Crime and Weather in Chicago

Yuval Shavit, David Barda, Aviv Yaish

Introduction

Problem Description

The city of Chicago is widely considered[1] one of the most dangerous in the US - Chicago had more homicides than any other US city in 2015[2], with homicide and shooting rates only increasing[3].

In contrast, Chicago's weather is less extreme. Compared to other places in the US, all four seasons are distinct, climate wise, with hot summers, cold winters, mild autumns and cool springs[4].

Although crime and weather might seem disparate at first, it is logical to assume that there is a close connection between the two. As such, our hypothesis is that **there is a connection between crime and weather in Chicago**, and this project is our attempt at evaluating this.

Data

The city of Chicago has many ambitious initiatives set on turning the 179 year old city into a young "smart" city. As part of these plans, the city collects a lot of data covering many different aspects of daily life in the municipality[5].

Among the data collected, Chicago's police department collected and published data on all crimes committed since 2001[6]. This huge dataset, weighing 1.5GB, contains info on more than $6,000,000$ crimes.

This dataset doesn't contain the weather data for each crime. Thankfully, NOAA, the National Oceanic and Atmospheric Administration, collects weather data for the US. But, they don't share this data publicly, and in order to obtain it we've had to go through their authorization process. After being authorized, we were granted access to Chicago's 2005-2016 weather data.

This dataset, weighing 38MB, contains at least one climate measurement per hour, and overall contains almost $150,000$ such measurements.

Processing

Both the crime and weather datasets, although excellent by themselves, require further work in order to be unified into one workable dataset containing both the crime data and the temperature for each crime. This was our main impediment, and we addressed it like so:

1. Both datasets contained missing values, so we've dropped them.
2. The weather dataset contained numerical values as strings, so we've converted them in order to facilitate compatibility with machine learning frameworks.
3. We supplemented the human-friendly time stamps provided in each dataset with a machine-friendly epoch time time stamp, and also with the name of the day and hour, which are useful when trying to group the data.
4. Then, we sorted the crime dataset according to time. The weather dataset was already sorted. Then, we removed all crimes committed before the first weather sample. The weather dataset contains weather samples covering the rest of the crimes (including the latest one).
5. For each crime, we looked for the weather sample taken closest to it, and attached the relevant temperature to it. In order to perform this efficiently, we've written functions using parallelization-ready methods provided by pandas.

These preparations result in a 1.4GB weighing dataset with more than $4,000,000$ records.

The code for this is provided in prepare_data.py. Instructions for running this are provided in the README.

General Statistics

The crime dataset is huge, and contains a lot of interesting information. Trying to visualize it can help at forming interesting hypotheses.

[1] "Trump mentioned Chicago's violence again. Police say they hope the president 'finally' sends help", Mark Berman, CNN, 2017. Very bad. We should build a wall and make Chicago pay for it.

[2] "Chicago violence, homicides and shootings up in 2015", Jeremy Gorner, Chicago Tribune, 2016

[3] "Chicago's 762 homicides in 2016 is highest in 19 years", Azadeh Ansari, CNN, 2017

[4] "Climate of Chicago", Various authors, Wikipedia

[5] City of Chicago Data Portal

[6] Crimes - 2001 to present, City of Chicago Data Portal

Let's for example look at the crime distribution among time, drilling down from months to hours. Looking at Figures 1, 2, 3, we can see that the crime distribution in the time period covered by our dataset isn't uniform among months, days or hours. Looking closer, it seems that there is a dip in crime during the cold seasons (specifically, during December - March). Drilling down to hours, there is a dip also during the cold hours of the late night-early morning. Of course, this is just an intuition, and the hypothesis needs to be rigorously tested.

It is also interesting to look at the distribution of crime types, as seen in Figure 4. Does the temperature affect this distribution? This is an interesting thought, that will be elaborated on in the next section.

Goals, Methods and Results

As a general goal, we wanted to find a connection between the weather conditions (specifically temperature) and crime in Chicago. We have decided to pursue this goal in multiple paths, trying to validate various hypotheses statistically, and trying to produce tools for the visualization of this connection which will be useful to the people of Chicago, and to Chicago's police-force and city council.

Goal 1

Goal: Find a connection between the temperature and the distribution of crimes.

