Weather or Not, There is Crime

Relationships between crime and weather patterns in Chicago

Jason Stock, Tom Cavey, Amber Lee

Colorado State University
Dr. Sangmi Pallickara
November 26, 2018

I. Problem

We aim to help the Chicago Police Department in strategizing the allocation of their officers and resources to better mitigate crime in the City of Chicago. Being able to anticipate certain categories of crimes during key weather events can help prepare officers retroactively, with the hopes of reducing crime rates. In addition to gaining a preemptive stance, we believe this data will provide logistical insights as well. Currently the Chicago PD consists of 12,244 officers and 1,925 employees [1]. There are 6.72 million crimes committed from 2001 to 2018, and 400 different crime codes which the Chicago PD use in their notation [3]. We believe the Chicago Police Department can utilize correlations in this data to better organize their officers in areas where needed, based on weather conditions.

This project will examine correlations in weather data, such as temperature, precipitation, humidity, and wind-chill with crime types and crime rates. The data is a set of weather station reports in Cook County as well as a complete criminal report of every crime in the city of Chicago dated from January 1st, 2001 to October 17th, 2018. Utilizing this crime and weather data obtained from surrounding area we aim to find a relationship between weather conditions and crime rates in the Chicago area. Our goal is to be able to analyze these trends and provide insight in which the Chicago Police Department can use to anticipate and prevent occurrence(s) of crime.

There are many factors which contribute to individuals being victim too, or committing crimes. Our project is important because we aim to uncover trends which can affect individuals livelihood who live in the City of Chicago. We hope that the correlations we find can reduce the number of crimes committed. We also think that the police department can benefit from this information, in knowing where they can place their officers and resources.

This is an interesting Big Data problem because we are looking at two datasets that typically are not joined together. Since we have tools such as Apache Hadoop's

MapReduce, we are able to look at these two large datasets and find correlations. We are hoping that our findings can be used by the Chicago Police Department, and people living in the Chicago area.

II. Methodology

Once we collected the raw data files, we used Apache MapReduce to comb through and pre-process the data. Our approach to pre processing was to first aggregate the weather and crime files. The weather files were in CSV format, and contained data from 2001-2018 and included multiple stations within the Chicago area. Each row was parsed and sorted by date allowing the reduce side join to compute the average of each chosen weather feature per day. The final result of the weather job produced a CSV file where each line contains the date and averages for the weather features for that day.

The crime files were also in CSV format, where each line contained a record which had been created as a result of a crime incident. The line was parsed and a filtering pattern was applied to collect the date, location, and desired crime data. Then each crime was sorted by date and crime type so they could easily be summed up. After summing the total number of crimes per day the output had to be formatted by each day so that each line contained the date, district, and the sum of every crime feature that we're considering for that day.

In order to run this data through a Machine Learning model, the weather and crime output files had to be aggregated. This was done by utilizing multiple inputs and a reduce side join. It was performed twice: once with a daily weather and crime output, the other with a weekly outlook by calculating the average of weather for the week, and summing the crimes for the week. Initially we had thought that aggregating the data by day was best, however we outline the reason we included week in the issues section of this report. This produced the appropriate file output for us to then apply our machine learning models. This sums up our total use of MapReduce for this project.

With the data prepared for the next step we began running our Machine Learning models. We experimented with many different types of frameworks, tools, and models. Our machine learning algorithms include Scaled Conjugate Gradient, Stochastic Gradient Descent, Adam (PyTorch), Linear Regression, Ridge, Lasso, SVR (linear, polynomial, SBR), and BoostedTreeRegressor (CoreML). We used both Python and Swift to run these algorithms to produce our models. Our frameworks include Apache MapReduce, Apache Hadoop, PyTorch, Python, CoreML, and Scikit-learn.

GitHub was used for collaborating and version control. This allowed us to easily share our output files and correspond on code. Jupyter notebooks were used to run Python and machine learning algorithms. Xcode was used to run the CoreML framework in Swift which allowed for the use of the BoostedTreeRegressor machine learning algorithm. IntelliJ and Atom were used in the MapReduce programming with Java. All of these tools were utilized to assist collaboration and help share work with one another.

