Crime in Chicago

A research project on predicting crimes in Chicago using Business Intelligence models such as Linear Regression and Classification

CAIO DINI
THOMAS HAYES HICKEN
JUN JIN
MURILO GUSTINELI
PHILLIP LEAL

Introduction

Being the third largest city in the United States and a major world financial center, Chicago is not only known for its culture such as deep-dish pizza but also its high crime rate. Chicago is experiencing violent crimes, and people are calling it Chiraq, equating it to a war zone. To explore factors that might be related to the crime, we found an extensive dataset of crimes in Chicago from 2001 to 2017 extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The dataset contained various information, including Case Number, Date, Primary Type, Description, Community Area, FBI Code, Location, and so on. We wanted to see if we could predict crime based on different factors and further find ways to apply these findings to business ideas.

Data Preparation

Choosing relevant data

Data selection is not only an important data analysis process, but also absolutely necessary for accurate results. Thus, the process of selecting suitable data for a research project can impact data integrity. The main objective of data selection is the determination of appropriate data type, source, and method(s) that allow analysts to appropriately answer research questions. In order to achieve accurate results, we had to condense some fields in our data set to eliminate ambiguity and irrelevancy. Fields such as crime ID and case number were unnecessary for our analysis in predicting crime rate. Also, most of the location fields meant the same thing, so we just kept the latitude and longitude to eliminate replication. We started with 8 million records. After eliminating records with invalid data or missing data, we ended up with 7.8 million records. Moreover, we decided to focus on violent crimes because they seemed to be more relevant for the research project. Assault, Battery, Sexual assault, Homicide, Intimidation, Kidnaping, Robbery, and Sex offense are all violent crimes according to the FBI. We dropped minor and regular crimes, resulting into 2,2 million violent crimes over the course of 16 years in Chicago.

Our model did not work. Now what?

Spoiler alert. Crime, date, and the location were not enough to predict where and when crimes were going to happen. Day of the week or day of the month showed no correlation with crime. We needed to come up with a new approach to this problem. We came across a paper arguing that ice cream sales lead to higher homicide rates. That's not because criminals like ice cream. It is due to higher temperatures. When the weather gets warmer, more people go out in the streets, leading to more general violence, and more homicides. We got the temperature data from the National Center for Environmental Information (part of NOAA) of every day for the past 16 years and attached to our data set. The results were very promising. Refer to the Model section below for more in depth explanation.

Model

Multilinear Regression

To approach predicting crime we started with a Multilinear Regression. In the data set we had three continuous variables (latitude, longitude, and max high temperature) and a continuous result that we were trying to predict (number of crimes). Using the sklearn Multilinear Regression package we produced our first model. Looking at the residuals we had a Mean Squared Error (MSE) of 900 and a Variance score of .06 or 6% of the variance could be explained by the model. This wasn't much better than just taking the average. We needed to approach the problem differently. After speaking with the group we determined that the number of crimes was really irrelevant without some sort of benchmark. Was six crimes good, normal, bad, or anarchy? We needed a better approach to the problem.

