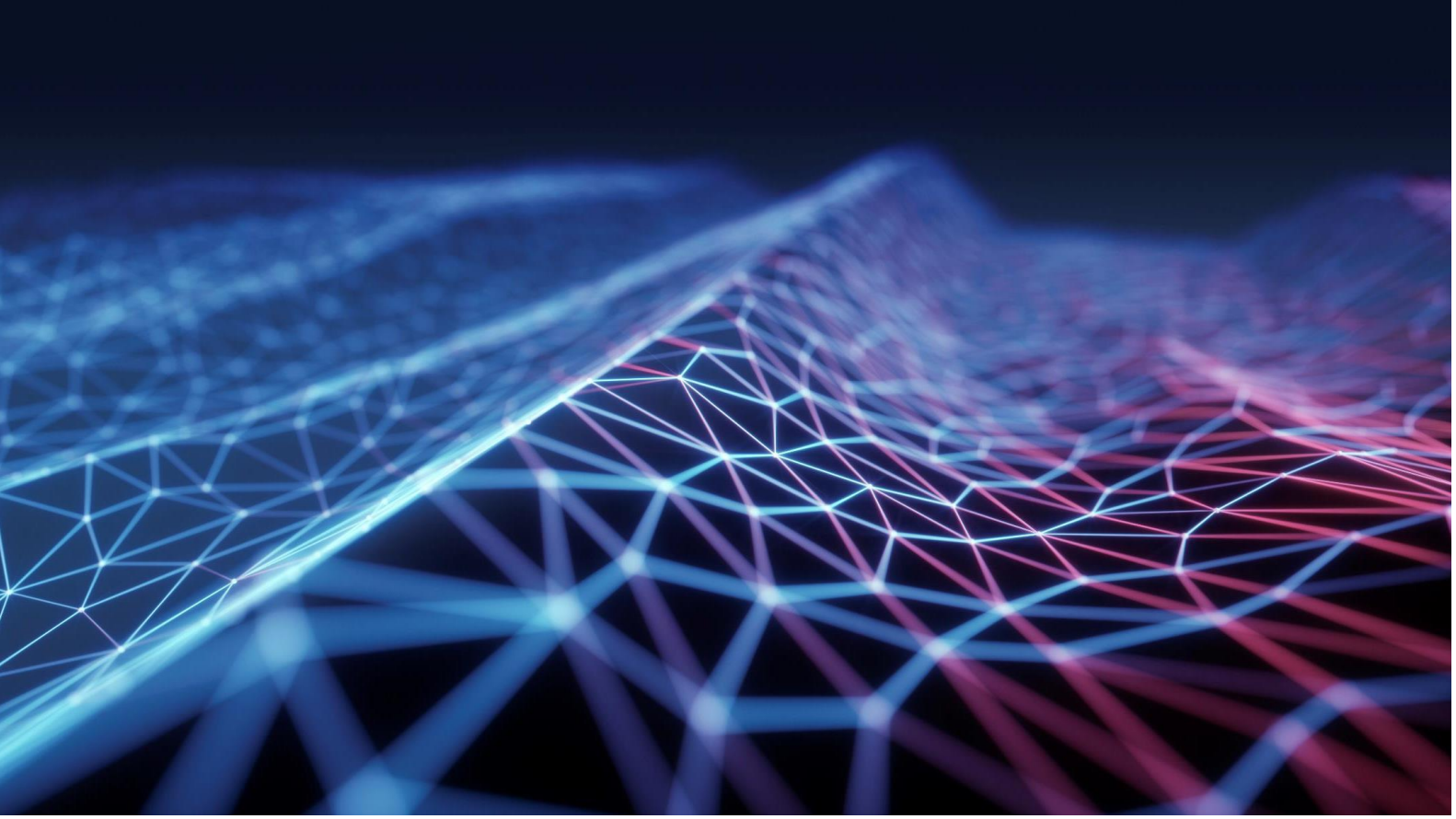


CHICAGO CRIMES

MACHINE LEARNING

MACHINE LEARNING BEAT

- Divya Vishnumurthy

- Ramesh Kalagnanam

- Kevin Kao

- Frances Doyle

- Harshvardhan Suresh

- Joe Strawinski

AGENDA

- What are we doing
- Chicago crime dataset
- Machine learning workflow
- Infrastructure walk through
- Summary of outcomes
- Crime Plots

