Predicting Arrests in the City of Chicago

By Justin Zhao, Natalie Tarn, Tyee Pomerantz, & Aaron Chai

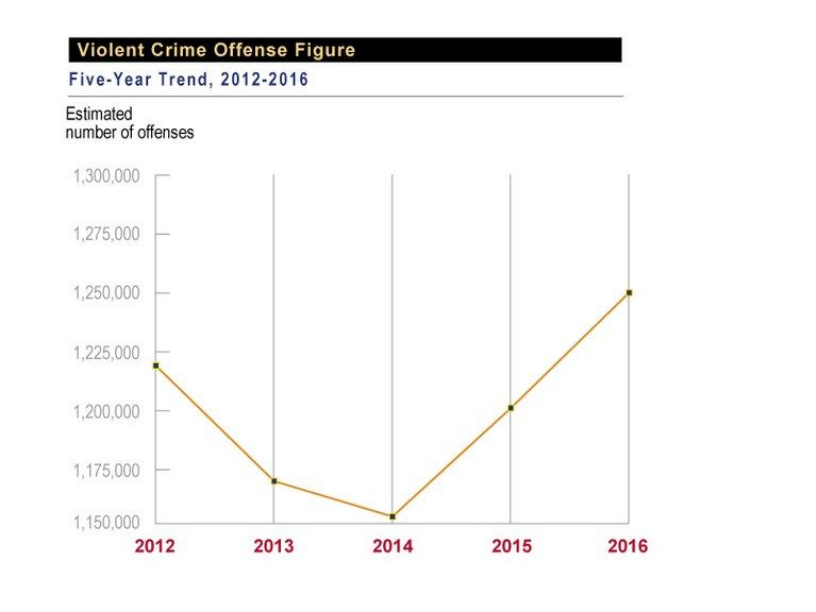
Topic Overview/Introduction

1,248,185

Number of violent crimes in the United States in 2016, roughly a 4.1% increase from 2015.

Introduction

- Since 1990, crime rates in the 30 biggest cities in America has dropped by 64%.
- However, from 2014 to 2016, crime rates across America rose significantly, with the biggest increase in violent crime (10.6% rise from 2014 to 2015).
- This issue has recently come to light in the political sphere, with many politicians addressing it, most notably included in Donald Trump's "Make America Great Again" campaign.



Source: ucr.fbi.gov

How can data science help?

- ➔ It is extremely important that our policymakers and law enforcement officials do their best to keep the American public safe.
- ➔ Of equal importance to reducing crime is making arrests when crime does occur.
- ➔ By looking at data on crimes and arrests, we can identify the factors that are the biggest indicators to predicting whether an arrest is made.
- ➔ We can then make the criminal justice system more efficient/effective in arresting and prosecuting criminals or reducing crime rates.

The Data

2

Our dataset - Why Chicago?

- The dataset we will be looking at is the "Crimes - 2018" dataset from the Chicago Data Portal.
- The data itself is a csv file with 229,359 observations across 22 variables.
- Each observation in the dataset is a separate reported crime that was committed in the city of Chicago in 2018 (up to mid-November), while the variables contain information about the reported crime, such as time, location, type of crime, etc.
- As one of the largest cities in the US, Chicago also has one of the highest rates of crime in the US, meaning that it will be a relevant microcosm for us to conduct our study.

CHICAGO DATA PORTAL

ID	Case Number	Date	Block	IUCR	Primary Type	Description
11525808	JB540604	12/04/2018 11:45:00 ...	031XX N KOLMAR AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
11525830	JB540603	12/04/2018 11:45:00 ...	020XX W NORWOOD ST	0820	THEFT	$500 AND UNDER
11525859	JB540599	12/04/2018 11:43:00 ...	067XX S LANGLEY AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
11525806	JB540595	12/04/2018 11:40:00 ...	011XX N NOBLE ST	0910	MOTOR VEHICLE THEFT	AUTOMOBILE
11525798	JB540583	12/04/2018 11:38:00 ...	018XX S LEAVITT ST	051A	ASSAULT	AGGRAVATED: HANDGUN

Data Organization

Our research question we are interested in answering with our data is:
What factors can be used to predict arrests in the city of Chicago?

