Helping Chicago Communities Identify Subjects Who Are Likely To Be Involved in Shootings

Group 6:

Josiah Argo, Brendan Sigale, Dylan Leigh, Connor McDevitt, Keenan Daly

Table of Contents

Overview

Shooting rates in Chicago are among the highest nationwide. We will be using the Chicago Police Department's Strategic Subject List to develop a model which accurately predicts the score for a person based on various factors in order to extrapolate the data to new people, as the data only contains listings for people from 2012-2016. We will also generate clusters to estimate the likelihood of a person being involved in a shooting if the user has a lack of detailed information. This data consists of almost 400,000 anonymized arrest records from the City of Chicago and a score for each person as to whether they are likely to be involved in a shooting, either as a victim or a perpetrator. The SSL score found in the data represents the person's propensity to be involved in a shooting on a scale of 0 (extremely low risk) to 500 (extremely high risk). The original model only used eight features, but we plan to use more, such as their neighborhood or TRAP (Targeted Repeat Offenders Program) status. This data can be used by community activists to identify people in their reach who they can specifically target for their programs.

This report will detail our exploratory analysis of the data, followed by which models we chose to test, what parameters we used in those models, and which model we chose to move forward with. Then we will evaluate our model, including its weaknesses and possible improvements. We will then describe our clusters, which can be used to generalize people based on factors which can be learned easily. Finally, we will go over what insights can be gained from our model and how it can be used in the future. "Inner-city gun violence mostly occurs within small networks of people who know each other, as allies or enemies, and are at high risk for becoming violent. Those groups typically account for 0.5 percent or less of a city's population" (Bower).

Background Information

Chicago has always been known for its gun violence, but definitely has of late. "The Chicago Police Department has recovered 7,000 guns per year that had been illegally owned or associated with a crime between 2013 and 2016" (Kight and Sykes). BUILD is one of many community groups trying to stop the violence and we believe our model can help them. Which it deeply needs right now because in 2017 only 17% of murders in Chicago were solved (Kight and Sykes). A 27-year-old man was killed at a house party in South Side Chicago because of a "war of words over social media" (Bower). Our model would be a stepping stone in the right direction for the Chicago Police Department and community organizations because they will be able to identify who is at risk to be involved with gun violence. This June in Chicago fifty-six people were shot in one weekend alone and more than 100 people got shot two years ago on the Fourth of July weekend (Pascus). This is just a couple of examples of gun violence in Chicago that we are trying to help prevent or help solve crimes.

Exploratory Analysis

Predictors

Age group was the first attribute we analyzed to determine its influence on the score. The histograms of the factor show extremely little overlap in the distributions of each level. From this visual, we can

conclude that age will be a great predictor of score. Furthermore, there is a perfect inverse relationship between age and score. Ages under 20 consistently yield scores closer 500 and as the age group increases, the scores proportionally decrease. This is evidence that our predictive model will be heavily influenced by the age of an offender.

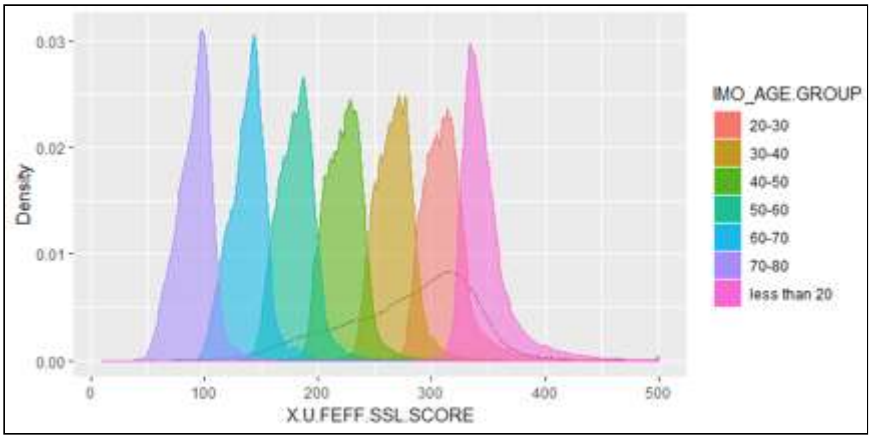

Figure 1 Histograms of Age Group

Next, we focused on the previous weapon charge attribute. This feature holds a boolean value of "Yes", the offender has unlawfully been charged of possessing a weapon, or "No", they have not been charged. One would logically assume that if a past-convict has been charged with such a crime, that they would be more likely to commit a violent crime in the future. The cumulative curve for this feature certainly backs this assertion, as it reveals the "Y" group does generally have a higher score than the "N" group. In conclusion, previous weapon charge can be labeled as a good predictor of the score.