WHAT ARE WE DOING

Solution statement goals:

- **Define scope (including Data Source)**
- **Define target performance**
- **Define context for usage**
- **Define how solution will be created**

Solution statement:

Use the machine learning workflow to process and transform the Chicago crime dataset to create a prediction model. This model must predict if a crime will result in an arrest with 70% or greater accuracy.

CHICAGO CRIME DATASET

- Data is extracted from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting system
- Reported crime between 2001 and present (less the most recent seven days)
- As of 10/27/19, 6.9 million records
- https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

Database Features:

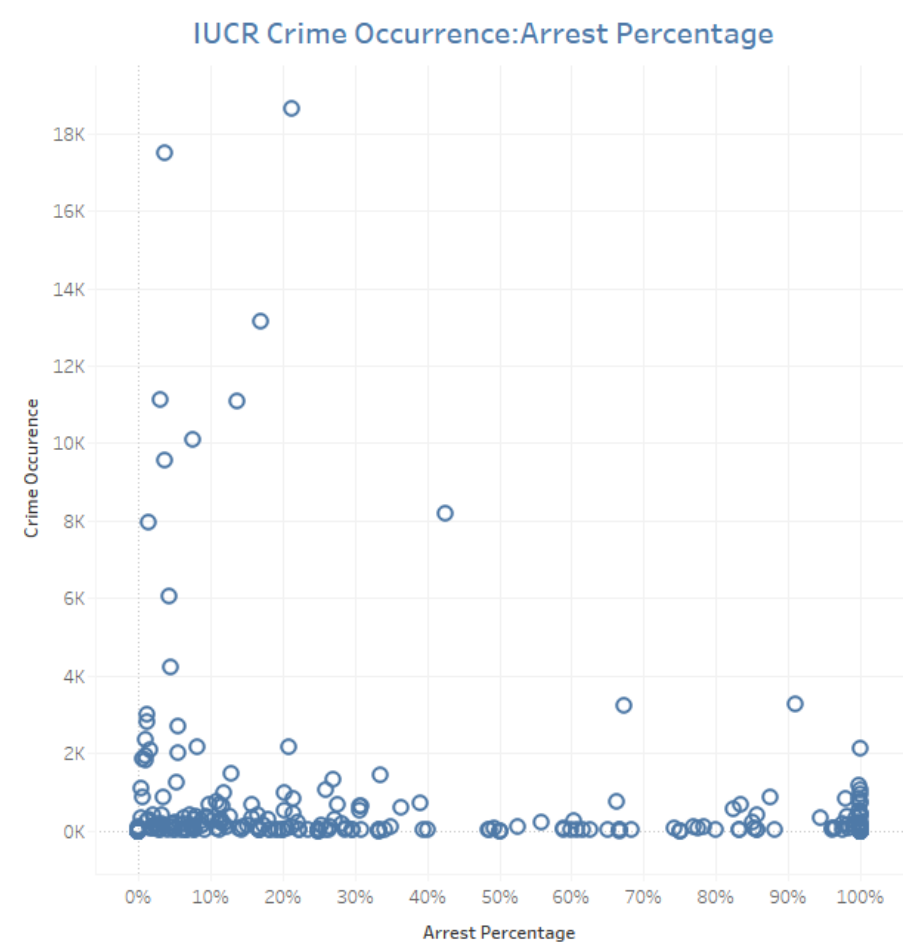
Self-Explanatory columns:

- ID
- Case Number
- Date
- Location Description
- Arrest?
- Domestic?
- District
- Ward
- Community Area
- X Coordinate
- Y Coordinate
- Year
- Updated On
- Latitude
- Longitude
- Location