A couple touch-ups...

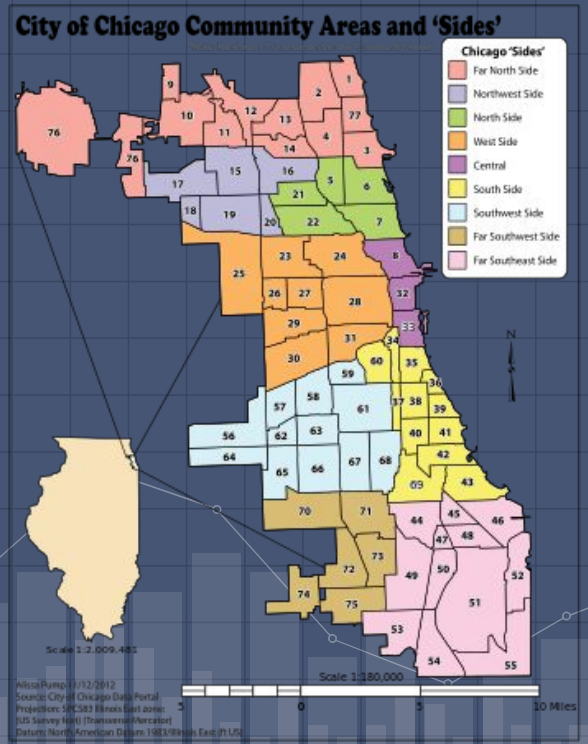
- We will only focus on the variables Date, Primary.Type, Location.Description, Domestic, and Community.Area for our explanatory variables.
- First, let's take a look at the Date variable. Currently, it is too complex and hard to grab information from:

```
##  chr [1:229359] "11/10/2018 19:11" "11/10/2018 1:39" "11/3/2018 0:00" ...
```

- As a result, we decided to create two new variables, month, which indicates the month in which the crime was committed, and time, which indicates the time of day (in military time) the crime was committed.
- However, our month and time variables still introduced too many levels for efficient logistic regression. We decided to create two new variables to break down the variables further:
  - season: indicates which season the crime was committed
  - daytime: a logical variable which indicates whether the crime was committed during daylight hours (6AM - 6PM).

A couple more...

- Next, we had to break down the Community Area variable, as it had 77 levels.
- Similarly, we created a new side variable which classifies all 77 community areas as one "neighborhood" or "side" of Chicago in accordance to the map to the right:
- Thus, we were able to go from 77 levels in Community Area to 9 levels in side.



Some final details...

- Finally, we can look at our Primary Type and Location Description variable.
- Since both of these variables go very into detail and have many levels, we decided to only focus on the following levels for each:
  - Primary Type: Assault, Battery, Criminal Sexual Assault, Homicide, Robbery, Theft, and Other Offense (violent or frequent crimes).
  - Location Description: Street, Sidewalk, Apartment, Residence, Other (top 5 most common crime locations).
- We then filtered for only these levels and put the results in a new dataset to have a fresh set of relevant data to work with.

```
## Observations: 88,051
## Variables: 10
## $ `Case Number`          <chr> "JB511275", "JB508160", "JB518326", "JB...
## $ `Primary Type`         <chr> "HOMICIDE", "THEFT", "OTHER OFFENSE", "...
## $ `Location Description` <chr> "APARTMENT", "STREET", "STREET", "RESID...
## $ Arrest                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
## $ Domestic               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE...
## $ month                  <chr> "November", "November", "November", "No...
## $ time                   <chr> "19:11", "10:00", "9:00", "9:00", "23:0...
## $ season                 <chr> "Fall", "Fall", "Fall", "Fall", "Fall",...
## $ daytime                <lgl> FALSE, TRUE, TRUE, TRUE, FALSE, FALSE, ...
## $ side                   <chr> "Southwest", "Far Southwest", "South", ...
```

