EECS 349 Project Final Report

Predicting Crimes In Chicago From Weather

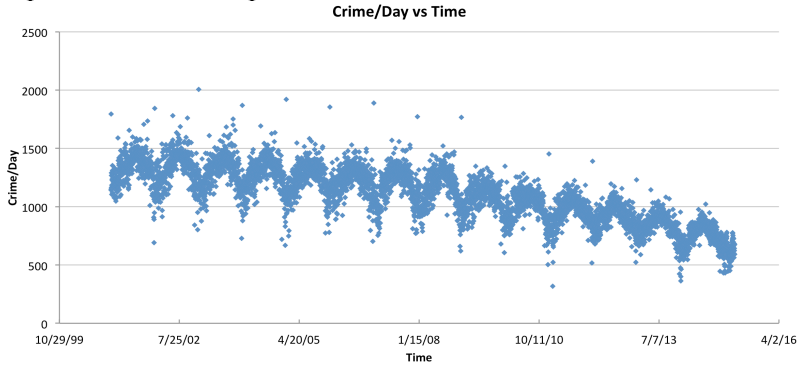
Alan Fu | Northwestern University EECS 349 | 2015-06-08

I. Problem

In this project, my task is to predict the crime rate in Chicago in a certain day given the weather parameters in that day. In a more fundamental level, I wish to investigate the relationship between weather and crime rate, if there is any. The motivation for this project emerges from my experience with Chicago's weather since my attendance at Northwestern. I have noticed significant effect of weather on my mood and behavior, so have my close friends in Northwestern. I wish to further explore this effect in a more quantified and rigorous manner, and I decided to use crimes as the specific measurement of human behavior (specifically, how violent human turn to each other). Important applications may arise if the program can predict crimes in Chicago in reasonable accuracy from weather. The effect of weather on crime can be a valuable addition to the existing crime-predicting tools such as *PREDPOL*[1] to achieve greater accuracy in predicting crimes, enabling law enforcement agencies to stop crimes better, faster, or even before they happen. The prediction made by the program can also be used in other psychological or sociological studies on weather and human behaviors to gain further insights.

II. Data Preparation

Two main sets of data are used for this project, one is the weather data recorded at the Midway Airport,[2] and the other is the crime data from *City of Chicago Data Portal.*[3] Initially the raw weather data contains around 144,971 hourly recordings over the period of 2001-01-01 to 2015-03-29, each recording contains the values of timestamp, temperature, dew point temperature, humidity, wind direction, wind speed, precipitation, atmospheric pressure and other miscellaneous variables. To prepare the weather data for training and testing, first all the miscellaneous variables that are deemed certainly irrelevant are deleted (leaving the variables I mentioned above), then the recordings are consolidated into daily averages. The raw crime data initially contains 5,765,192 instances of crime in Chicago over the period of 2001-01-01 to 2015-03-29; each recording contains the value of timestamp, crime type, arrested, and other miscellaneous variables such as crime ID. To prepare the crime data for training and testing, all the irrelevant variables are first deleted. Then the numbers of instances of crimes in total happened in each day are counted, along with the crime counts for each individual type of the total 35 different types of crime. To further consolidate the data, all the crimes are categorized into the following 11 general types: sex offense, theft, assault and battery, burglary, rubbery, substance, homicide, gambling, prostitution, arson, and others. At this point the total crime counts each day over time is plotted as the following graph to gain a general grasp of crime count's patterns.