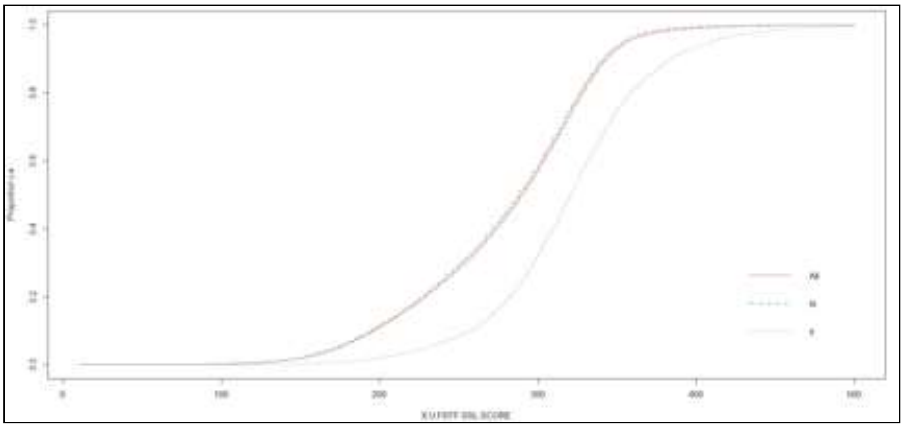

Fig. 2 Cumulative Curve of Age

Racial Biases

In our exploratory analysis, we observed a bias large enough in scale that it warrants acknowledgement. In the City of Chicago, Whites and African Americans use drugs at similar rates (NAACP). While the demographic breakdown of the city is 49.14% white and 30.51% African American (World Population

Review), there are significantly more African Americans in our dataset who have been convicted of a drug offense. Furthermore, the proportion of African Americans charged with a drug offense is approximately five times higher than any other race group.

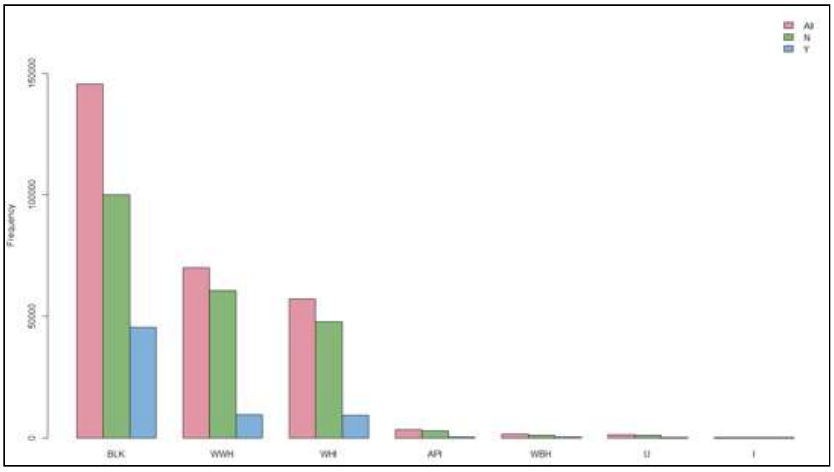

Fig. 3 Bar Plots of Race Group

An analysis of the drug offense histogram reveals that high scores can be influenced by a prior drug charge. In conclusion, the bias of African Americans disproportionately being persecuted for drug offenses may affect the accuracy of our predictive model because high scores will not be assigned from a objectively-convicted collection of observations.

