Weather and Its Effect On Crime (Chicago)

Chicago is one of the most well known cities in the world. With many attractions such as the Willis Tower, Navy Pier, Soldier Field, and some of the best deep dish pizza around. Although with all those good comes another thing Chicago may also be known for its crime. What if there was a better way to be prepared for crime such as knowing when the weather is perfect for a severe crime. In this project I will be looking into the correlations between weather data and crime data to see if we can find patterns and help predict whether a crime was severe or not.

1. The Data

I used a combination of 2 datasets for this project. The first is the crime dataset for chicago, I got that from the Chicago Data Portal it lists all crimes from 2001 - present. Then for the weather dataset I went to weather.gov and requested a dataset of all weather data from Midway Airport between 2001-2017.

- Crime Data - Features: Date, Primary Type - Crime type, Arrest, Domestic, District Name, and many other features although.

- Weather Data - Features: Average daily wind speed, Fastest mile wind time, Peak gust time, Precipitation, Snowfall, Snow depth, Average/Min/Max Temp, Fastest 2 and 5 min wind, Fog/Ice fog, Heavy Fog, Thunder, HailSmoke/Haze, Tornado

2. The Idea

For This project I decided it would be best to create a model that would predict when a crime was severe or not. I classified a severe crime as any of these offenses (OFFENSE INVOLVING CHILDREN, CRIM SEXUAL ASSAULT, BATTERY, NARCOTICS, BURGLARY, SEX OFFENSE, ROBBERY, PROSTITUTION, HOMICIDE, KIDNAPPING, ARSON, HUMAN TRAFFICKING) and also anytime someone got arrested. So now we can see how the weather for an individual day may make a crime more likely to be severe or not. In the future I think other approaches may prove better such as weather and affect on location and also if we had more states and cities to compare I think more insights can be found.

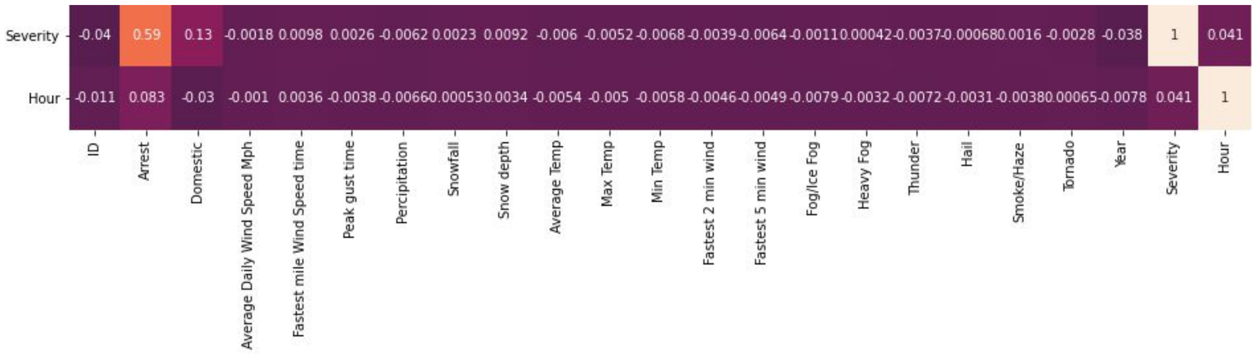
3. Data Cleaning

Crime Data - For the crime data the first matter was to drop all the useless and repetitive columns. First I dropped all columns filled with na values that I could not fill or did not need like longitude and latitude. Then for the repetitive columns for the location I used the districts since there were not as many na values compared to community areas and other location descriptions were vague pretty vague. For the offense descriptions I only kept the primary description which is what the crime is classified as not what happened since that info was vague as well. Finally I had to make sure to change all the district names since they were numbers and I had to make the dataframe end on the date 2017-12-31 since my weather data only went to that date.