Methods: We have formulated two hypotheses "classes":

1. If the temperature falls below X, does it affect the observed probability of a crime of a certain type Y happening?
2. If the temperature falls below X, does it affect the distribution of all crimes among the days of the week (Sunday, ..., Saturday)?

We have a certain "untreated" observed discrete random variable, and a "treated" observed random variable, where the treatment is the temperature being below a certain value, X. Both random variables receive the same values. For the first class each random variable has two possible values (two categories): either the crime type is Y, or it isn't. For the second class there are 7 - one for each day of the week. The χ^2 is a good pick to test the above hypotheses - it allows us to compare the two random variables, and check if the observed "treated" random variable's distribution fits with the distribution that is expected according to the "untreated" random variable.

We have performed the χ^2 hypothesis test on each class for each observed temperature (and for the first one also for every observed crime type). This includes performing all the necessary checks to verify that the hypothesis can indeed be used, as can be seen by the code.

The code for this is provided in hypothesis_test.py. Instructions for running this are provided in the README. Running the code outputs all necessary information, including:

1. The name of the hypothesis currently tested.
2. The various checks performed to verify that the test can be performed, how the checks are done and their results.
3. The result of the hypothesis test, if indeed we can perform it.
4. Supporting information.

Note that there are a lot of different observed temperatures and crime types, so running the code might take some time, and it is advised to output the results to a file, as there is a lot of text.

Uses: This might prove very useful for the police department of Chicago -

1. If indeed the temperature affects the probability of a crime of a certain type happening, the police can prepare itself accordingly and in advance to predicted temperature, for example by sending on patrols policemen who specialize in such a crime type.
2. If indeed the temperature affects the distribution of crimes among days, the police can reduce the active workforce on days with relatively fewer crimes, and increase it on days with relatively more crimes.

Results: There were many hypotheses being tested, and all results are included in the file /pdf/hypotheses_tests_output.txt. As examples, we'll go over one example for each of the hypotheses classes.

Example for class 2: Let's look at the crime distribution among days above and below -6.1 Celsius:

	Sun.	Mon.	Tues.	Wed.	Thu.	Fri.	Sat.
above	0.13	0.141	0.14	0.143	0.142	0.150	0.143
below	0.13	0.127	0.14	0.155	0.147	0.161	0.125