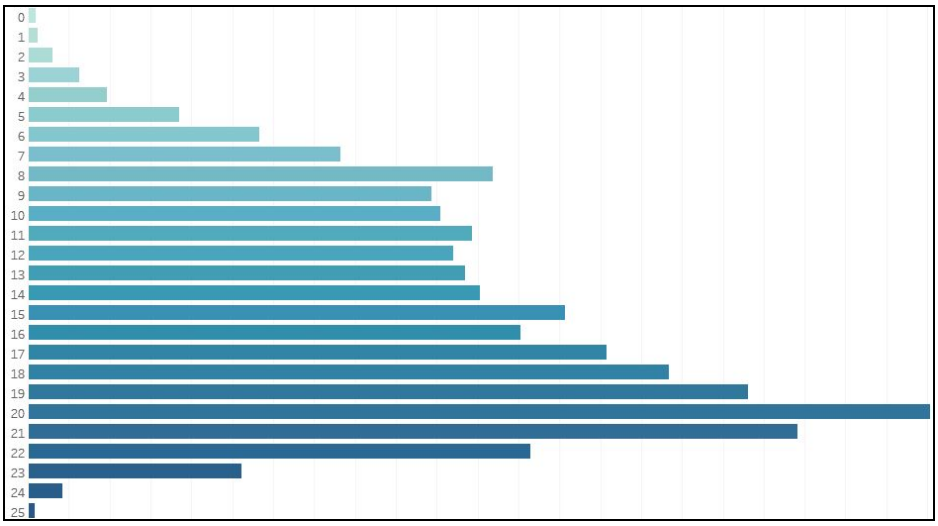
Classification

Instead of a regression problem, this was a classification problem. We determined that we should bucket the number of crimes into low, normal, and high. We determined each bucket by taking the quadriles of the data set and determining the buckets based on which quadrille the number of crimes resided. It was determined that "High" would be in quadrille 4, "Low" would be in quadrille 1, and "Normal" would be quadrille 2 and 3.

We also wanted to bucket the variables to make the classification easier. We split the latitudes and longitudes of the data set into a 100 x 100 matrix. For validation, we made a heat map of the data to see what it looked like:

Our heatmap looked like Chicago, even down to the void in the bottom corner which is a lake. Satisfied with the results, we moved on to temperature. Splitting the city into different sections did something else. One of the issues we had was the lack of variables for the prediction. There is an old saying “Birds of a feather flock together” so by segmenting the area we also segmented the people in that area and captured the demographics of that area. This includes ethnicity, household income, average age, and even education level. These variables were embedded into each 100x100 block.

We split temperature into 4 degree buckets and plotted that just to be sure:



Seeing that the graph looked like we would expect, we moved on to the Classification model. Using the DesissionTreeClassifier in the sklearn package, we did a 60/40 split on the train/test set. Having such a large dataset allowed us to have a larger test set to validate against. We decided to start at 4 branches. Here were the results:

	Precision	Recall	F1-Score
High	.54	.43	.48
Low	.60	.64	.62
Normal	.38	.40	.39
Accuracy: .51			

This was better than the MLR. We could do better still. First we needed to determine the best number of branches to use. So we wrote a python program that looped and made models based on an increasing number of branches then testing that based on our test set. What we found was 8 was an ideal number where the model was not overfit.

Lastly, we looked if each outcome was equally desired. Determining being able to predict a high crime area more important than just predicting that the crime was normal. Same thing about Low. Being that these were the two extremes, being in the top and bottom 25%, we changed the weights to High(5), Normal(1), and Low(4).

So after running the model with the adjusted branches and weights, the results were:

	Precision	Recall	F1-Score
High	.45	.74	.56
Low	.56	.77	.65
Normal	.00	.00	.00
Accuracy: .51			

And applying the model to the test set:

	Precision	Recall	F1-Score
High	.45	.74	.56
Low	.55	.77	.64
Normal	.00	.00	.00
Accuracy: .51			

Results

By giving high crime and low crime areas more weight, we increased the accuracy of the model just enough to get it over 50%. Although this change was only slightly better, we saw a significant increase in the model's recall ability. The model is now ready to produce correct information 77 percent of the time when searching for low crime areas and 74 percent of the time when searching for high crime areas.

Conclusion

We feel confident that our model will help predict which areas of Chicago are prone to high crime rates or low crime rates. The accuracy is not quite where we want it to be, but the recall gives us reason to trust the model to predict crime. The information provided by the model is not only useful to the Chicago's Law Enforcement as a monitoring tool but also we believe that our

model can be used by real estate investors to know which areas to avoid and which areas to pursue when investing in properties based on crime rates. As everyone might guess, properties in neighborhoods with higher crime rates tend to not appreciate according to the market, it usually have a lower rental price, and have a higher tenant turnover, leading to higher maintenance costs than properties in relatively safer areas and most of the times, a bad investment. In conclusion, using our recall model would allow us to assist real estate investors on pointing out areas with a lower crime.