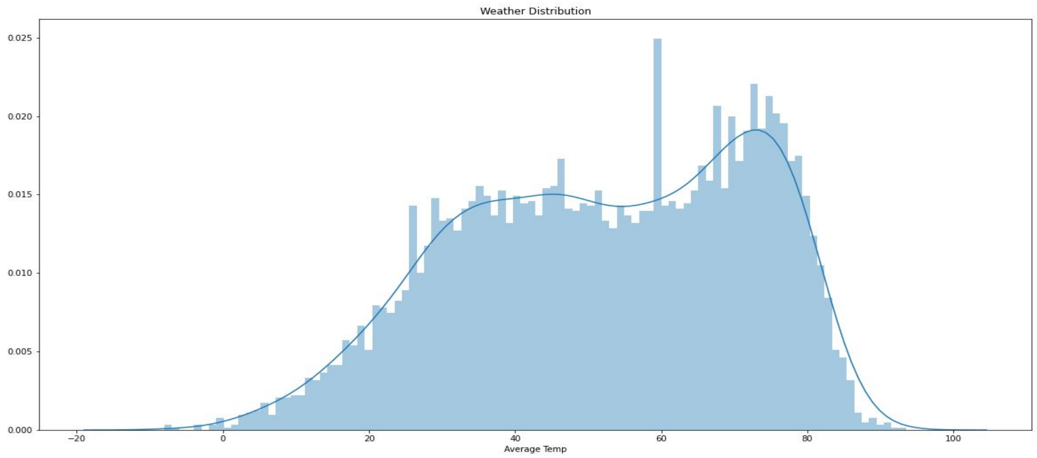
Weather Data - the weather data was a lot easier to clean and sort through than the crime data. I only needed to change the column names from the abbreviations since only a meteorologist would know what they all mean. I also had to fill the average temp column since it had so many missing values while all the others had no missing values. To do this I used an lambda function with

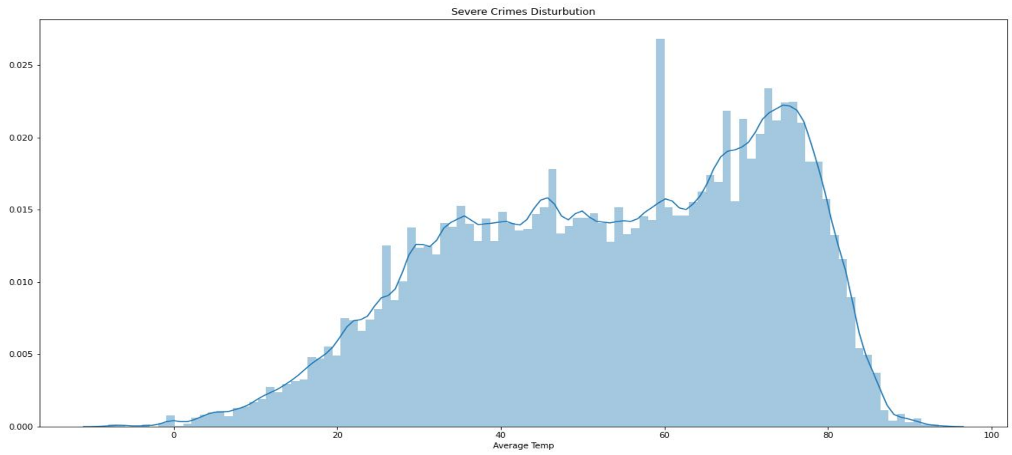
.apply to add the min and max temp together then divide them by 2 and take that value to replace the nan value if it was nan.