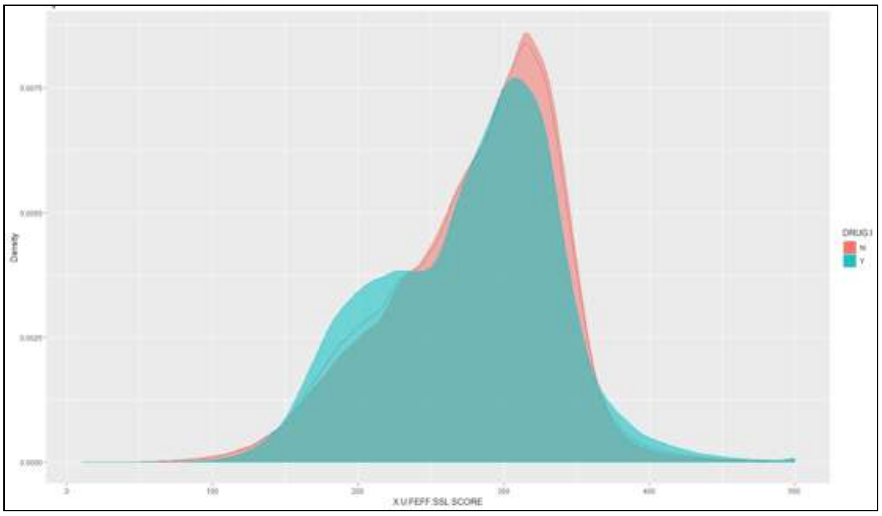

Fig. 4 Histograms of Prior Drug Offense

Data Cleaning

The data cleaning process took place in both Excel and Rattle.The original dataset had over 40 features, of which many were deleted because they would not directly pertain to our inquiries.

Once the dataset was imported into the rattle application with a 70/15/15 partition, it held 279,078 rows and 15 features. Eight of the features were categorical and 7 were numeric. with our target variable, SSL score, being numeric. The numeric variables of domestic arrest count, weapon arrest count, and narcotics arrest count all contained null values for instances where the subject had 0 (e.g. no narcotics arrests). Therefore, null values were imputed as a 0 so that these features would be accurately represented. After imputation, the dataset contained 0 missing values so there was no need to delete observations with missing data.

Lastly, there was additional data transformation that took place before the ANN model. All categorical variables were transformed into indicator variables to ensure that the dataset would be compatible with a neural network model. In addition, numeric variables were re-scaled from 0-1 so that high-values would not compromise the integrity of the model.

Data Science

Modeling

We used Rattle to determine the best model to predict a person's score based on several input variables. We tested all four regression models: Decision Tree, Random Forest, Neural Networks, and Linear Regressions with various tuning parameters to identify the best combination of regression type and parameters.

The results (R^2 and MAE) with their tested parameters were as follows:

Decision Tree

	Complexity	MAE	R2
Model 1	**0.0025**	**11.4798**	**0.9097**
Model 2	0.0075	13.932	0.8812
Model 3	0.01	13.932	0.8812
Model 4	0.015	14.7498	0.8683
Model 5	0.02	14.7498	0.8683

Linear Regression

	MAE	R2
Model 1	**10.7862**	**0.9207**

Random Forest - Impute = ON

	Trees	# Variables	R2	MAE
Model 1	100	2	0.9145	**12.7146**
Model 2	50	2	**0.9147**	12.8245
Model 3	47	2	0.9139	12.9127
Model 4	45	2	0.9137	12.9848
Model 5	35	2	0.9141	13.1000

ANN

	Nodes	MAE	R2
Model 1	5	10.6424	0.9214
Model 2	3	10.6812	0.921
Model 3	7	14.1757	0.8895
Model 4	10	11.0895	0.9162
Model 5	**20**	**10.5925**	**0.9226**

Fig. 5 Optimal Tuning Parameters

The Decision Tree's ideal complexity was **0.0025**. Random Forest's ideal amount of Trees was **100**. ANN's ideal amount of nodes was **20**. Linear Regression didn't have different parameters to test.