Data Analysis

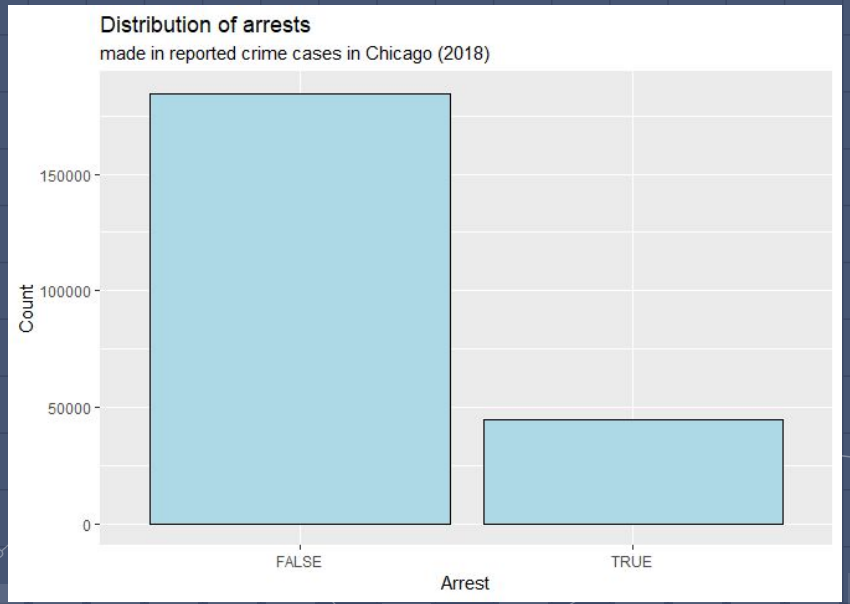
Logistic Regression w/ AIC Model Selection

4

A look at the variables

- For our logistic regression, our response variable is the binary Arrest variable, coded as True/False, and we are interested in cases where an arrest has been made (Arrest = True).
- Let's take a look at the initial distribution of this variable:
- Our explanatory variables will be Primary Type, Location Description, season, daytime, side, and Domestic.

```
## # A tibble: 2 x 2
##   Arrest  prop
##   <lgl>  <dbl>
## 1 FALSE   80.5
## 2 TRUE    19.5
```



Logistic Regression

- We started by creating the full logistic regression model with all the variables.
  - The summary output is as follows:

```
## Call:
## glm(formula = Arrest ~ season + daytime + Domestic + `Primary Type` +
##     `Location Description` + side, family = "binomial", data = crimesdata)
##
```

```
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -1.71830    0.06355 -27.038  < 2e-16 ***
## seasonSpring                              0.07953    0.02990   2.660 0.007821 **
## seasonSummer                              0.05760    0.02922   1.972 0.048658 *
## seasonWinter                              0.22444    0.03403   6.596 4.24e-11 ***
## daytimeTRUE                              -0.16991    0.02129  -7.983 1.43e-15 ***
## DomesticTRUE                              0.04086    0.02477   1.649 0.099079 .
## `Primary Type`BATTERY                     0.34531    0.03107  11.114  < 2e-16 ***
## `Primary Type`CRIM SEXUAL ASSAULT        -0.90751    0.14895  -6.093 1.11e-09 ***
## `Primary Type`HOMICIDE                   -0.11449    0.15519  -0.738 0.460675
## `Primary Type`OTHER OFFENSE               0.41840    0.03618  11.565  < 2e-16 ***
## `Primary Type`ROBBERY                    -1.23121    0.06508 -18.919  < 2e-16 ***
## `Primary Type`THEFT                      -1.85263    0.04808 -38.533  < 2e-16 ***
## `Location Description`OTHER              -0.37754    0.05870  -6.432 1.26e-10 ***
## `Location Description`RESIDENCE          -0.38983    0.03085 -12.636  < 2e-16 ***
## `Location Description`SIDEWALK            0.25440    0.03535   7.197 6.13e-13 ***
## `Location Description`STREET              0.28366    0.03104   9.139  < 2e-16 ***
## sideFar North                            -0.11313    0.06062  -1.866 0.062001 .
## sideFar Southeast                         0.24947    0.05612   4.445 8.79e-06 ***
## sideFar Southwest                         0.02769    0.06198   0.447 0.655074
## sideNorth                                -0.18419    0.06767  -2.722 0.006492 **
## sideNorthwest                            -0.20103    0.06707  -2.997 0.002725 **
## sideSouth                                -0.06635    0.05462  -1.215 0.224476
## sideSouthwest                             0.02195    0.05446   0.403 0.686889
## sideWest                                 -0.19156    0.05213  -3.675 0.000238 ***
## ---
```

