



Capstone Project: Chicago Crime Predictions

Darius Smith | BrainStation | April 11, 2023



Introduction:



What is safety?

Safety is a state in which hazards and conditions leading to physical, psychological or material harm are controlled in order to preserve the health and well-being of individuals and the community.

- Is safety a deciding factor for you when moving into another neighborhood?
- What if we could look at 'Arrests' in a city or community to not just avoid those areas but to make it better?

The problem statement for this project?

“Using machine learning, how may we predict how 'Arrests' change with respect to time , location description, and other associated features so that we can make communities in Chicago safer for everyone?”

Data Collection:

Where was the data collected?

- Data was retrieved from the city of Chicago website.
- <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2>

How did I collect the data?

- Data was collected simply by going on the site and downloading the 'csv' that was provided.
- From there saving the file to the Desktop took a matter of seconds.



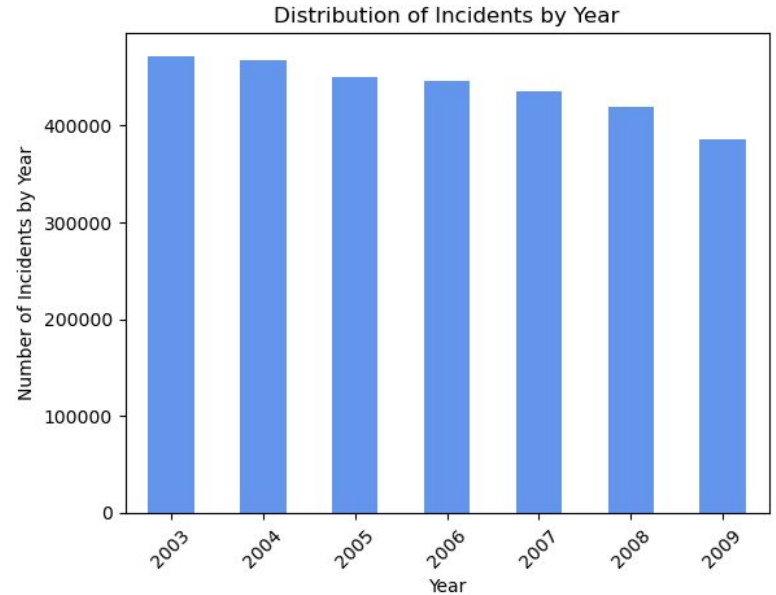
Data Description:

The dataset contained the following:

Reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present.

The columns represent where incidents took place, what incidents took place, and if there was an arrest during those incidents, and if they were domestic.

Target Variable: 'Arrests'



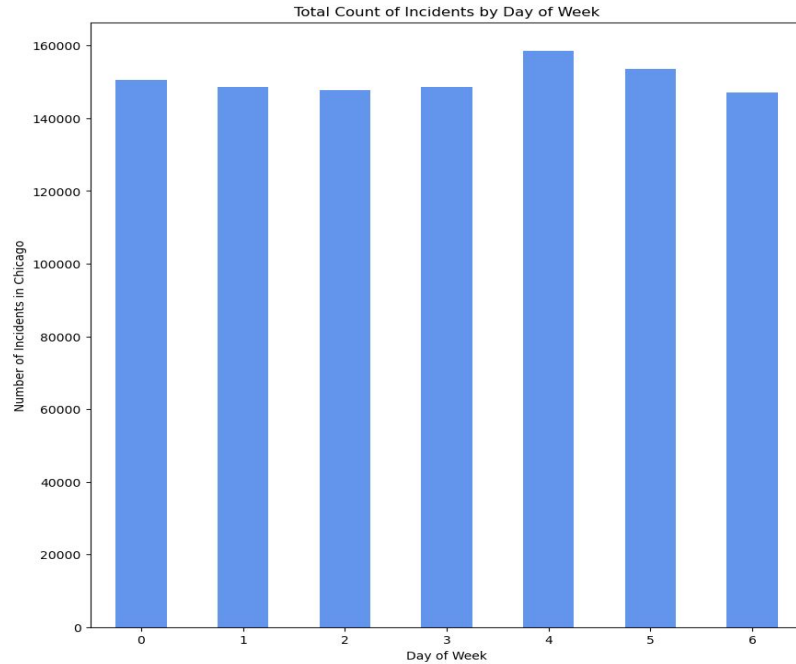
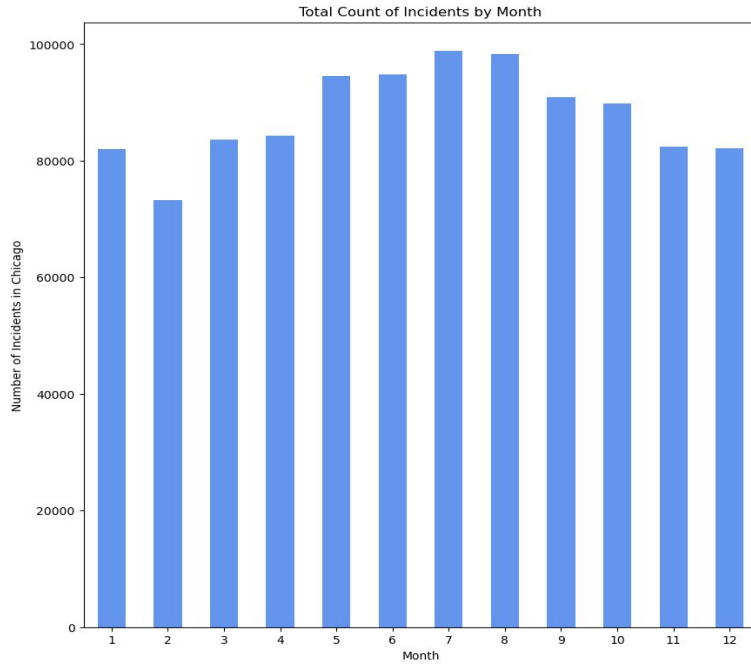
Incidents of Crime Visual