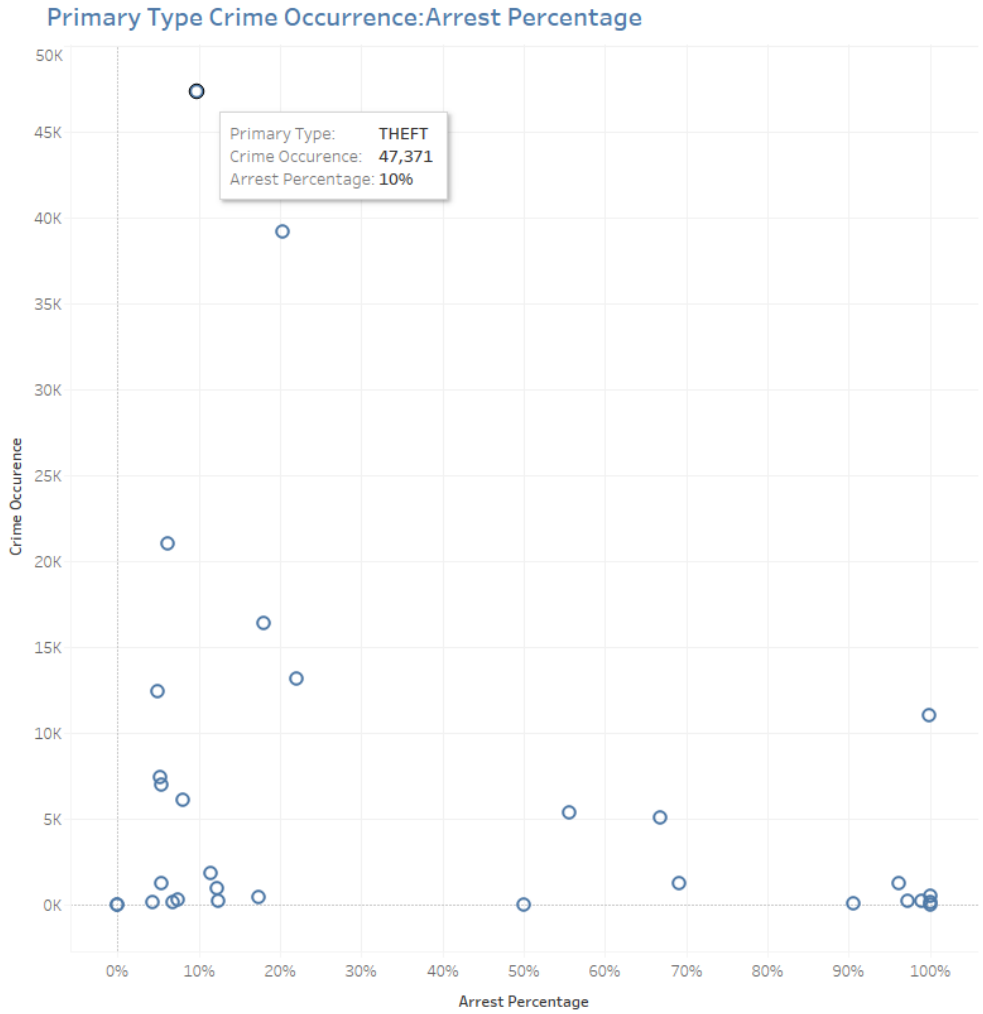
Column definitions:

- Block - partially redacted address
- IUCR – Illinois Uniform Crime Reporting code
- Primary Type – description of IUCR
- Description – secondary description of IUCR
- Beat – smallest police geographic area
- FBI Code – FBI crime classification