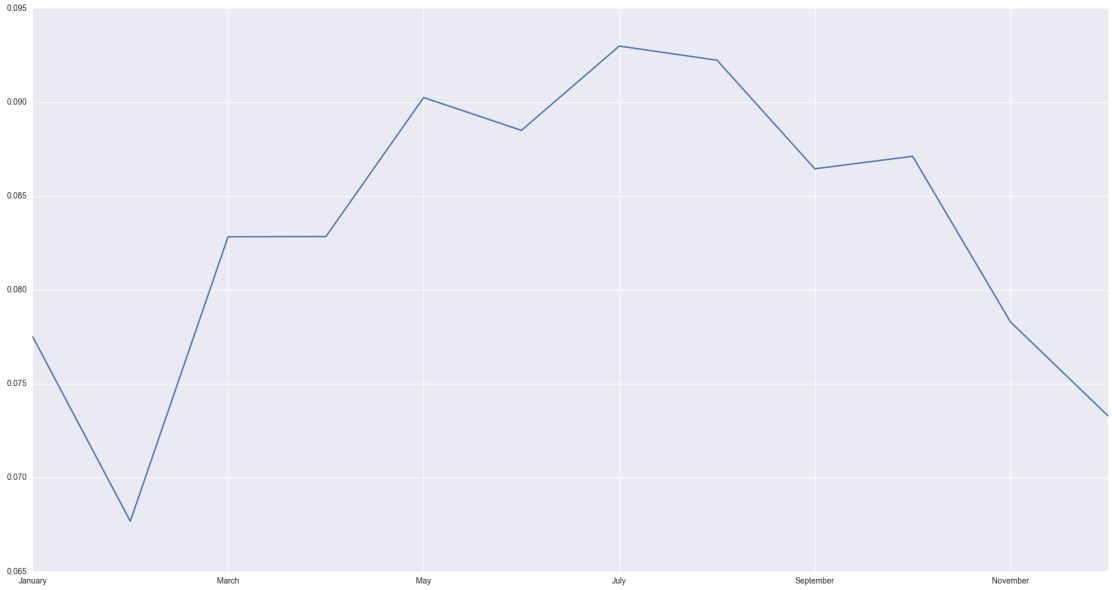

Figure 1: Distribution of crime among months

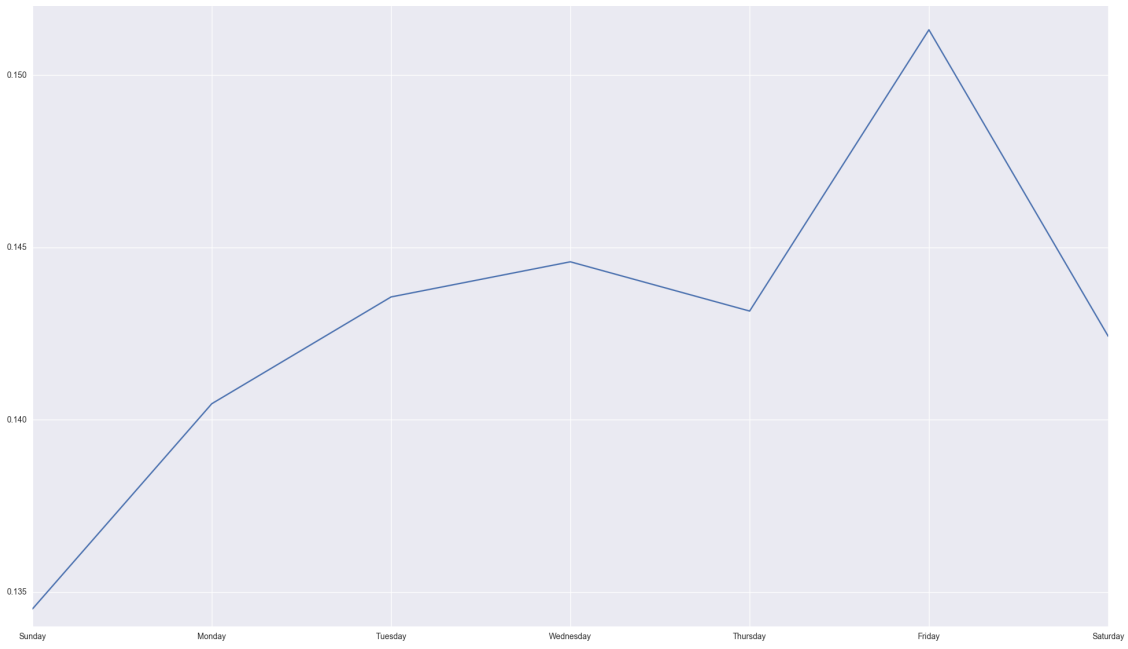

Figure 2: Drilling down to distribution of crime among days

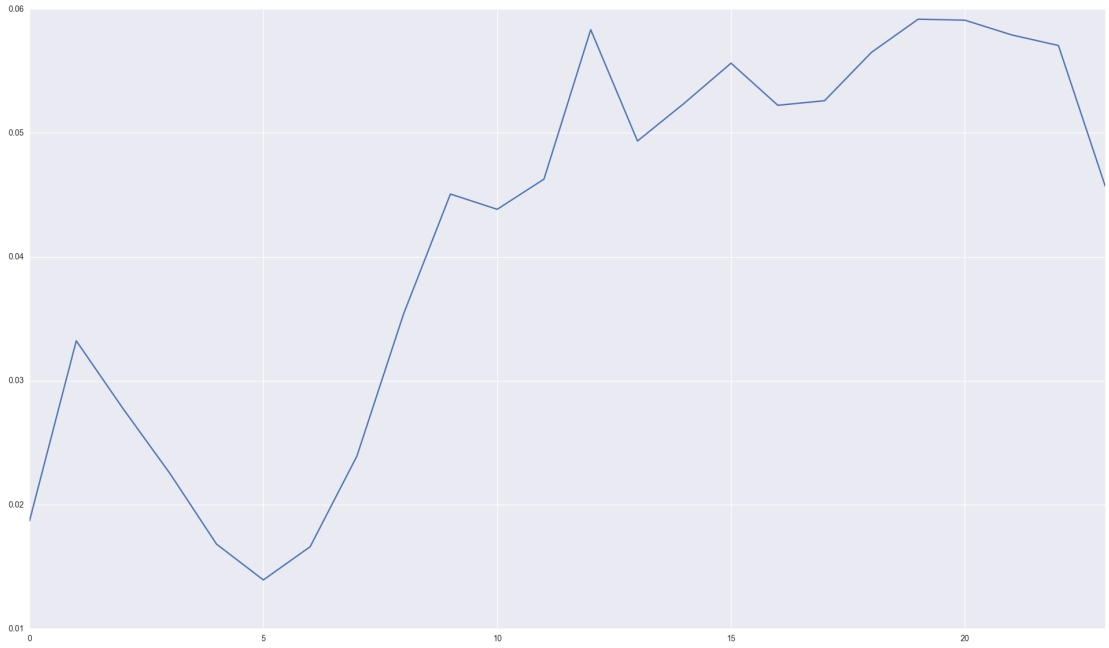

Figure 3: Drilling (further) down to distribution of crime among hours

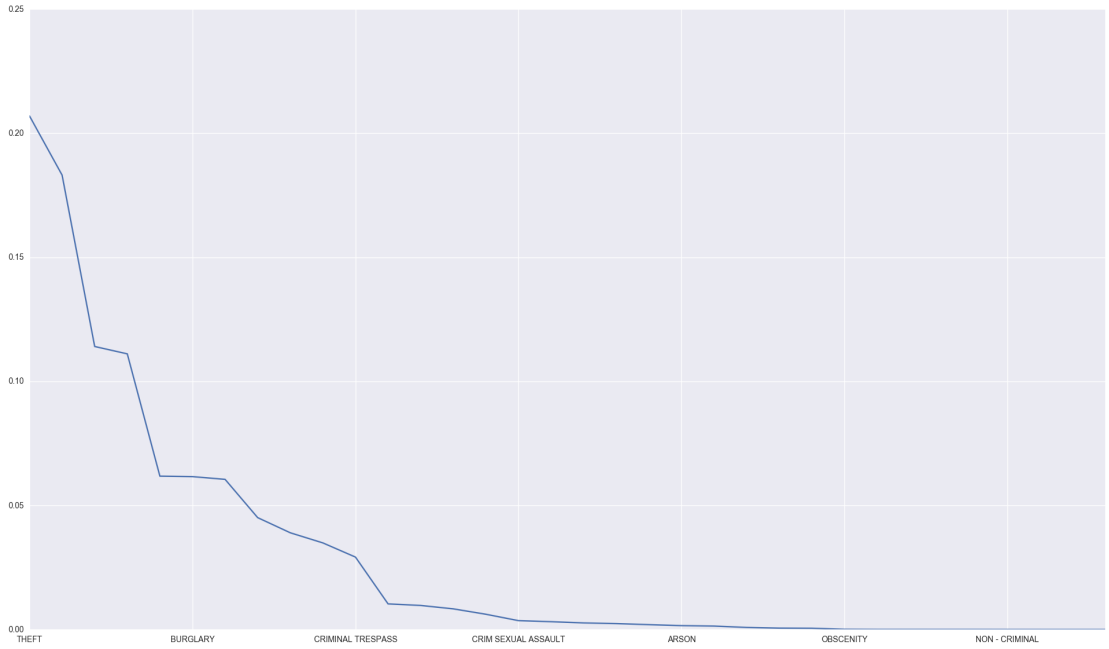

Figure 4: Distribution of crime types

We can see that there are three obvious changes between the distributions:

$$\frac{Friday - above}{Friday - below} = 0.93$$
$$\frac{Monday - below}{Monday - above} = 0.9$$
$$\frac{Saturday - below}{Saturday - above} = 0.87$$

So, there might be a reason for Chicago's police force to deploy their forces differently when it's colder or warmer than -6.1 Celsius. Let's test if the difference in distribution is indeed statistically significant using the χ^2 test with $\alpha = 0.05$. The test can be performed on the data, because the expected frequencies (the "untreated" distribution times the number of "treated" samples) are all above the minimal value for the test. The critical value for the test on this data is 12.59. The χ^2 value result received for the test was $1658.7 > 12.59$, with a p-value very close to $0 < 0.05$, so - the difference between the distributions is statistically significant.