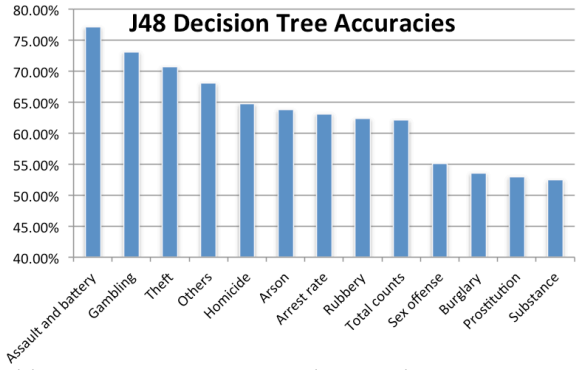


[1] https://www.predpol.com/technology/
[2] http://mesonet.agron.iastate.edu/request/download.phtml?network=IL_ASOS
[3] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Clearly the graph resembles the pattern of a sinusoidal wave with downward-sloping horizontal axis, which tells two things: first, crime count apparently oscillates periodically on a yearly basis. This is encouraging because temperature also follows the same pattern throughout the year. Second, the absolute counts of crime has been steadily decreasing over the past 15 years, which means that absolute crime count is not the best variable to predict. To counter this problem, all the crime counts variables (total counts and counts for each type) are changed to binary variables: 1 meaning the count is lower than or equal to the median count in the 180-day-before to 180-day-after period and 2 means higher. This way the effect of external variables such as improving education and government policies on crime counts is mitigated, and our data is ready for decision tree modeling.

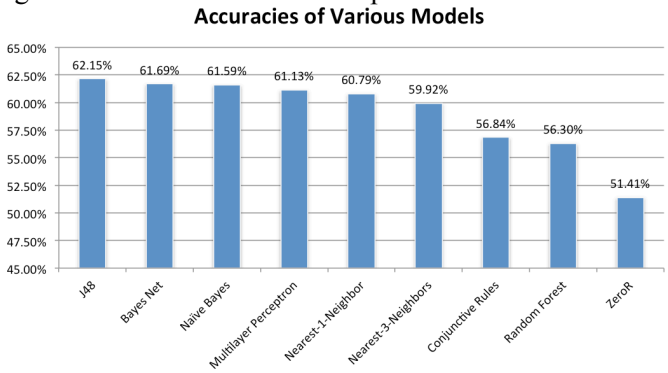
III. Modeling and Testing

Two packages are utilized for modeling and testing, one is the decision tree classifier offered in the python library *scikit learn,*[4] and the other is *weka*. Using 10-fold-cross-validation, the decision tree model in *scikit learn* can only achieve an accuracy of 55.03% when predicting total counts while the J48 model offered by *Weka* can achieve an accuracy of 62.15%.[5] Therefore *Weka* J48 is chosen as the primary model. The accuracies of using the same J48 model to predict the count of each type of crime is plotted as below:



As we can see, the assault and battery type counts per day can be most accurately predicted from weather parameters (an accuracy nearly 78%.). Consider how many potential factors there are that might affect crime rates in a city in a given time, we can consider there to be a reasonably strong connection between weather and crime rates. To test whether crime counts are more related to something else that is correlated to time of the year other than weather, the same J48 model is used to predict total crime counts solely from the month of the year, achieving an accuracy of only 60.5734%. Therefore, we can tell that weather is the stronger predictor for crime counts instead of time of the year.

In order to confirm the selection of J48 as the primary model for this project, a range of other machine learning models offered in *Weka* are explored to predict the total crime counts. The accuracies of each of these models using 10-fold-cross-validation are plotted as below:

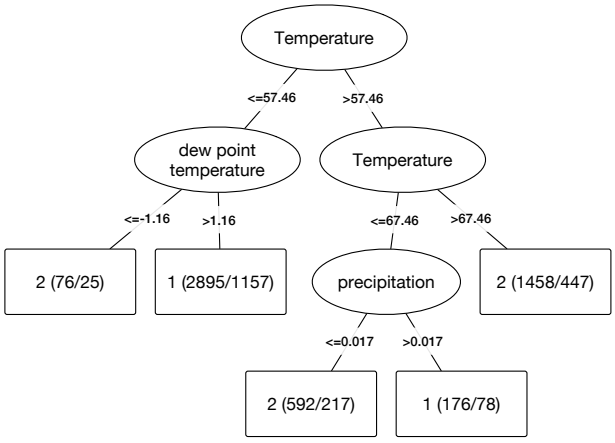


[4] http://scikit-learn.org/stable/modules/tree.html
[5] Configuration: weka.classifiers.trees.J48 -C 0.2 -M 2

As shown, J48 has the highest accuracies among all models explored (though still not great). Relatively accurate and easy to interpret, J48 is clearly the best model for this task.

