Capstone Project: Final Report

Darius Smith | April 11, 2023

Problem Statement & Background:

The city of Chicago's **'Crimes – 2001 to Present'** is a dataset that uses data from police reports, and areas from around the city to report on the number of incidents or crimes that have occurred around the city. The data is extracted from the **'Citizen Law Enforcement Analysis and Reporting'** system which is a division of the police department that uses this data to contribute to ongoing measures to increase public safety.

My interest in this dataset comes from me being a Chicago native. I was born and raised in the city. I saw a lot of violence, poverty, and hopelessness that proved to be a breeding ground for crime in my neighborhood. As someone who is now an adult and saw these issues growing up, my mind shifted to the future. *"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?" "Could I be someone to help make this city safer?" "What could this mean for other cities in the United States?"*

It was from here I decided to explore this dataset with the hopes of creating a machine learning model or models that could predict an **'Arrest'** so that crime could be reduced in the city. The intention was to use this data in an ethical way to not create a **'Minority Report'** situation, but to understand where community organizations, resources, and police can work together.

About the Data:

Finding a large dataset that reported on crime was easier than anticipated. In the beginning, I was able to find many **'csv.'** files on the city of Chicago's website. However, they only captured one year whereas I needed multiple years of data if I was going to create a successful prediction model. With a little time, I was able to find the data I needed.

The dataset had over 7,000,000 rows and 22 columns. The columns included date, description, location description, arrest, domestic, district, ward, community area, year, and many other columns some of which was created by me for modeling – **'day', 'day of week', 'month', and 'week.'** Fortunately, I had the column I needed to begin making predictions which was **'Arrest.'** Before doing any of that, I knew I needed to dive in an get a better understanding of this data.

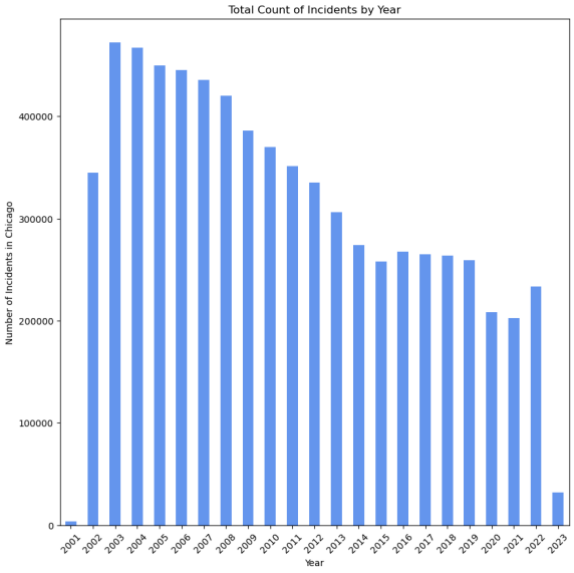
Data Cleaning and Exploratory Data Analysis:

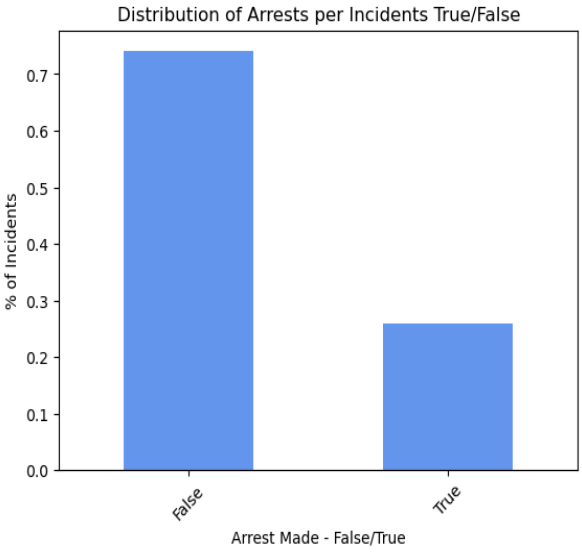
After collecting my dataset, one of the first things I wanted to do was check for missing values and duplicated data. Fortunately, there were no duplicated rows, but there were many missing values, **1,663,647** to be exact. Given this, there were a couple of methods that I could use to address this, I could (1) remove the missing values, or (2) impute them, meaning fill in the missing values.