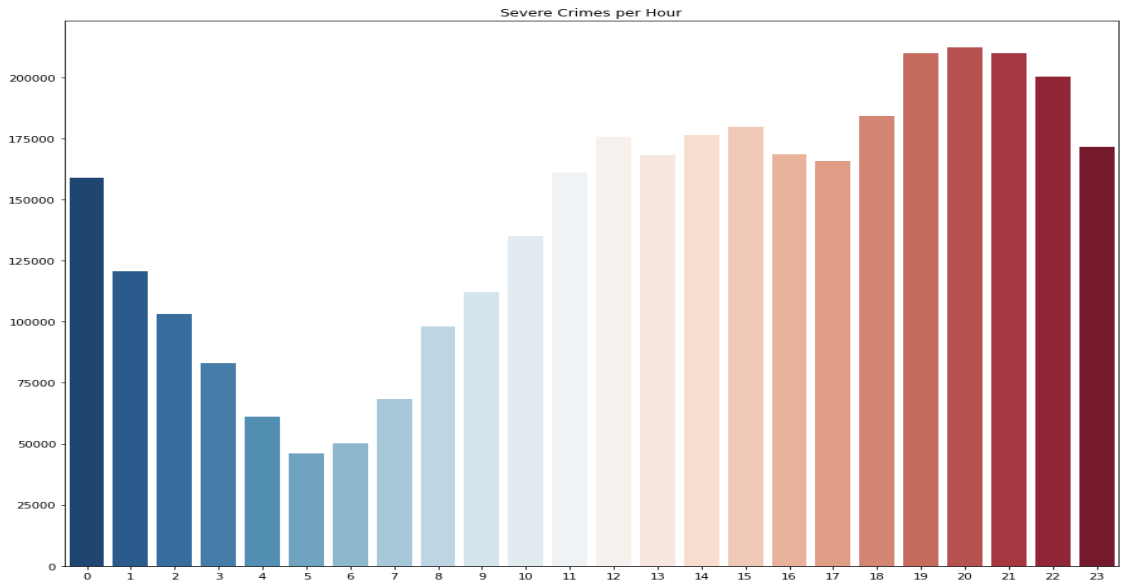
4. EDA

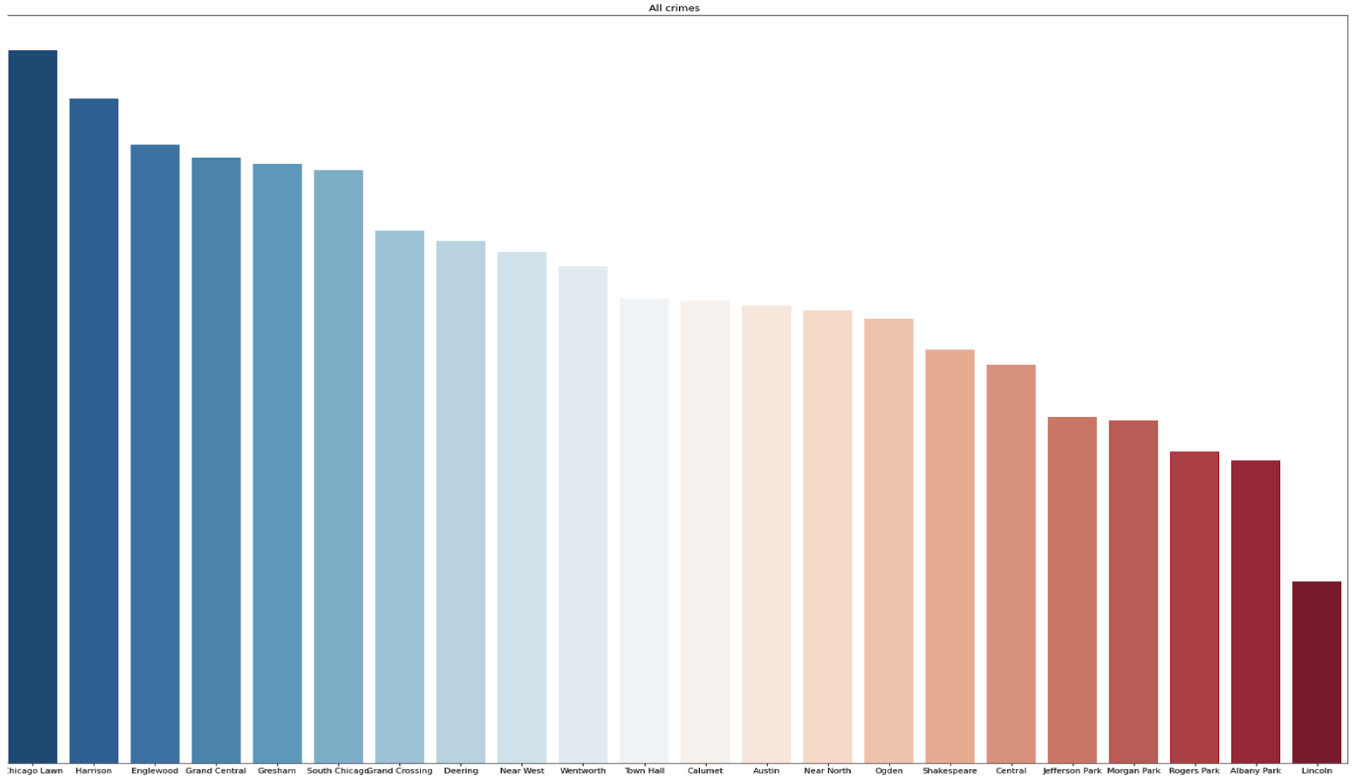


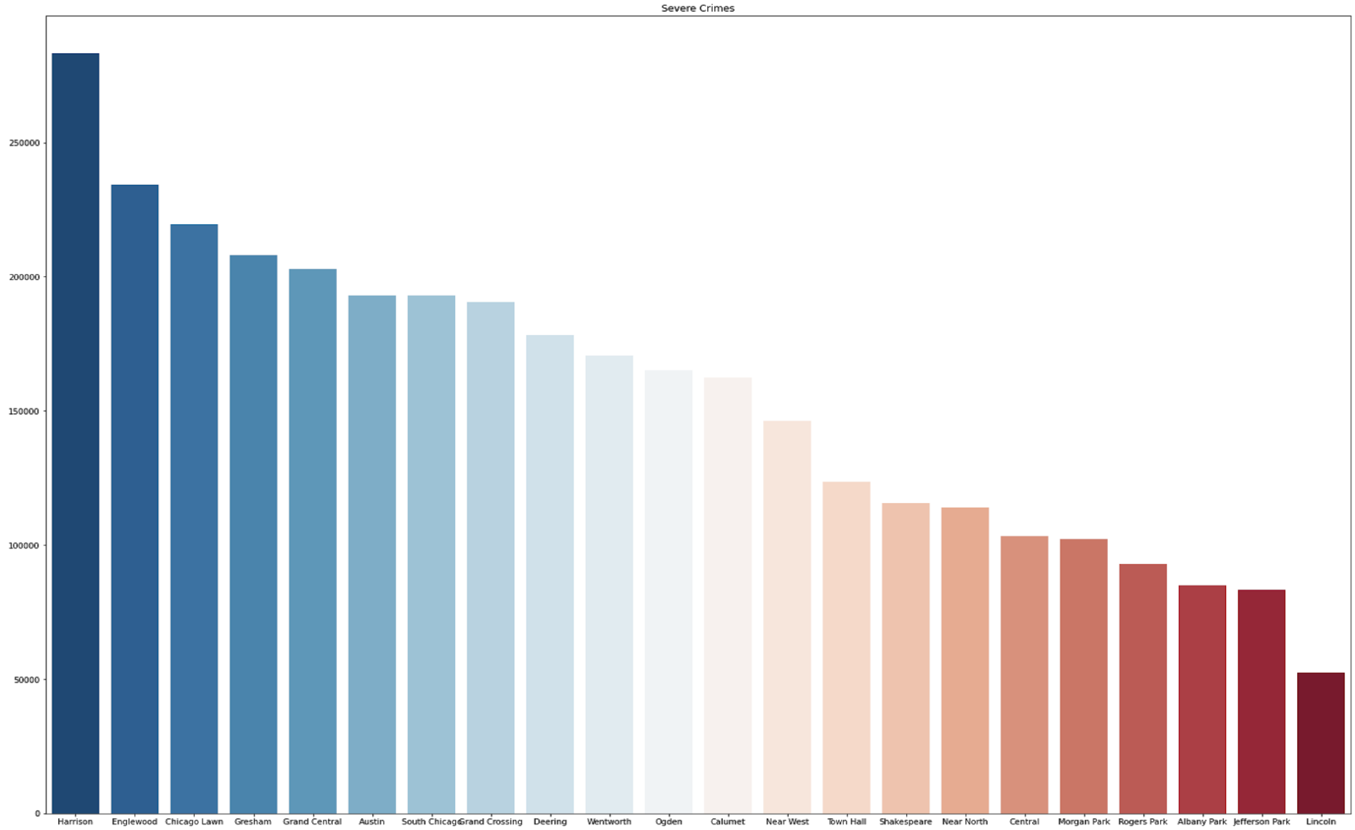
Beginning my EDA I started with this correlation heatmap and it told me a lot about the weather and severity correlation which there was not much to be seen as you can see. The fastest mile scored .0098 and snow depth scored .0092 which were the most significant scores from the weather. These scores almost reached .01 so you can say there may be some correlation but not very much. This told me that probably extreme outliers in the weather for those categories problem had the most effect considering that if there is enough snow depth sometimes people can physically get outside or go somewhere. A few surprising finds in the heatmap were the domestic and hour features I did not think that they would be that great of predictors. Although the weather data seemed like it didn't have much correlation I still wanted to look so I compared the amount of days for each temp and the amount of crimes per temperature. As you may expect the graphs indeed did look pretty identical except for the ends of the graph where you see the weather data go out a little further again showing that the extremes of the weather cases may have a little effect. Going into the modeling I knew that there was not going to be much predictive value in the weather features. Although I did find other interesting trends in my data such as crime seems to be at its lowest point at 5am which it then over triples the amount of crime peaking at 8pm. Another interesting find was also the districts some districts appeared higher on my severe crime graph than my all crime graph meaning some districts tend to have more severe crime.