III. Data

The dataset includes eight files of Local Climatological Data (LCD) from the National Oceanic Atmospheric Administration. This data was pulled from four weather stations located within Cook County. The link for this data can be found at: https://www.ncdc.noaa.gov/cdo-web/datatools/lcd. [2]

Weather stations include the following locations, Chicago Lansing Municipal Airport, IL US; Chicago Palwaukee Airport, IL US; Chicago Midway Airport, IL US, and Chicago O'Hare International Airport, IL US.
Each weather dataset reports hourly on various weather attributes from January 2001 to present. Each row includes a station ID, station name, elevation, latitude, longitude, date, report type, and hourly dry bulb temps, wet bulb temps, relative humidity, and wind speed, among several other attributes that we deemed irrelevant for our project.

Other attributes within the weather sets were missing large sections of values or would have otherwise contributed to noise.

The criminal dataset is a 1.6 GB file reporting on every crime reported from the Chicago Police Department from January 2001 to present. Data is extracted from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR) system [3]. There are 6.72 million rows of crime incidents within this file. Data was downloaded from: https://catalog.data.gov/dataset?res_format=CSV&tags=crime&page=3 [4].

Each row includes an ID, District/Community Location, Date/Time, IUCR (Illinois Uniform Crime Reporting) codes, Primary type of crime, and a description of crime among other attributes that were deemed irrelevant for this project. IUCR codes are used to aggregate types of criminal cases for statistical purposes [5].

After an initial sweep of the data, as well as a discussion on which variables should be considered when attempting to correlate weather with crime rates, we chose to join our weather datasets on date, and use daily dry bulb temperatures, daily wet bulb temperatures, daily wind speed and daily relative humidity for our weather attributes.

Chicago is a large city with unique neighborhoods defined by social-economic factors. It made sense to attempt to categorize and correlate crimes in each individual district [6]. There are roughly 50 primary crime categories. For the scope of this project, we chose to limit crime categories to those that we felt had the most negative impact on victims or those which demanded more police resources.

IV. Analysis

Prior to the data being reduced and aggregated there was difficulty interpreting correlations between the weather and crime rates. This was primarily due to the large quantity of samples that spanned nearly 17 years. Once merged and reduced, the help

of statistical models, numerical analysis, and machine learning algorithms assisted in better understanding the data.

Analysis of the features were seen by numerically and visually computing distribution and probability densities across different inputs and target outputs. It was seen that certain weather characteristics had a large standard deviation and while others did not. This is important because characteristics such as dry / wet bulb temperature and humidity extend a greater range of values while wind speed did not. Wind speeds shifted plus or minus two miles per hours which indicates that the majority of the samples are consistent. Whereas the temperatures had a higher variances between samples, and can be used to help distinguish which inputs may or may not be be of more importance for the machine learning algorithms.

Since temperature has the highest variance the individual crimes could be plotted against the temperature to find a linear line of best fit. A line that shows a positive or negative tangent relates to that crime having a wider variation in quantity of crimes over temperature ranges. Contrary to this, a slope of zero or undefined displays little to no variation in crime rates as the temperature changes. Specifically, battery, assault, and theft show the steepest tangent over all the districts for all the weather conditions, i.e., the rate of crime varies the most over different conditions. As a result, this allows better understanding of how machine learning algorithms react to the data, and let conclusions be made.

Various frameworks from PyTorch, CoreML, and Scikit-Learn were used to compute regression analyses; however, a naive Python implementation of scaled conjugate gradient (SCG) outperformed the former. Since there were nine unique districts, corresponding models were computed to minimize biased noise. SCG is a supervised neural network algorithm that requires deep exploration to numerous networks to find one that best suits the data, hence, the need of various models for each district. Finding the most accurate model was done by computing the root mean squared error

compared across trials. This approach was quite successful, and largely possible because of the formulation and preprocessing of data.

It was found that by predicting the most accurate models for each district the Pearson's product-moment correlation coefficients of each crime category can confirm the findings in the feature analysis. Specifically, the top networks that performed the best displayed a positive correlation of 0.73, 0.65, and 0.55 for assault, battery, and theft respectively. Thus, the trained neural network structures for each district could be serialized and saved such that they can be reused to predict criminal activity given weather conditions. This means that for a given week, the trained model can confidently predict specific crime rates across the city of Chicago.