The missing values were in the **'ward', 'community area', 'latitude', 'longitude', and the 'X and Y coordinate'** columns. The goal was to fill in the 'ward' and 'community area' that had missing values with the proper 'latitude' and 'longitude.' However, after searching for these online, I noticed that there were locations in the data that were not in Chicago. It was at this point I elected to drop all missing values and to proceed with EDA.

Below is the preliminary analysis of the data with some visuals below:

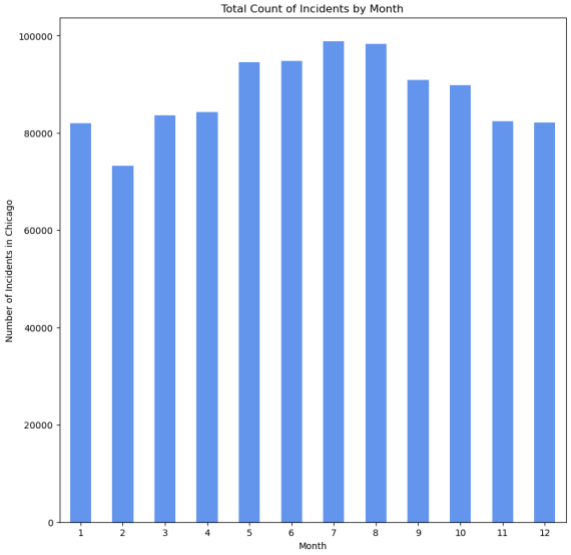
- 2003 saw the highest number of incidents with (**471,986**).
- There seems to be a downward trend in incidents after the year 2003.
- There were (**1,813,335**) incidents that occured on the street that is 26% of incidents.
- There were (**5,221,926**) incidents where someone was not 'Arrested.' This accounted for 74% of total incidents.

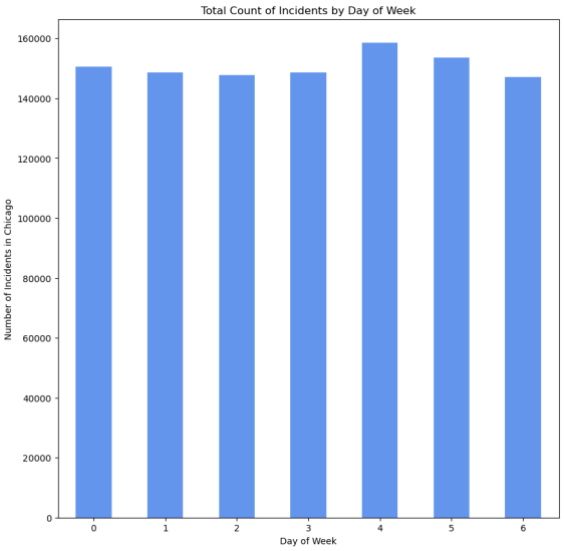




After this basic EDA I then proceeded to reduce the data for modeling. I decided to choose the **'Year'** column as it proved easiest to narrow the dataframe. It was from here where I decided to create a new dataframe for modeling, and to save as a new 'csv.' file.

Before modeling there were many categorical columns that needed to be converted to numerical for modeling. The **'Date'** column proved to be most insightful as it provided additional EDA that was helpful to grasp a picture of incidents/crime in the city. *The summer months saw the highest number of crimes, while colder months showed a decrease in incidents*. Also, the weekends showed the highest number of incidents.