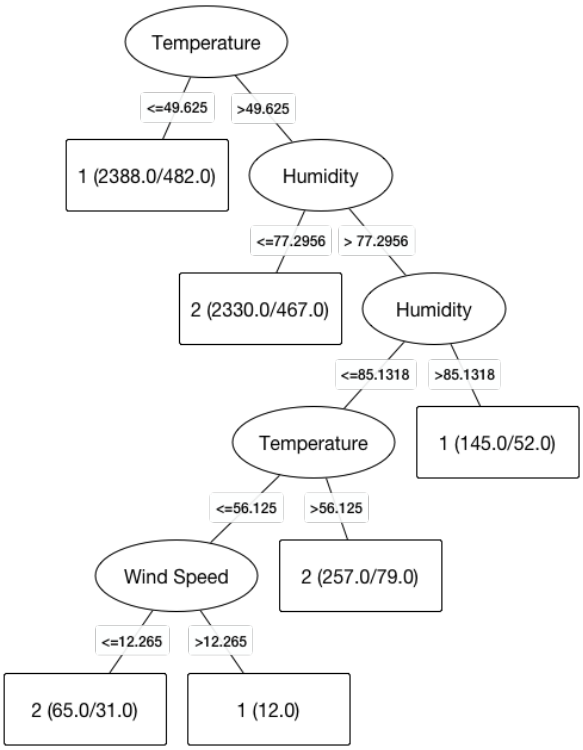
IV. Results And Analysis

The decision tree for total counts can be visualized as below:



As seen in this simple tree, temperature shows the most dominant relationship with total crime counts in a given day. Contrary to as many would expect, crime counts do not simply increase as temperature increase, but are likely to be higher than yearly average when the temperature is either too high or too low. This insight matches the research results shown in a 2013 New York Times article,[6] which concludes "episodes of extreme climate make people more violent toward one another". While temperature is within a "comfort range", precipitation steps in and predicts that the higher the precipitation, the more likely crime counts will be lower than yearly average.

The decision tree for assault and battery counts (the most accurately-predicted counts) can be visualized as below:



[6] http://www.nytimes.com/2013/09/01/opinion/sunday/weather-and-violence.html?_r=0

This tree is slightly more complicated than the one for total counts. Temperature is still the most dominant variable in this tree, while humidity and wind speed also enter the picture. The general prediction of this tree is that assault and battery crimes are less likely to happen in the weather of low temperature, high humidity, or high wind speed.

The types of crimes that can be most poorly predicted from weather parameters are prostitutions and substance, which is reasonable because the majority cases of these two types of crimes take place indoor; moreover, prostitute customers' need for sex and addicts' need for drugs are highly inelastic under any circumstance.

V. Suggestions For Future Work

In this project the variable being predicted (crime counts) is reduced to the simplest possible binary form (lower or higher comparing to yearly average), which might not contain enough information to be useful for other academic research or crime-predicting tools. A carefully hand-tuned Bayes network might be able to resolve this problem by predicting the full probability distribution of crime counts in a given day, instead of simply how the crime counts will compare to the yearly average. The raw crime data also contains the district names in where each crime took place. To further enhance this project's usefulness for the police, learning can be applied to each specific district to provide more relevant predictions for district local police force. At the end, however, to predict crime counts with substantially higher accuracy regardless of which machine-learning algorithm is used, more variables other than weather parameters have to be thrown in because crime is a complicated human behavior after all.