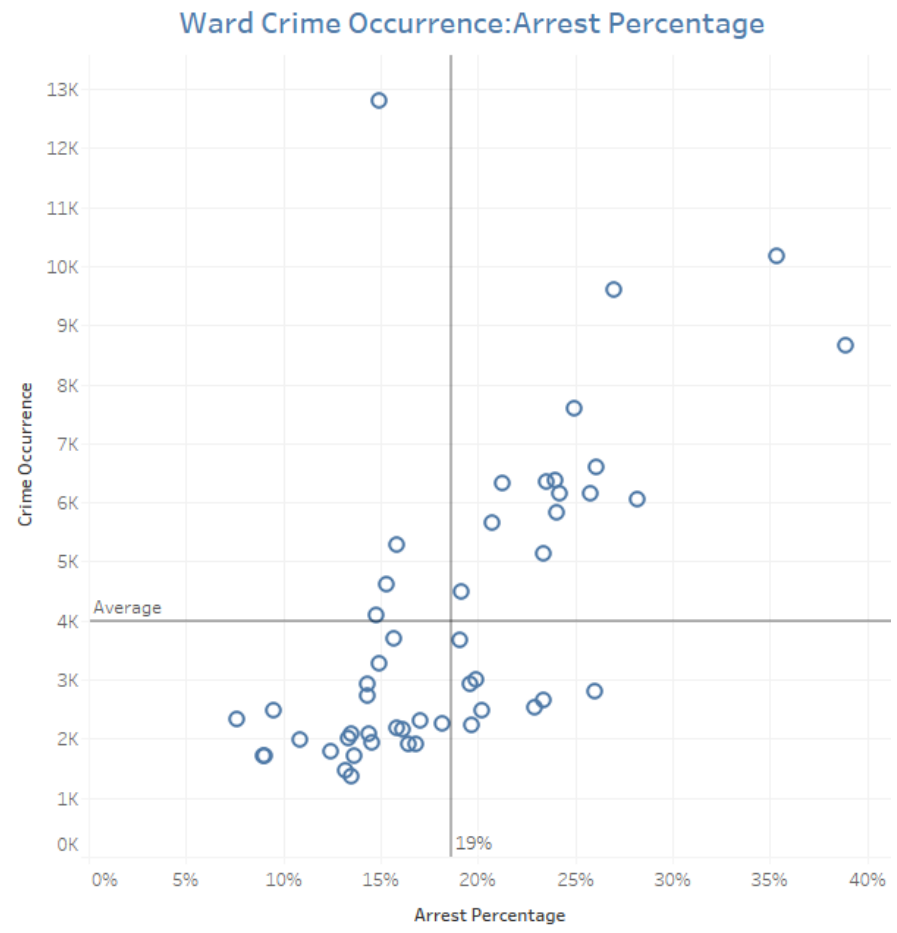
CHICAGO CRIME DATASET



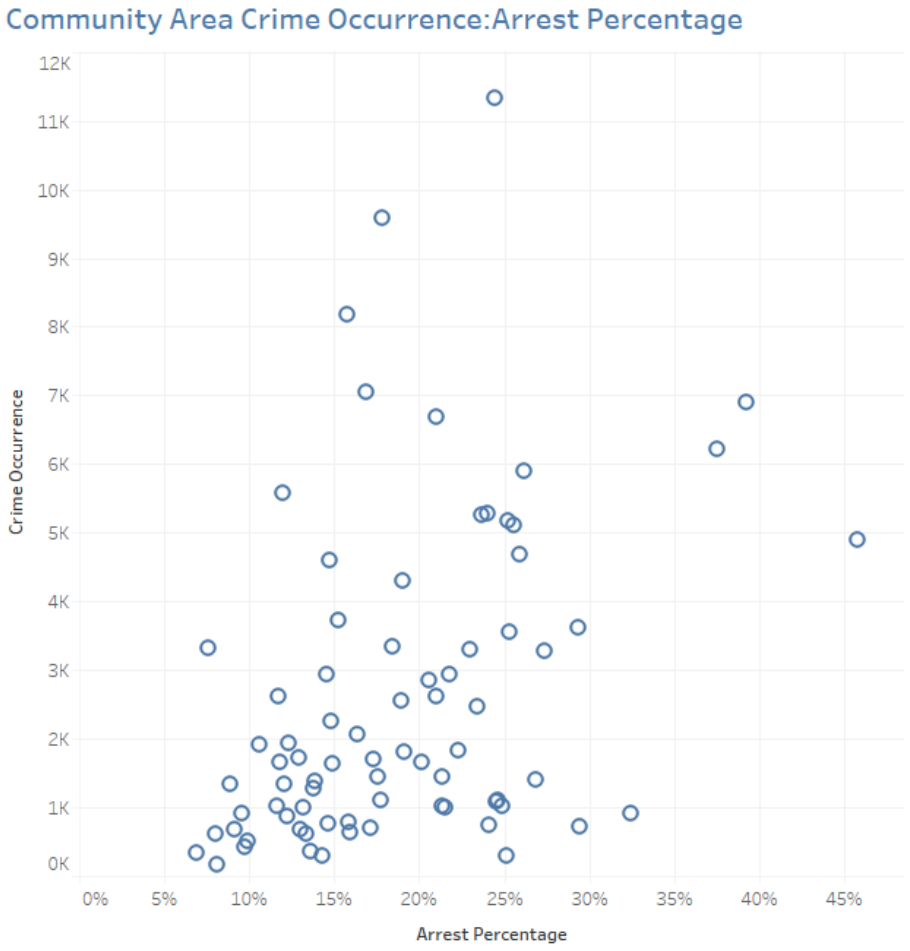
Sum of Arrest Percentage vs. sum of Crime Occurence. Details are shown for Iucr.



CHICAGO CRIME DATASET



Sum of Arrest Percentage vs. sum of Crime Occurrence. Details are shown for Ward.



Sum of Arrest Percentage vs. sum of Crime Occurrence. Details are shown for Community Area.

MACHINE LEARNING WORKFLOW

ASKING THE RIGHT QUESTION

PREPARING THE DATA

SELECTING AN ALGORITHM

TRAINING THE MODEL

TESTING THE MODEL