Example for class 1: Let's look at the the percentage of crimes that are labeled 'THEFT' crimes of all crimes above and below 23.9 Celsius:

$$\text{above: } \frac{THEFT - above}{ALL - above} = 0.241025$$
$$\text{below: } \frac{THEFT - below}{ALL - below} = 0.200768$$
$$\Downarrow$$
$$\frac{THEFT - below}{THEFT - above} = 0.832$$

This is a pretty major reduction, and according to the hypothesis test, it is statistically significant: the test was performed with $\alpha = 0.05$. Again, the test can be performed on this data, because the expected frequencies are above the minimal value for the test. The critical value for the test on this data is 3.841. The χ^2 value result received for the test was $30760 > 3.841$, with a p-value very close to $0 < 0.05$.

Goal 2

Goal: Visualize the data in order to make it more understandable and useful.

Methods and Results: We've produced an interactive heat map of Chicago separated into it's community areas, where each community area's color is "hotter" if more crimes were committed there. The map presents only the crime count for the current temperature, which is controlled using a slider. To see this map, go to Figure 5

To create this map we've used OpenStreetMaps' map of Chicago, and the community areas JSON overlay was pulled again from the city of Chicago's data web site. In order to generate the heat map for each different temperature, we've used a nifty library called *folium* which allows automated generation of maps using python, and tin the end we've tied everything together using interactive JS code.

Here is a demonstration of the visualization.

Here is a link to the online visualization.

Note that each map weighs a few megabytes and that our server is bandwidth-constrained, so it might be a bit slow.

Uses: As before, this is useful for the citizens and police of Chicago - good visualization of the data facilitates better understanding of it. By understanding the three way relationship between crime, weather and location in Chicago, citizens can understand better when they should avoid certain areas, and the police can understand where to place more cops.

The code for the creation of the various maps is provided in viz.py, while the page which bundles them all together with the temperature slider is located in ./static/heatmap.html. Instructions for running this are provided in the README.

Goal 3

Goal: Predict the type of crime that will happen on a given location, during a given time, when the temperature is X.

Methods: Our problem is a multiclass prediction problem - predicting the crime type from a vector that looks like so: $\begin{bmatrix} \text{altitude} \\ \text{longitude} \\ \text{time} \\ \text{temperature} \end{bmatrix}$.

We have decided to try to perform the prediction using Machine Learning. Specifically, the Scikit library offers many useful tools for this. Using it, we tested the following classifiers:

1. Decision Tree Classifier
2. Naive Bayes Classifier for Multivariate Bernoulli Models
3. Gaussian Naive Bayes Classifier
4. Random Forest Classifier

Each trained and tested on the same sets, where the train-test split is $70\% - 30\%$. Note that all our data is labeled. Because of the size of the dataset, we've abandoned the idea of performing cross validation tests, as it would've taken many days of computation. After all the data was trained and tested, we picked the model with the highest accuracy, and pickled it for further use.

As there are 34 possible crime types, we thought it might be better to unify similar crime types under "umbrella" types, thus reducing the number of types and hopefully increasing the accuracy. But, this was deemed as unworthy - detailed information about the crime type is important for the regular police work.

Regarding the time feature, we've decided to remove the year from the time stamp, and convert all dates to values in $[1, 365]$ - meaning we retained the information about the day in the year of a crime.

Uses: As before, by producing a prediction model, Chicago's police can more accurately prepare itself for possible crimes.

Results: Because there are 34 different types of crime, the naive method of giving a random prediction will give an accuracy of $\frac{1}{34} = 0.029 = 2.9\%$, and this will be our baseline. The accuracy we received for each classifier was:

1. Random Forest - 25%
2. Gaussian Naive Bayes - 22.1%
3. Bernoulli Naive Bayes - 20.7%
4. Decision Tree - 20.5%

As you can see, the Random Forest one outperformed them all, with an accuracy 8.5 times better than the naive predictor.

Working with a Random Forest has it's benefit - after training a model, we can extract from it the importance it attaches to each feature of the input samples. The importances are:

1. Latitude - 0.293
2. Longitude - 0.279
3. Temperature - 0.214
4. Time and Day in Year - 0.212

Note that although the importances are very roughly equal, the model places most importance on the location, and less on the temperature and time. Although this is no concrete statistical proof that the location is more important than the others, it might be of value to continue research in the (logical) connection between location and crime in Chicago.

