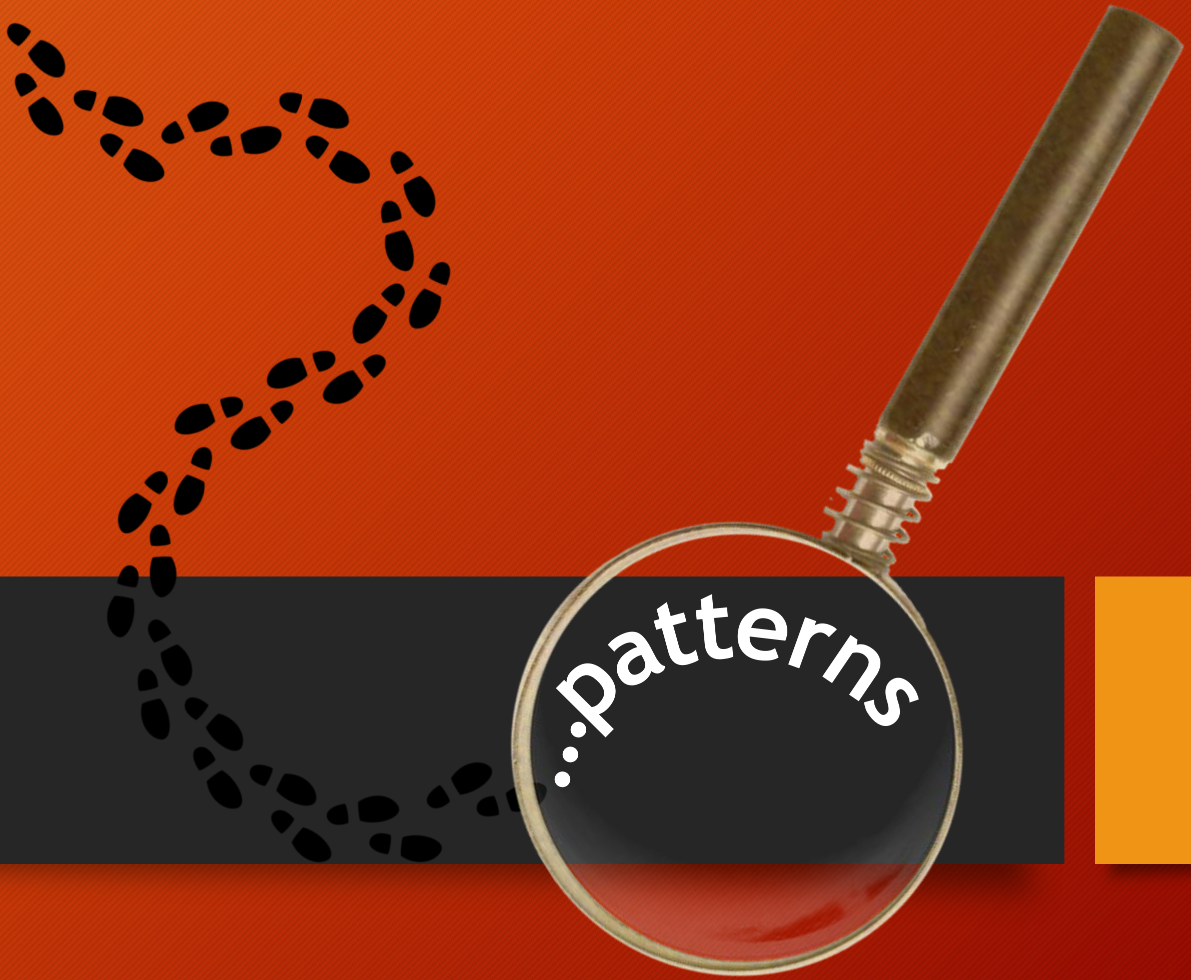
Predicting Daily Crime in Neighborhoods (B)

Random Forest Regressor

- How does the model benefit?
  - The benefits of calculating neighborhood daily crime assists both:
    - Chicago Police Department
      - Staffing on the micro-level
      - Precision for crime fighting
    - Chicago neighborhood resident
      - Informs the resident when conditions are relatively dangerous
      - Empowers them in their choice of living location

- Thinking Ahead?
  - As before, this model can be assessed for efficacy and adjusted accordingly on an annual basis
    - **For the CPD**: model automated and run on *weekly* and *daily* bases by continually pulling from the CLEAR database with 7-day forecast
    - **For the resident**: model provided by CPD in abbreviated format to inform residents of crime activity with precision

While crime does not have a zip code, it does have…



…patterns

Questions & Answers

https://github.com/paulbenschmidt/chicago-crime