MACHINE LEARNING WORKFLOW

PREPARING THE DATA

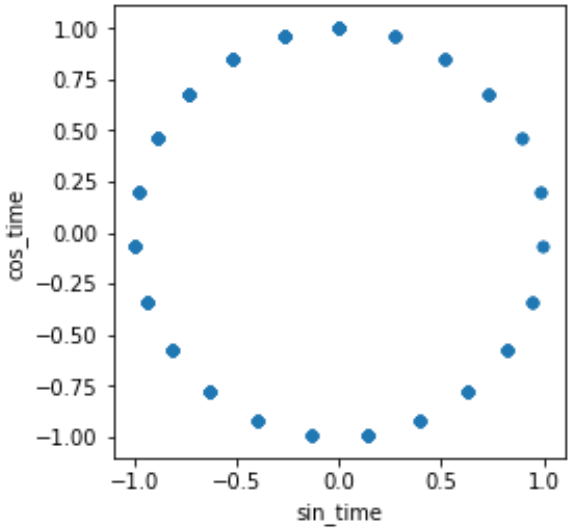
Preparing the data is a major step in machine learning workflow. 50 to 80% of time is spent in this step.

Inspect data, Clean, Load and convert it into Tidy data.

Tidy datasets are easy to manipulate, model and visualize, and have a specific structure.

- Each variable is a column,
- Each observation is a row,
- Each type of observational unit is a table.

Dropped nulls, Convert categorical data into numeric, eliminated outliers, conversion to cyclical data.