Evaluation

After selecting the best parameters we evaluated each model with different seeds:

Decision Tree		
Seed	MAE	R2
42	11.4798	0.9097
342190	11.3768	0.9111
40599	11.4110	0.9105
AVG	**11.4225**	**0.9104**

Random Forest		
Seed	MAE	R2
42	12.7146	0.9145
342190	12.9583	0.9132
40599	13.0118	0.9128
AVG	**12.8949**	**0.9135**

Linear Regression		
Seed	MAE	R2
42	10.7862	0.9207
342190	10.808	0.9211
40599	10.8371	0.9213
AVG	**10.8104**	**0.9210**

BEST MODEL

ANN		
Seed	MAE	R2
42	10.5925	0.9226
342190	10.5385	0.9238
40599	10.6179	0.9237
AVG	**10.5830**	**0.9234**

Fig. 6 Model Testing

ANN was selected as the best model as it had the lowest MAE **- 10.583** and highest R2 at **0.9234**. An R^2 of 0.9234 and a MAE of 10.583 is extremely accurate for this data, as the score can be anywhere within 0-500. A mean average error of 10 translates to a 2% difference in score, which is not indicative of a huge shift in a subject's likelihood to be involved in a shooting. The difference between a subject with a score of 380 and a subject with a score of 370 is essentially negligible. However, we are not able to determine if this is the best model achievable. Due to technical limitations, we were not able to test the model parameters to the full extent we would have liked. For example, we were limited to only two variables and a small amount of trees for the random forest testing, as tests with three variables spent over an hour processing with no progress, even with only 10 trees. That being said, achieving a more accurate fits runs the risk of overfitting, which would be even worse than having inaccurate data. For these reasons, we were satisfied with the results of our regression models.

Clustering

We sought to generate clusters for the various subjects to learn if there were any generalizable results. This would allow community organizations to estimate the likelihood of a member being involved in a shooting based on facts they might know about the person, such as their arrest history, gender, race, and involvement with weapons and drugs.

Using k-means clustering, we determined that three clusters was the optimal amount of clusters for this data. This was done by using the cleaned data in rattle and using the iterate function to generate an elbow plot.

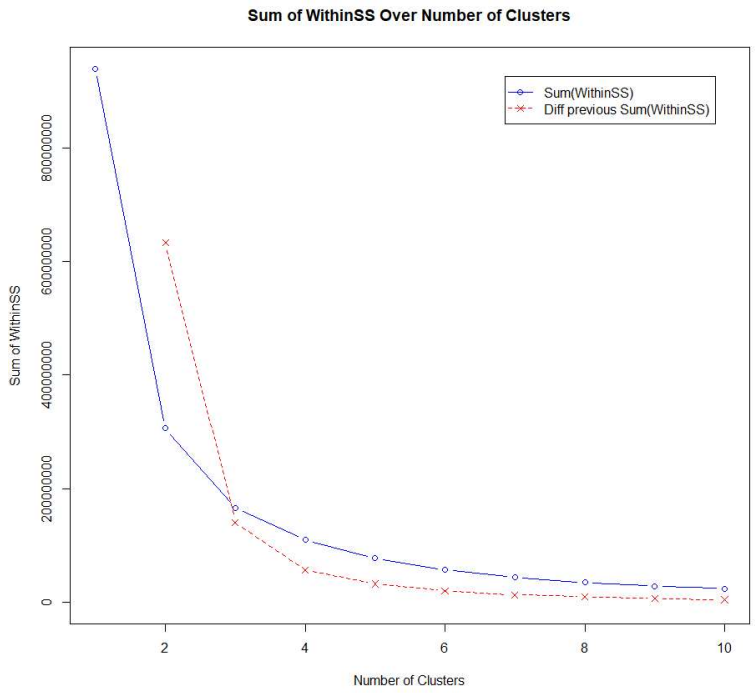

Fig. 7 Elbow Curve

After generating three clusters, we proceeded to analyze them based on their feature means.