Exploratory Data Analysis:

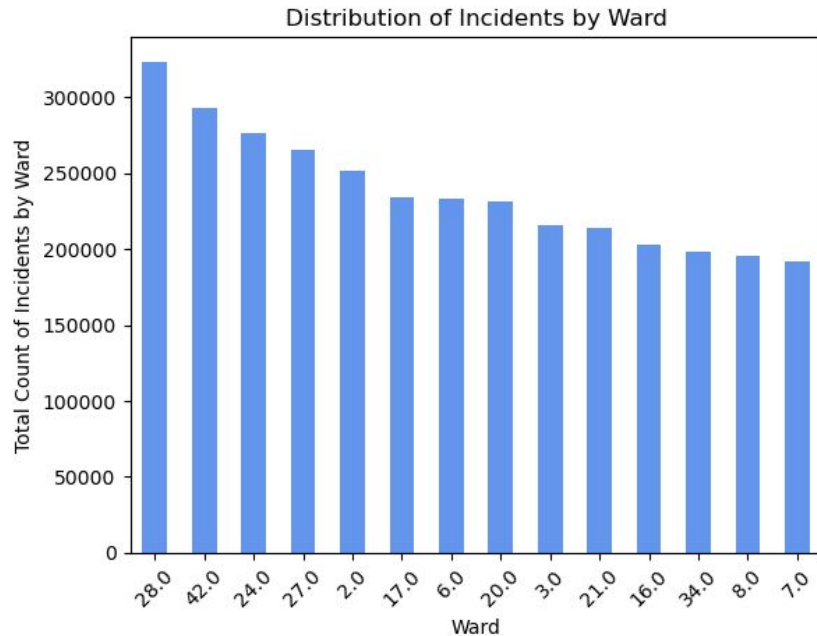


- 2003 saw the highest number of incidents with (471,986).
- There seems to be a downward trend in incidents after the year 2003.
- There were (5,221,926) incidents between 2001 - 2023, where someone was not '**Arrested.**'
This accounted for 74% of total incidents.
- '**Month**' - Summer months produced the highest numbers of crime/incidents.
- '**Day of Week**' - Friday and Saturdays produced the highest number of crime/incidents.
- '**Hour**' - 12:00pm, 6:00pm, and 7:00pm had the highest number of incidents.