MACHINE LEARNING WORKFLOW

ASKING THE RIGHT QUESTION

PREPARING THE DATA

SELECTING AN ALGORITHM

TRAINING THE MODEL

TESTING THE MODEL

MACHINE LEARNING WORKFLOW

SELECTING AN ALGORITHM

Decision factors for picking our algorithms

- Learning Type – Supervised, Unsupervised
- Result Type – Classification, Regression
- Complexity of the algorithm
- Basic Vs Enhancement

- ✓ Naïve Bayes
- ✓ Random Forest
- ✓ Decision Tree
- ✓ Logistic Regression

MACHINE LEARNING WORKFLOW

ASKING THE RIGHT QUESTION

PREPARING THE DATA

SELECTING AN ALGORITHM

TRAINING THE MODEL

TESTING THE MODEL

RANDOM FOREST

Features: IUCR, Time of Occurrence, Location (modified Latitude/Longitude)

	Predicted "No"	Predicted "Yes"	
Actual "No"	**TN = 1,279,397**	**FP = 90,149**	**1,369,546**
Actual "Yes"	**FN = 183,435**	**TP = 332,810**	**516,245**
	1,462,832	**422,959**	**1,885,791**

Precision	Accuracy
When it predicts "Yes", how often is it correct? TP/(TP + FP) = 332,810/422,959 = 78.7 percent	Overall, how often is the classifier correct? (TP + TN)/total = (332,810+1,279,397)/1,885,791= 85.5 percent

Evaluating the model with new dataset

- Tested with 500 new records
- 425 out of 500 ("Predicted" True = Actual True)
- Accuracy = 425/500 = 85 percent

LOGISTIC REGRESSION

Features: IUCR, Time of Occurrence, Location (modified Latitude/Longitude)

	Predicted "No"	Predicted "Yes"	
Actual "No"	**TN = 1,006,257**	**FP = 363,289**	**1,369,546**
Actual "Yes"	**FN = 206,925**	**TP = 309,320**	**516,245**
	1,213,182	**672,609**	**1,885,791**

Precision	Accuracy
When it predicts "Yes", how often is it correct? TP/(TP + FP) = 309,320/672,609 = 45.9 percent	Overall, how often is the classifier correct? (TP + TN)/total = (309,320+1,006,257)/1,885,791= 69.7 percent

Evaluating the model with new dataset

- Tested with 500 new records
- 332 out of 500 ("Predicted" True = Actual True)
- Accuracy = 332/500 = 66.4 percent

LOGISTIC REGRESSION

Features: Crime type, Domestic crime(Y/N), District, Ward, Community Areas, Beat

	Predicted "No"	Predicted "Yes"	
Actual "No"	TN = 50,984	FP = 1,305	52,289
Actual "Yes"	FN = 7,167	TP = 5,949	13,116
	58,151	7,254	65,405

Precision	Accuracy
When it predicts "Yes", how often is it correct? TP/(TP + FP) = 5,949/7,254 = 82 percent	Overall, how often is the classifier correct? (TP + TN)/total = (5,949+50,984)/65,405 = 87 percent

Evaluating the model with new dataset

- Tested with 203,786 new records
- 176,662 out of 203,786 ("Predicted" True = Actual True)
- Accuracy = 176,662/203,786 = 86.7 percent

NAÏVE BAYES

Features: IUCR, Community Area, Police Beats, Hour, Month

	Predicted "No"	Predicted "Yes"	
Actual "No"	**TN = 236,478**	**FP = 14,428**	**250,906**
Actual "Yes"	**FN = 59,046**	**TP = 4,621**	**63,667**
	295,524	**19,049**	**314,573**

Precision	Accuracy
When it predicts "Yes", how often is it correct? TP/(TP + FP) = 4,621/19,049 = 24.3 percent	Overall, how often is the classifier correct? (TP + TN)/total = (236,478+4,621)/314,573= 76.6 percent

Evaluating the model with new dataset

- Tested with 500 new records
- 373 out of 500 ("Predicted" True = Actual True)
- Accuracy = 373/500 = 74.6 percent

RANDOM FOREST
Features: IUCR , Community Area, Police Beats, Hour, Month

	Predicted "No"	Predicted "Yes"	
Actual "No"	**TN = 237,048**	**FP = 13,858**	**250,906**
Actual "Yes"	**FN = 30,034**	**TP = 33,633**	**63,667**
	267,082	**47,491**	**314,573**

Precision	Accuracy
When it predicts "Yes", how often is it correct? TP/(TP + FP) = 33,633/47,491 = 70.8 percent	Overall, how often is the classifier correct? (TP + TN)/total = (33,633+237,048)/314,573= 86 percent

Evaluating the model with new dataset

- Tested with 500 Records
- 437 out of 500 ("Predicted" True = Actual True)
- Accuracy = 437/500 = 87.4 percent

EFFICIENCY METRICS OF CHICAGO PD

The goal was to determine the efficiency of a police beat based on the crime type, police district and ward.