V. Challenges

During debug of the MapReduce code that we used to join our weather and crime datasets, we discovered that there were some missing values for the crime community identifier, specifically in the year 2001. Although it was not discovered how extensive this was until other challenges arose as the project progressed.

Initially we aggregated our data based on daily crime rates. Once we had our data, we began work on neural nets. We found that we were not seeing any correlations. Histograms were plotted so that we could more closely look at the data by year. The plotted histograms helped us to discover that we were dealing with sparse data and that many categories contained values of zero. In other words, the crime rates were scattered and our model had too much variation to decipher between. This prompted us to rewrite our MapReduce code to aggregate crimes on weekly averages and rates versus daily averages and rates.

It was discovered then that we needed to further modify our MapReduce code to ensure that we were accurately reporting the correct week number. In order to redesign the MapReduce code to reflect the number of crimes per week than day, a

few changes had to be made. One challenge we ran into when making this change was how too aggregate the number of days to correspond to the appropriate week. It seemed to be working fine by counting the week number, however it proved to be much simpler to utilize the Java.time package included in Java 1.8. Once calculated, the year was appended to each week number. The weather features were averaged and the crime features were summed for the week. Finally, after aggregating the crime by week number and year each district had better looking data for our model.

Histograms were replotted and we then discovered the extensivity of the missing crime data in the first 70 weeks. This caused extreme variance with the histograms showing large gaps with no data inter-dispersed with large crime rates. We discovered that the crime data set was missing values for community number in these early weeks and therefore decided to truncate our dataset to exclude the poorly documented weeks.

VI. Conclusion

Chicago is a diverse city with each community showing a unique ecology of crime. Therefore we trained neural networks on 9 separate districts or communities within the city. While we discovered some moderate to strong correlations between certain crime rates and our chosen weather attributes, we also discovered many correlations that were weak or at times close to zero. The findings were informative in that some of the districts with the highest rates of crime, such as District 7, known as the south side of Chicago by name, were more weakly correlated with weather than others examined. In other words, in some of the worst parts of the city for crime rates, temperature and humidity does not seem to play as prevalent a role in rates of crime.

A data summary is provided in the table below. It is important to note that the test data is selected randomly by our algorithm and the Pearson's product-moment correlation coefficients, R-values, may vary between runs. This represents the actual values of crime rates against those predicted for each category. Below is a snapshot of a run on each district.

Reported RMSE is the average root mean square error of our test data on each network. We strove to find networks in which RMSE of test data was near equal to our training data to prevent over-fitting or under-fitting our training model. As discussed above, we were able to run more complex networks and lower our RMSE on our training data significantly, but found that we were over fitting our model to the data. In these cases, RMSE on our test data would increase significantly and restrict our predictive capabilities.

In general, a larger RMSE value indicates a less linear relationship between actual vs. predicted. This was seen and noted in our scatter plots. By comparing RMSE values across each district we can conclude that the districts which produced neural nets with lower RMSE values are doing a more accurate job and associated R-values are more statistically reliable.

In general, comparing actual versus predicted values, we saw a few values near or just above 0.7 which would indicate a strong correlation. Overall, 5 of 9 districts had Theft, Assault, and Battery in the top 3 correlated to our chosen weather attributes, with Burglary and Robbery close behind. Theft made top 3 for every district. Weapons Violation, Motor Vehicle Theft, and Homicide showed weak correlation, if not close to zero in some districts.

This summary concludes that we can use a neural net and machine learning algorithms to predict rates of crimes in certain categories for each district, with some districts more correlated than others. In general we note that Theft, Assault, Battery, and Robbery are most predictable according to our model.