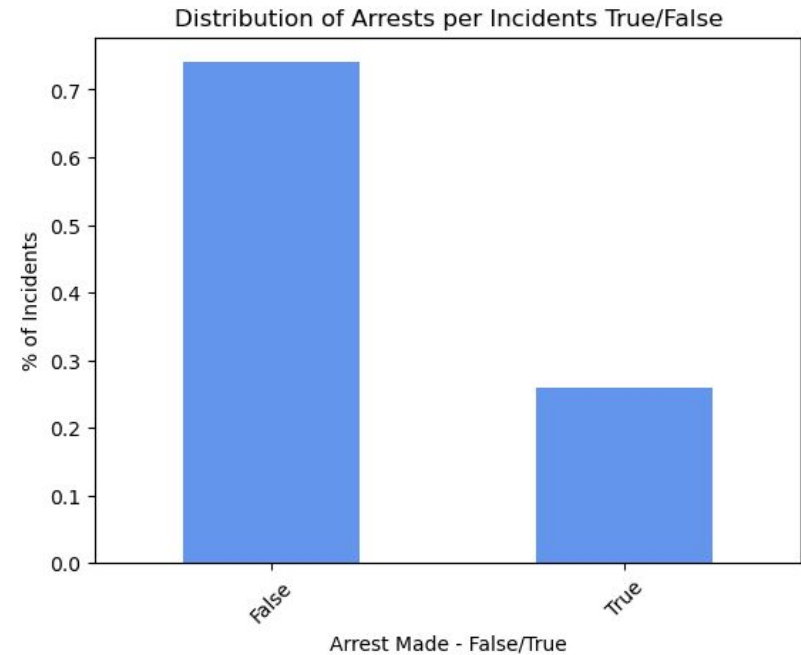
EDA Visualizations:



EDA - Additional Visualizations:



Distribution of Incidents by Ward Visual



Distribution of Arrests Visual

Data Clean Up:

Before cleaning, the dataset contained the following:

- (7,742,476) data points, (22) columns, (1,663,647) missing values, and (0) rows of duplicated data.

After cleaning, and narrowing the dataset down to the years 2016-2019, the dataset contained:

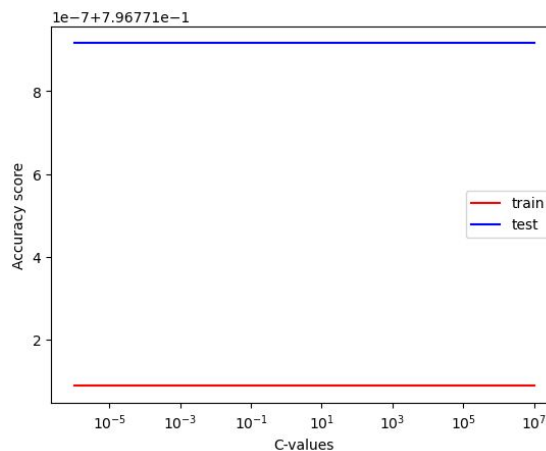
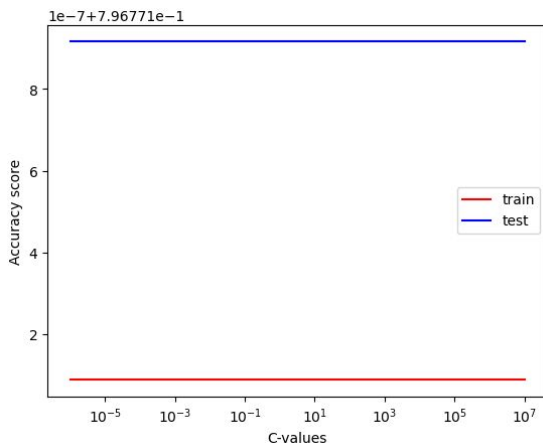
- (1,011,679) data points, (16) columns, (0) missing values, and (0) rows of missing values.

	Arrest	Domestic	Beat	District	Ward	Community Area	X Coordinate	Y Coordinate	Year	Latitude	Longitude	Month	Day	Week	Day of Week	Hour
0	0	0	132	1	3	33	1177560.0	1889548.0	2019	41.852248	-87.623786	9	24	39	1	8
1	0	0	1221	12	26	24	1160005.0	1905256.0	2019	41.895732	-87.687784	10	13	41	6	20
2	0	0	1224	12	27	28	1166986.0	1900306.0	2019	41.882002	-87.662287	10	5	40	5	18
3	0	0	1922	19	47	6	1164930.0	1923972.0	2019	41.946987	-87.669164	10	13	41	6	19
4	0	0	2033	20	47	3	1167380.0	1934505.0	2019	41.975838	-87.659854	10	13	41	6	14