Logistic Regression

- From our full model, we can see that there are many coefficients in our original logistic regression. We can perform backward model selection on this model to try and choose a better model with a lower AIC.
- Below are the results of our backward model selection process:

```
Start:  AIC=60330.77
Arrest ~ season + daytime + Domestic + `Primary Type` + `Location Description` +
    side

                        Df Deviance   AIC
<none>                       60283 60331
- Domestic               1    60285 60331
- season                 3    60328 60370
- daytime                1    60346 60392
- side                   8    60460 60492
- `Location Description` 4    60882 60922
- `Primary Type`         6    64769 64805
```

Logistic Regression

- From the results of our backwards AIC model selection, our original full model was still the best logistic regression model of our results.
- This can be explained by the fact that most of our variables had multiple levels, and removing that variable from the full model would also lose a lot of specificity and diversity in the data, thereby increasing the AIC.
- Knowing this, we looked at the coefficients of our full logistic regression model to get a better idea of which factors influence our response variables the most.

Logistic Regression

- ➔ However, the current coefficients of our full logistic regression model are the increase in log odds for an arrest being made per change in explanatory variable, which is quite hard to conceptualize.
- ➔ Instead, we can exponentiate the coefficients to make them easier to understand, using the exp() function in R.

```
##                   (Intercept)                   seasonSpring
##                     0.1793707                     1.0827820
##                   seasonSummer                   seasonWinter
##                     1.0592940                     1.2516249
##                   daytimeTRUE                   DomesticTRUE
##                     0.8437371                     1.0417042
##          `Primary Type`BATTERY `Primary Type`CRIM SEXUAL ASSAULT
##                     1.4124294                     0.4035286
##        `Primary Type`HOMICIDE      `Primary Type`OTHER OFFENSE
##                     0.8918221                     1.5195316
##         `Primary Type`ROBBERY             `Primary Type`THEFT
##                     0.2919377                     0.1568244
##     `Location Description`OTHER  `Location Description`RESIDENCE
##                     0.6855464                     0.6771721
##  `Location Description`SIDEWALK    `Location Description`STREET
##                     1.2896923                     1.3279855
##                 sideFar North                 sideFar Southeast
##                     0.8930343                     1.2833433
##               sideFar Southwest                     sideNorth
##                     1.0280759                     0.8317736
##                 sideNorthwest                     sideSouth
##                     0.8178855                     0.9358023
##                 sideSouthwest                      sideWest
##                     1.0221950                     0.8256727
```

A closer examination

- To begin, we looked at the season variable. Of the levels in the summary output, only the coefficient for winter was statistically significant, with the odds of an arrest being made increasing by a factor of 1.25 if a crime was committed in the winter vs if a crime was committed in the fall.
- If we recall from earlier, January and February are the months to which we assigned the season winter.
  - We looked at the relative arrest rates in these two months compared to the yearly average:

```
## # A tibble: 2 x 4
## # Groups:   month [2]
##   month    Arrest     n prop
##   <chr>    <lgl>  <int> <chr>
## 1 February TRUE    3837 22.3%
## 2 January  TRUE    4148 20.5%

## # A tibble: 1 x 2
##   Arrest  prop
##   <lgl>  <dbl>
## 1 TRUE    19.5
```