Weather Distribution
0.025
0.020
0.015
0.010
0.005
0.000
−20
0
20
40
60
80
100
Average Temp


Severe Crimes Disturbution
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
Average Temp


Severe Crimes per Hour
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


All crimes
hicago Lawn
Harrison
Englewood
Grand Central
Gresham
South Chicago
Grand Crossing
Deering
Near West
Wentworth
Town Hall
Calumet
Austin
Near North
Ogden
Shakespeare
Central
Jefferson Park
Morgan Park
Rogers Park
Albany Park
Lincoln


Severe Crimes
250000
200000
150000
100000
50000
0
Harrison
Englewood
Chicago Lawn
Gresham
Grand Central
Austin
South Chicago
Grand Crossing
Deering
Wentworth
Ogden
Calumet
Near West
Town Hall
Shakespeare
Near North
Central
Morgan Park
Rogers Park
Albany Park
Jefferson Park
Lincoln

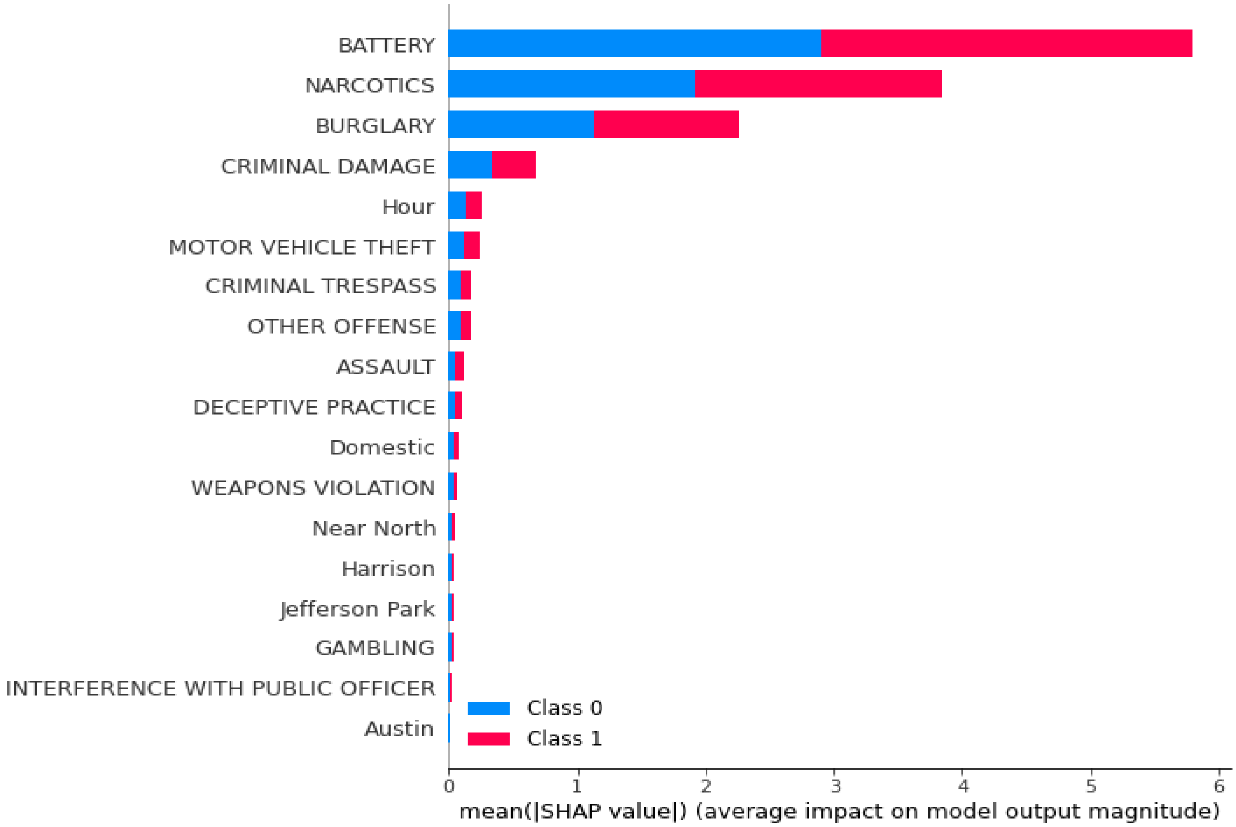
5. Machine Learning

Before I could start training and testing I had to create my dummy features first. I created dummy features for every feature that held names such as primary type, month name, day name, etc. I also removed the arrest feature since all arrests are considered severe crimes. Once all my features were right I created a dataframe that had 70,000 severe cases and 70,000 non severe cases from the original 6 million cases. Which I then used to create my train and test splits with the test size being 30%. After creating the splits I ran my my features through a function and took out all features that did not have a .008 to .8 correlation because all the other features would have little or too much impact on the model like the arrest feature. Once my splits ran through that I was ready to try out my different models. I used 3 models the LightGBM Classifier one being the best, the XGboost Classifier was pretty much the same, barely worse and the RandomForest Classifier was the worst out of the 3. I tuned the parameters by hand and found the best learning rate= .1, n_estimators=100, and max_depth=10. For some reason as well the RandomForest Classifier predictions had a lot less non severe crimes flagged as severe crimes than the other models but on the other hand it had way more severe crimes flagged as non severe crimes than the other models

1.LightGBM Classifier