After choosing the best classifier for the task (as detailed in the Results section), we've pickled it and produced a website and a backend so that anyone could access our prediction engine. Given a request regarding a certain location, the backend pulls the current time and temperature in Chicago from various web services, and passes the data to the classifier. Then, it returns the crime type prediction to the user. You can see the GUI for the prediction engine on Figure 6.

When clicking on a point on the map, a prediction for the crime type that will happen on the time of the click and the temperature in Chicago at that time will pop up.

Here is a demonstration of the prediction engine (right after the demonstration of the visualization).

Here is a link to the online prediction engine demo.

The code for this is provided in machine_learning.py and server.py. Instructions for running this are provided in the README.

Future Work

Scientific future work

There is much future work possible to be done:

1. We saw that the Random Forest Classifier gives the geographic location a preference over the other features. This intuition requires further scientific work to prove that there is a connection between crime rate and location.
2. On the machine learning side of things:
   (a) We've tried "classic" machine learning. Today, Neural Networks are all the rage, and for good reason - they produce excellent results in some classification tasks. They, and other machine learning techniques, might also be of use here.
   (b) The "intuition" of the Random Forest classifier about the importance of location should be further explored: for example, it is possible to add location-type info from the map as features, like - was the crime committed in a park? Private residence? On the shoreline of a lake? In a pub? This might improve prediction accuracy - the crime type "LIQUOR LAW VIOLATION" probably is more frequent in pubs than in private residences.
   (c) Other possible features to use are the precipitation amount and type (rain/snow/etc'). It seems logical that there is less crime when it's snowing.