Cluster	**1**	**2**	**3**
X.U.FEFF.SSL.SCORE	0.489273	0.647191	0.503367
TIN_SEX.CODE.CD_F	0.232589	0.306950	0.170677
TIN_SEX.CODE.CD_M	0.767213	0.692922	0.829202
TIN_SEX.CODE.CD_X	0.000198	0.000129	0.000122
TIN_RACE.CODE.CD_API	0.013586	0.011883	0.009362
TIN_RACE.CODE.CD_BLK	0.456290	0.510339	0.611983
TIN_RACE.CODE.CD_I	0.000982	0.000594	0.000474
TIN_RACE.CODE.CD_U	0.005201	0.005165	0.003477
TIN_RACE.CODE.CD_WBH	0.004940	0.005709	0.004353
TIN_RACE.CODE.CD_WHI	0.241685	0.190122	0.181255
TIN_RACE.CODE.CD_WWH	0.277317	0.276189	0.189097
TIN_WEAPON.I_N	0.964431	0.955863	0.932765
TIN_WEAPON.I_Y	0.035569	0.044137	0.067235
TIN_DRUG.I_Y	0.130664	0.003067	0.636263
TIN_IMO_AGE.GROUP_20.30	0.157190	0.590520	0.290982
TIN_IMO_AGE.CURR_20.30	0.000000	0.763179	0.275858
TIN_IMO_AGE.GROUP_30.40	0.624320	0.000000	0.028000
TIN_IMO_AGE.CURR_30.40	0.681057	0.000000	0.052365
TIN_IMO_AGE.GROUP_40.50	0.000000	0.000010	0.498134

TIN_IMO_AGE.CURR_40.50	0.100494	0.000020	0.421586
TIN_IMO_AGE.GROUP_50.60	0.169136	0.000000	0.093995
TIN_IMO_AGE.CURR_50.60	0.143071	0.000000	0.186580
TIN_IMO_AGE.GROUP_60.70	0.043402	0.000000	0.015757
TIN_IMO_AGE.CURR_60.70	0.065061	0.000000	0.026760
TIN_IMO_AGE.GROUP_70.80	0.005911	0.000623	0.000644
TIN_IMO_AGE.CURR_70.80	0.010318	0.000010	0.001654
TIN_IMO_AGE.GROUP_less.than.20	0.000042	0.408847	0.072487
TIN_IMO_AGE.CURR_less.than.20	0.000000	0.236792	0.035198
TIN_ICN_PAROLEE.I_Y	0.027580	0.012486	0.064657
TIN_ICN_PAROLEE.I_N	0.972420	0.987514	0.935343
TIN_TFC_ICN_TRAP.STATUS_.0.0.	0.995583	0.994558	0.985264
TIN_TFC_ICN_TRAP.STATUS_.0.3.	0.003540	0.004571	0.011660
TIN_TFC_ICN_TRAP.STATUS_.3.4.	0.000877	0.000871	0.003076
R01_YEARS_SINCE_LAST_CONTACT	0.196619	0.169575	0.172311
R01_ICN_WEAPONS.ARR.CNT	0.006833	0.008944	0.013802
R01_ICN_NARCOTICS.ARR.CNT	0.011985	0.000187	0.063653
R01_ICN_DOMESTIC.ARR.CNT	0.012463	0.007255	0.012284

Fig. 8 Feature Means of Clusters

These feature means allow us to separate the clusters into three distinctly different groups of people, based on gender, race, age, and narcotics and weapons history.

- **Cluster 1 – Some Drug Involvement**
  - Most varied race, primarily male, least victimized by weapons, some drug use, 30-40 years old, some arrests
- **Cluster 2 – Some Weapons Involvement**
  - Less varied race, more females, some victims of and involvement with weapons, few drugs, under 30, large population, very few arrests
- **Cluster 3 – High Weapons and Drug Involvement**
  - Notably more black people, overwhelmingly male, not victims of weapons but arrested for weapons, mostly in 40-60 years old range, proportionately higher on TRAP, significantly more narcotics arrests.