A closer examination

- Next, we looked at the daytime variable.
- According to the output, the odds of a criminal being arrested for a crime committed between 6AM and 6PM decrease by a factor of 0.16 when compared to a crime committed between 7PM and 5AM (more likely to be arrested for a crime committed at night).
- This is pretty surprising, given that one would expect there to be less police presence during night hours.

```
## # A tibble: 2 x 4
## # Groups:   daytime [2]
##   daytime Arrest     n prop
##   <lgl>   <lgl>  <int> <chr>
## 1 FALSE   TRUE   19002 20.5%
## 2 TRUE    TRUE   25712 18.8%

## # A tibble: 1 x 2
##   Arrest  prop
##   <lgl>  <dbl>
## 1 TRUE    19.5
```

A closer examination

➔ Since the Domestic variable yielded no statistically significant coefficients, below are the arrest proportions for the Primary Type and Location Description Variables.

```
## # A tibble: 7 x 4
## # Groups:   Primary Type [7]
##   `Primary Type`      Arrest     n  prop
##   <chr>               <lgl>  <int> <dbl>
## 1 BATTERY             TRUE    8659 19.9
## 2 OTHER OFFENSE       TRUE    2919 19.8
## 3 HOMICIDE            TRUE      93 18.6
## 4 ASSAULT             TRUE    2861 16.1
## 5 THEFT               TRUE    5330  9.57
## 6 ROBBERY             TRUE     621  7.44
## 7 CRIM SEXUAL ASSAULT TRUE      71  5.21
```

```
## # A tibble: 5 x 4
## # Groups:   Location Description [5]
##   `Location Description` Arrest     n  prop
##   <chr>                  <lgl>  <int> <dbl>
## 1 SIDEWALK               TRUE    6957 37.3
## 2 STREET                 TRUE   10675 20.9
## 3 APARTMENT              TRUE    4139 14.2
## 4 RESIDENCE              TRUE    4143 10.9
## 5 OTHER                  TRUE     847  9.23
```

A closer examination

- Finally, the distribution for the statistically significant coefficients of the side variable are as follows:
- As one can see from this distribution, the highest proportion of arrests occurred for crimes committed in the West and Far Southeast sides of Chicago.
- A significantly low proportion number of arrests were made for crimes committed in the North, Far North, and Northwest sides of Chicago.

```
## # A tibble: 9 x 4
## # Groups:   side [9]
##   side           Arrest     n prop
##   <chr>          <lgl>  <int> <chr>
## 1 West           TRUE   14763 25.8%
## 2 Far Southeast  TRUE    4691 22.5%
## 3 Southwest      TRUE    6195 20.4%
## 4 Far Southwest  TRUE    2742 19.2%
## 5 South          TRUE    5427 18.1%
## 6 Central        TRUE    3308 14.9%
## 7 Far North      TRUE    2829 13.9%
## 8 Northwest      TRUE    1578 13.3%
## 9 North          TRUE    1860 11.2%
```

Conclusion

5

Results

- Overall, from observing all the statistically significant coefficients from our selected logistic regression model and comparing the relative arrest proportions for each, we can conclude that the following factors are most closely associated with an arrest:
  - If the crime occurred within the first few months of the year
  - If the crime was committed during daylight hours (6AM - 6PM)
  - If the crime was a battery or other minor offense
  - If the crime was committed outdoors (street or sidewalk)
  - If the crime was committed in the West or Far Southeast side of Chicago

Applications

- With insights gained from our data analysis, we can suggest the following to city officials and law enforcement to increase the proportion of arrests made in Chicago overall.
  - Work to increase the number of arrests made later on in the year, especially during summer (a possible mid-year lull)
  - Increase police activity during daylight hours
  - Reform the way arrests for violent or serious crimes are made, reported, and processed
  - Increase efficiency in making arrests dealing with crimes committed in private residences
  - Increase police activity and presence, as well as the number of arrests made in the northern region of Chicago

Further Readings

Below is a link to an article by Business Insider about how the LAPD is also utilizing data science and data analytics to increase efficiency in dealing with crime:

https://www.businessinsider.com/the-lapd-is-predicting-where-crime-will-occur-based-on-computer-analysis-2014-6

References

https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-2014-crime-statistics

https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/violent-crime

https://data.cityofchicago.org/Public-Safety/Crimes-2018/3i3m-jwuy

https://stats.idre.ucla.edu/r/dae/logit-regression/

THANKS!

Any questions?