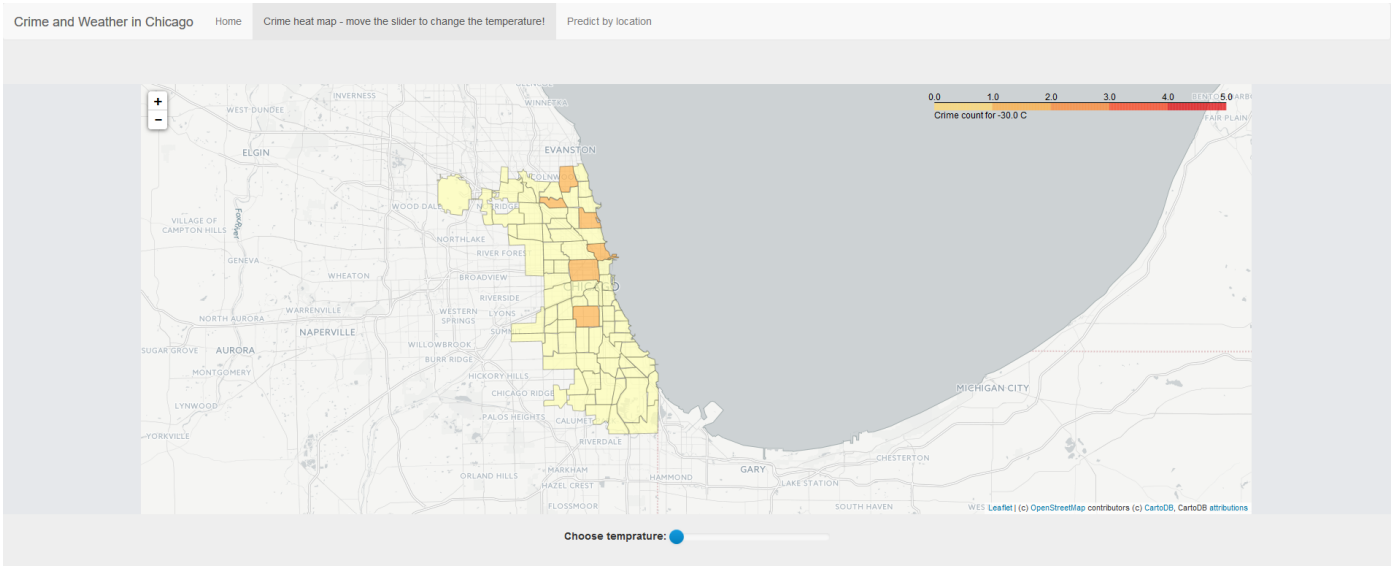

Figure 5: The GUI for our heatmap visualization

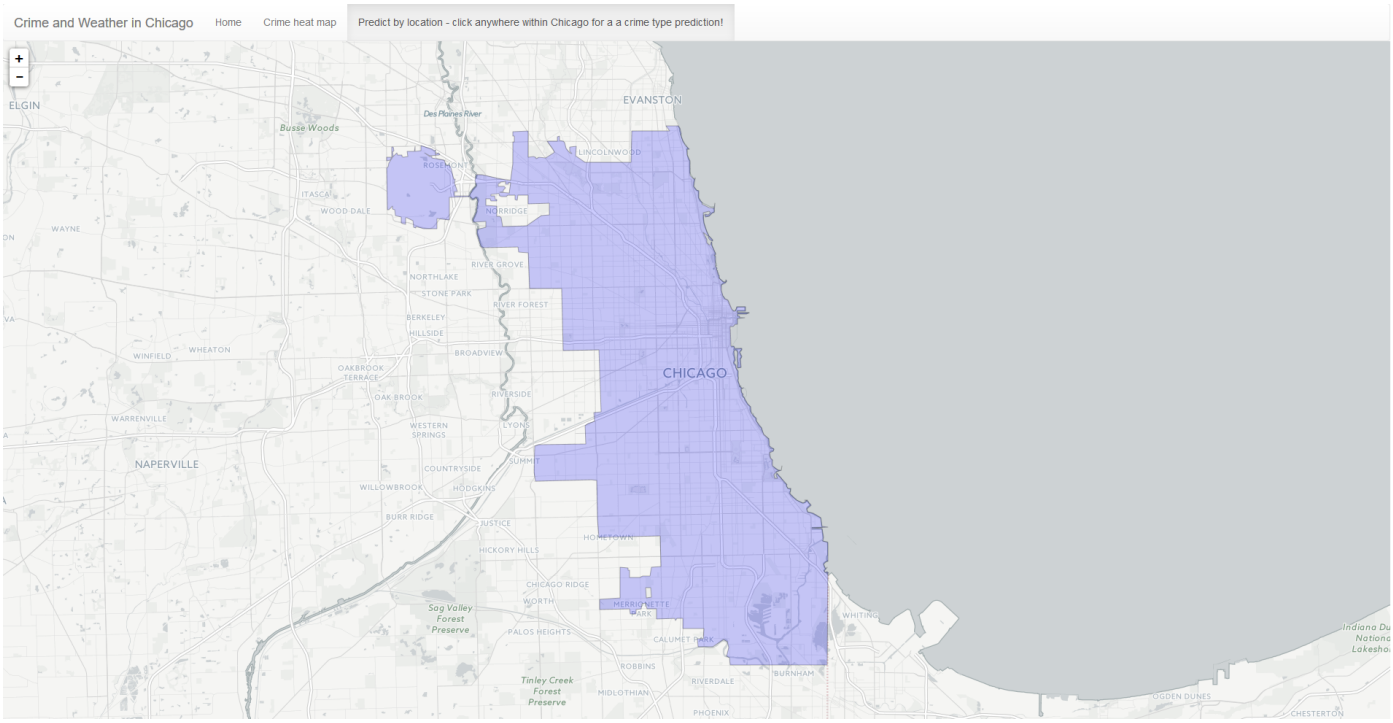

Figure 6: The GUI for our prediction engine

Social future work

Our data was made by humans, for use by humans, to help humans. We believe the power of big data should be used to help humanity, and as such we believe that although there is future work to be done on this data scientifically, maybe the more important future work to be done here should be civically, by Chicago's citizens, city council and police.

Again, maybe the importance of location for the Random Forest means something - maybe there are "dangerous" neighborhoods? Further work by the social services of the city of Chicago should be performed in order to better understand the relationship between crime and location in their city.

Conclusions

We have found that there **are** connections between weather and crime in Chicago. We have produced tools as demonstrated here to try to help the citizens and police force of Chicago deal with the severity of the crime there. Further work should be performed on the subject in order to better scientifically understand the relationship between crime and weather (and location), and hopefully eventually this will help protect citizens from crime, in Chicago, and elsewhere.

Acknowledgments

We would like to give our thanks to:

The City of Chicago Data Portal for publishing their data.

NOAA, for giving us access to their weather data.

The creators of Anaconda, folium, scikit, numpy, and all the other libraries we used, for producing excellent and useful libraries.