A beat is the smallest police geographic area with a dedicated police car. Three to five beats make up a police sector, and three sectors make up a police district.

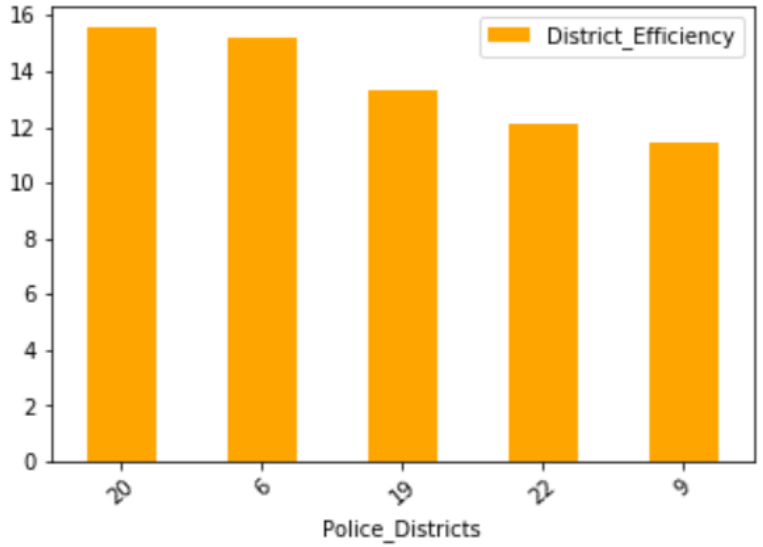
Chicago police hold a monthly regular policing strategy meeting for the public to provide safety feedback based on the recent crime events.

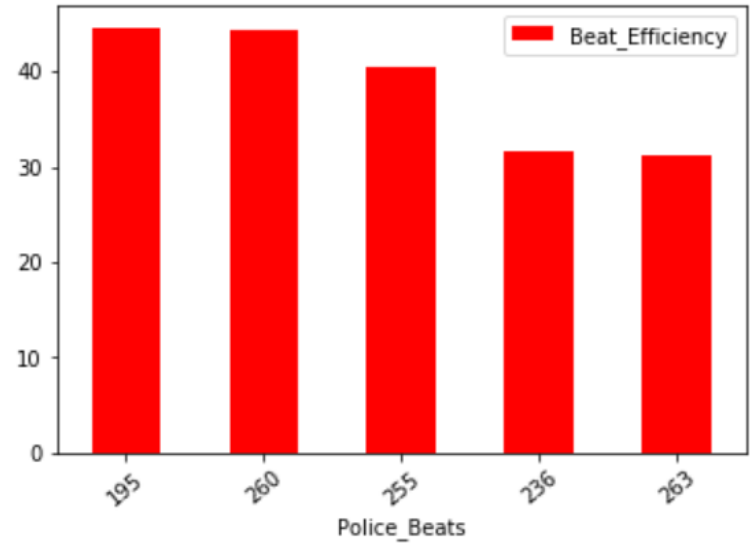
The intent is to use these metrics to determine the most effective police district based on the crime type. Data will be used to re-assign resources to improve arrest rates.

EFFICIENCY METRICS OF CHICAGO PD

- With features- Crime type, Ward, Police District and Police Beat, the model works with an accuracy of 87%

- Limitations in dataset:
  - The lag between event and arrest is unknown.
  - This missing data would help calculating efficiency better.