Final Dataframe

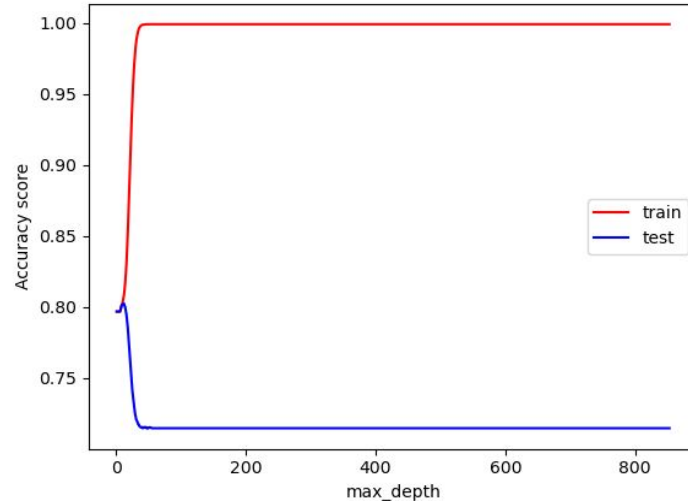
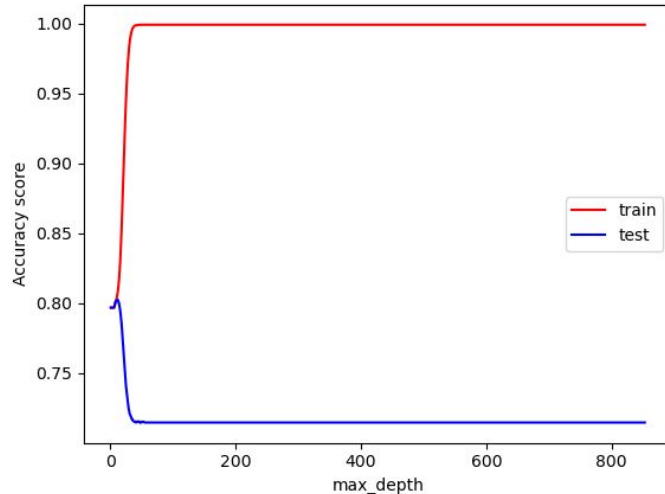
Modeling | Logistic Regression:

- Two Logistic Regression baseline models were ran. They both produced an accuracy score of approximately ~ 80%.
- An additional two attempts were made to make this score better by using modeling hyperization values. Below are the visuals of the attempts.
- They both reproduced the same score, approximately ~ 80%.



Modeling | Decision Tree:

- After multiple replicated results from the Logistic Regression model, a Decision Tree model was used to attempt to increase the accuracy score.
- They both produced the same score as the Logistic Regression model ~ 80%.
- Below are the results for both models ran.



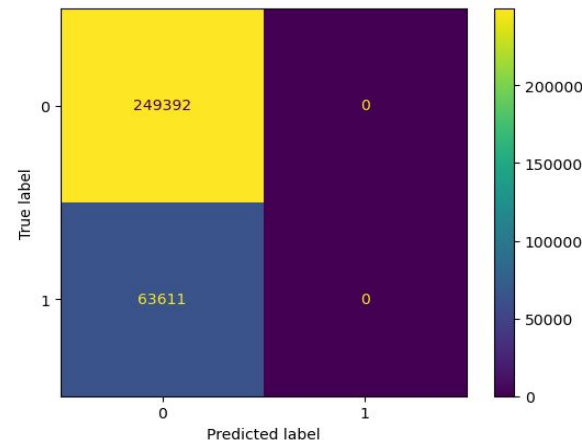
Model Interpretation:

What did this all mean?

Were my models able to successfully predict an arrest?


After running a classification report and matrix it was clear:

- My models were unsuccessful predicting an arrest.
- Not only were they unsuccessful, but it was conclusive.
- This was what the **80%** was alluding to.



	Predicted Non-Arrest	Predicted Arrest
True Non-Arrest	249,392	0
True Arrest	63,611	0

Conclusion:

- 
- There seems to be a downward trend in incidents after the year 2003.
 - There were (5,221,926) incidents between 2001 - 2023, where someone was not '**Arrested.**' This accounted for 74% of total incidents.
 - Predicting an arrest proved elusive...**for now.**
 - The next steps for this project will be to use additional ML models to address the problem statement.
 - I like all of you am interested in creating a safer and better world for future generations.
 - This is the beginning not the end.



Questions