roc/auc score: 0.8525
Matrix: ((20367, 588)
(5619, 15426))
2.XGB Classifier
roc/auc score: 0.8523
Matrix: ((20338, 617)
(5597, 15448))
3.RandomForest Classifier
roc/auc score: 0.8495
Matrix: ((20548, 407)
(5926, 15119))

6. Conclusions/Feature Importance

The idea that weather was going to help us predict severe crimes was a bit of a failure. The best correlations as seen above the values did not even reach .01 so I would say there does not seem to be much correlation between weather and crime at least in this model and dataset. With the feature importance the primary types that I turned into dummy variables are the best indicators which make sense because they are used to define the severity. Although we didn't see any weather variables to have high impact which I expected there were a few variables that were unexpected to me such as hour and domestic both were pretty good indicators for severity. Below all those we can see some district names pop up.



7. Future Improvements

Even though most findings between the weather and crime were basically non-existent in this model and dataset I think there can be multiple different approaches that we can go from here. One of the first improvements I would make is to get more cities with different weather data so we can compare 2 places on the same day with differing weather data. Maybe even add places like Seattle which has on average very high rainfall. Could also try an approach where you try to predict the amount of crimes in a day based on the weather. The feature location description I think you could also add back as there might be some correlation between where a crime might take place based on the weather. I also think that there could be some more data we could add to help predict crime such as maybe which politicians are in office at certain times or certain police chiefs. A little side note before I end the graphs for all crimes and severe crimes based on district would be great to help distribute funding to certain police stations in those communities which experience higher rates of crime.

Special thanks to Brooke Spodarek for helping with the idea for the project and also to my springboard mentor Rahul Sagrolikar who helped me on pretty much all of it.