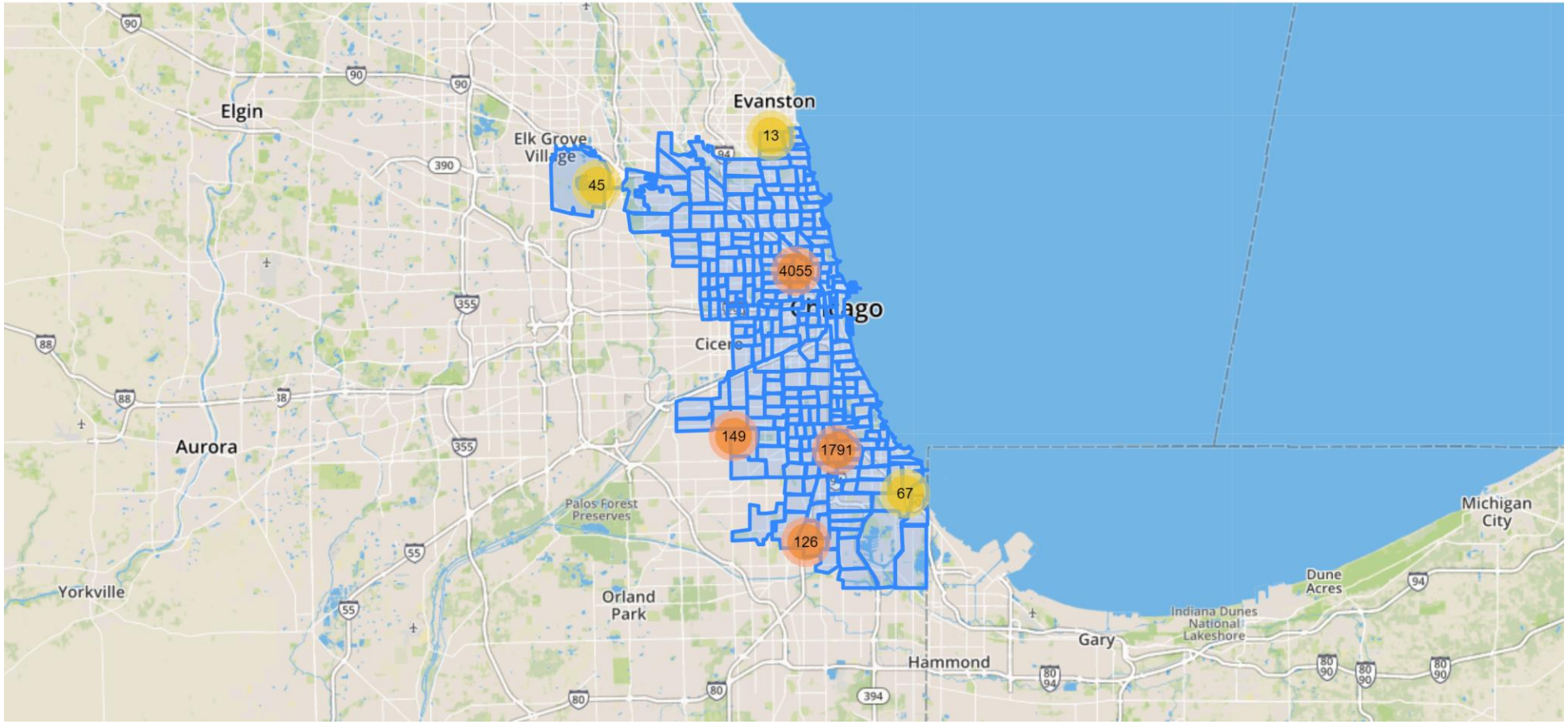




ALGORITHM WITH BEST OUTCOMES

- Random Forest 87%
- Logistic Regression 87%
- Naïve Bayes 74%
- Decision Tree 87%

DEMO


Elgin
Evanston
Elk Grove Village
13
45
4055
Aurora
149
1791
67
126
Palos Forest Preserves
Michigan City
Yorkville
Orland Park
Dune Acres
Indiana Dunes National Lakeshore
Gary
Hammond

THANK YOU