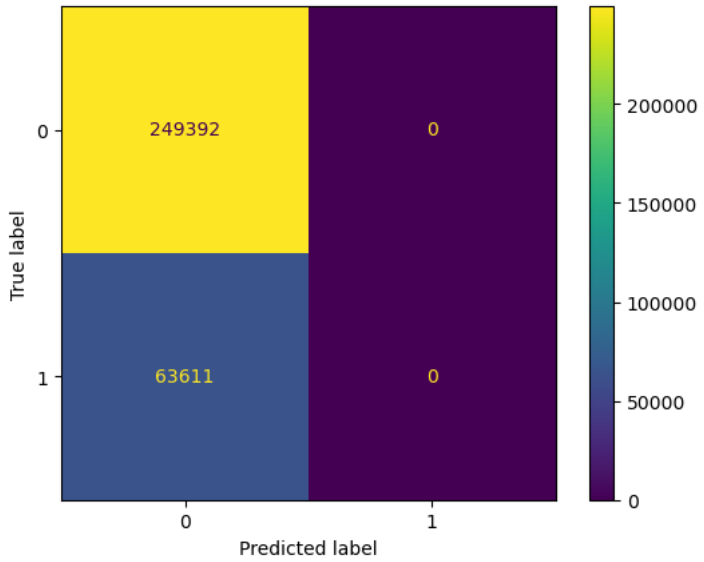
Modeling and Modeling Optimization:

After EDA, I moved to modeling. One thing to note, was that many categorical columns had to be dropped because they would have created a very wide dataframe for modeling. These included columns like **'Location Description'** and **'Location.'** The final dataframe used for modeling is below.

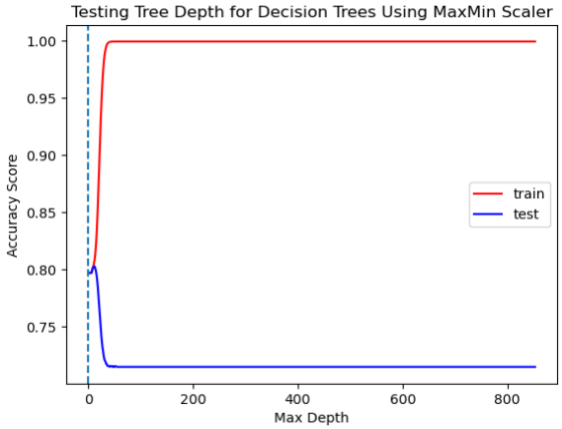
	Arrest	Domestic	Beat	District	Ward	Community Area	X Coordinate	Y Coordinate	Year	Latitude	Longitude	Month	Day	Week	Day of Week	Hour
0	0	0	132	1	3	33	1177560.0	1889548.0	2019	41.852248	-87.623786	9	24	39	1	8
1	0	0	1221	12	26	24	1160005.0	1905256.0	2019	41.895732	-87.687784	10	13	41	6	20
2	0	0	1224	12	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

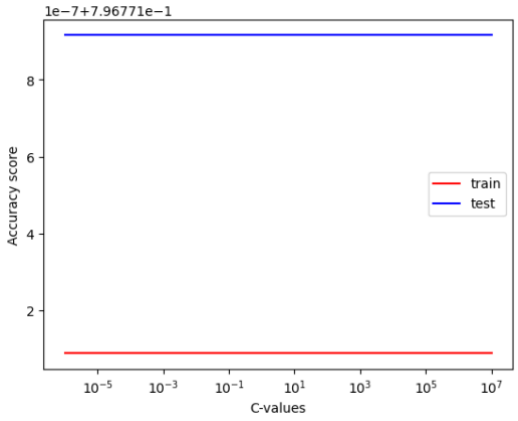
After separating the independent variables and dependent variable **('Arrest'**) two baseline models were ran using a logistic regression. The accuracy was approximately **~80%** for both. From here an attempt at interpretability and optimization was made using a confusion matrix. **(Figure below).**

All model optimization attempts showed an accuracy of 80%. This was both for the logistic regression and decision tree. It was also observed that, an **'Arrest'** could not be correctly predicted. C-Values were used to optimize the logistic regression, and the max depth was used for the decision tree. **(Visuals are below.)**



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





Conclusion and Next Steps:

This project provided many insightful trends and patterns about crime in the city of Chicago. These included findings such as:

- 2003 saw the highest number of incidents with (**471,986**), but years after this saw a decline in crime.
- There were (**5,221,926**) incidents where someone was not 'Arrested.' This accounted for 74% of total incidents.
- The summer months saw the highest number of crimes, while colder months showed a decrease in crimes.

The problem statement, ***"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?"*** could not be answered at this time and results were conclusive of this. However, other models will be used in the future to figure this out. It is not a matter if not now, but when we can predict. I like others am interested in creating a safer, and better world for future generations. Especially as a Chicago resident.