These clusters will make it easy for community organizations to determine how likely any new member is to be involved in a shooting. An interesting factor is that the mean of the SSL score does not vary significantly between the three clusters. This indicates the model predicts subjects on the list are somewhat equally likely to be involved in a shooting based simply on the fact that they have been arrested before. A list that included all residents of Chicago, regardless of whether they have been arrested before, would naturally create new clusters which would boost the means in the score category of the clusters above. This can be seen in microcosms within these clusters, such as the TRAP score. The Targeted Repeat Offenders (TRAP) List has only 3118 members, representing under 1% of subjects on the list. Because the TRAP status was processed as an indicator variable, we can easily see that subjects in Cluster

3 have more people on the TRAP list. Due to the slight difference in the cluster mean. Were there are new cluster of people not on the list with extremely low scores, the score means of the three clusters above would increase significantly.

Business Insights

This model tells us what factors often lead an individual to be involved in a shooting. Community Organizations in Chicago and specialized crime prevention units in the Chicago Police Department can use these insights to help predict who will be involved in shootings. They could even extend that beyond predicting shootings into starting to predict perpetrators and victims of other similar violent crimes such as assault or battery, provided they have more data to add to the model.

Limitations and Improvements

This project has implications for communities across the board. However, we were limited by some factors, such as processing power and the quality of the data. We would have loved to break this data down by location. The police data had the district of their last arrest, but this data was difficult to work with because much of the data was missing. In addition, this data is currently limited to people who already have an arrest history. While this works if the subject has an arrest history, the model and clusters are not generalizable to those without an arrest history. In addition, we would have liked to test additional model parameters. While this would have run the risk of overfitting our model, we can not know without seeing the data. Some of our models described above already took too long to run, such as any random forest with three variables. With more processing power, we could run more accurate models and assess them for overall fit and the risk of them overfitting our data.

Conclusion

After running our analysis, we have determined two different methods of estimating someone's likelihood of being involved in a shooting, which can be used by Community Organizations in Chicago and specialized crime prevention units in the Chicago Police Department. We are confident that a neural network with 20 nodes is the best method we were capable of testing to predict a community member's score given a large amount of information on that person, such as their age, race, gender, and prior involvement with illegal activity. This allows them to model their current community members to obtain accurate scores for these people, allowing them to more selectively target their community outreach programs. The second method we identified was to develop clusters based on their demographics and criminal history, allowing community organizations to make more general inferences about their community members. However, there are limitations to this method, as we are only able to cluster people who have an arrest history to begin with. While this helps cluster people who have an arrest history, it is difficult to generalize a determination of who has been arrested before, making this method difficult to implement on community members whom the organization has had little to no contact with. These tools can be used across Chicago to make their efforts to better their community easier, reducing shootings in Chicago and bringing communities out of violence across the board.

References

Bower, Bruce. “Can Neighborhood Outreach Reduce Inner-City Gun Violence in the U.S.?” *Science News*, 5 Nov. 2019, www.sciencenews.org/article/neighborhood-outreach-can-reduce-inner-city-gun-violence.

Chicago Tribune. “Tracking Chicago Homicide Victims.” *Chicagotribune.com*, Chicago Tribune, 3 Dec. 2019, www.chicagotribune.com/news/breaking/ct-chicago-homicides-data-tracker-htmlstory.html.

Chicago Tribune. “Tracking Chicago Shooting Victims.” *Chicagotribune.com*, Chicago Tribune, 3 Dec. 2019, www.chicagotribune.com/data/ct-shooting-victims-map-charts-htmlstory.html.

“Chicago, Illinois Population 2019.” *Chicago, Illinois Population 2019 (Demographics, Maps, Graphs)*, worldpopulationreview.com/us-cities/chicago-population/.

Chicago, City of. “Strategic Subject List: City of Chicago: Data Portal.” *Chicago Data Portal*, 1 May 2017, data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np.

“Criminal Justice Fact Sheet.” *NAACP*, www.naacp.org/criminal-justice-fact-sheet/.

Kight, Stef W., and Michael Sykes. “The Deadliest City: Behind Chicago's Segregated Shooting Sprees.” *Axios*, www.axios.com/chicago-gun-violence-murder-rate-statistics-4addeeec-d8d8-4ce7-a26b-81d428c14836.html.

Pascus, Brian. “56 people were shot in Chicago over the weekend, including four fatally.” CBSNEWS, www.cbsnews.com/news/chicago-gun-violence-56-shot-this-weekend-four-killed/