District	RMSE	Assault	Theft	Battery	Robbery	Burglary	Auto Theft	Weapons	Homicide
1 (Far North Side)	21.64	0.162	**0.402**	0.171	0.322	0.322	0.295	0.165	0.095
2 (North Side)	18.69	0.163	**0.442**	0.254	0.293	0.254	0.048	0.136	0.024
3 (NW Side)	18.69	**0.392**	0.205	0.325	0.058	0.072	0.149	0.099	0.052
4 (Central)	13.65	0.202	**0.581**	0.330	0.282	0.203	0.095	0.140	0.053
5 (West)	36.28	**0.731**	0.553	0.651	0.421	0.392	0.319	0.270	0.209
6 (SW Side)	29.74	**0.542**	0.275	0.515	0.303	0.123	0.074	0.335	0.273
7 (South Side)	5.43	0.351	0.326	**0.405**	0.209	0.209	0.123	0.199	0.106
8 (Far SW Side)	11.57	**0.592**	0.458	0.494	0.136	0.299	0.043	0.358	0.249
9 (Far West Side)	19.76	**0.617**	0.461	0.524	0.246	0.174	0.144	0.459	0.093

VII. Contributions

Amber Lee-

Collaboratively researched datasets, brain-stormed project ideas, and helped formulate and write project proposal with Jason and Tom. Participated in team meeting to outline, review data, and refine our plans for MapReduce patterns and specify our approach to train neural nets. I specifically wrote the MapReduce code to filter, aggregate, and sum crime rates on the crime dataset. Then I began working on an investigation of neural nets using a Scaled Conjugate Gradient algorithm as well as running trainings with PyTorch to see if various optimizers would improve RMSE. This investigation, along with a review of other team members work helped to identify that aggregating data daily was not a good approach for our dataset. I helped to identify that we should try to aggregate our data by week vs day and then continued investigation of neural net code utilizing SCG algorithm as PyTorch did not offer any improvement on this code for correlations. I focused in on identifying a way to quantify our correlations outside of using scatter plots to visualize our data. Helping to conclude that Pearson's coefficient could be used to show how strongly correlated our data was.

I, Jason Stock, contributed by writing various aspects of the project. In collaboration with my other teammates, we explored various datasets where I then found myself focused on the weather data to gather samples across the surrounding Chicago counties. Once the data was retrieved, I continued with writing the MapReduce job to conglomerate all the weather samples to one file averaged over the city by day. After all the pieces were in play and the final aggregation of crime and weather data were accessible the analysis and models could be built. I explored various Python modules from Seaborn, Matplotlib, Scikit-Learn, and other machine learning algorithms to compare regression models for our data. It was seen that our own scaled conjugate gradient algorithm outperformed the former, and I proceeded to further analyze this. In order to better the results and performance of the neural networks, feature analysis of both the inputs and outputs were computed. This helped display the results to my other teammates, and collaborate to make conclusions.

I, Tom Cavey, contributed to the project in a variety of ways. First, I helped brainstorm project ideas, researched datasets, and engaged in discussions which defined the project scope. Then, participating in team meetings which outlined the layout of the code of our project in terms of how the MapReduce jobs will work, what we want the output to look like, and what visuals and graphs we want to try and create. Researched the background information from the Chicago PD and related websites. I wrote the MapReduce Aggregator code for both the daily aggregator and the weekly aggregator. Along with debugging various aspects of the code and how we want the output file formatted. I set up and executed the Xcode CoreML machine learning model in Swift. This included implementing the LinearRegressor, BoostedTreeRegressor, RandomForest, and DecisionTree regressor cases. Although there were no graphs for this data, just the RMSE and other related metrics. I was helpful in debugging and figuring out the problems which we encountered during the use of the output files into the machine learning algorithms used within the Jupyter notebooks.

VIII.Bibliography

[1] Chicago Police Department. (2018, October 22). Retrieved from https://en.wikipedia.org/wiki/Chicago_Police_Department

[2] National Centers for Environmental Information, & NCEI. (n.d.). Data Tools: Local Climatological Data (LCD). Retrieved from https://www.ncdc.noaa.gov/cdo-web/datatools/lcd

[3] (n.d.). Retrieved from https://dev.socrata.com/foundry/data.cityofchicago.org/6zsd-86xi

[4] Data Catalog. (n.d.). Retrieved from https://catalog.data.gov/dataset?res_format=CSV&tags=crime&page=3

[5] Chicago Police Department - Illinois Uniform Crime Reporting (IUCR) Codes | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e

[6] Chicago Neighborhoods. (n.d.). Retrieved from http://www.thechicago77.com/chicago-neighborhoods/