Introduction and Agenda

Introduction - Slide 1
Hi, my name is Katie Mason and I will be presenting a predictive analysis of the Chicago crime dataset.

Agenda - Slide 2
The agenda for today's presentation will start with some background on Chicago crime. We'll then take a look at the objectives and success criteria for the predictive analysis. Next we'll take a look at the dataset and some pre-processing steps taken prior to modeling, followed by modeling and evaluation, and then, lastly, deployment.

Business Understanding - Slides 3-4

Background - Slide 3

Last year, Chicago experienced the highest levels of violent crime in almost two decades.(1,2) Homicides increased by over 50% from 2015 to 2016 and the number of victims from gun shootings increased by 46%. Some factors that may have contributed to this drastic increase are distrust of police force, low police morale, lower arrest rates, gang violence, and an influx of illegal firearms. While both violent and non-violent crimes have serious legal consequences, violence is especially pernicious to the health and well-being of communities. Although the definition of what constitutes a violent crime may vary, it is generally accepted that homicide, rape and sexual assault, robbery, aggravated assault and aggravated battery classify as violent crimes. To address the increase in violence this past year, the police department is looking to add predictive analytics to its arsenal for fighting crime.

Objectives - Slide 4

The objectives of this research are to better anticipate violent crime by examining its relationship with factors such as arrest rates and location using predictive analytics. Using the Chicago crime dataset, I hope to answer the following questions:

1. When does violent crime occur?

2. Where is violent crime occurring?
3. What are profiles of violent crime?

To answer these questions, we will use the following methods for analysis:

- Use classification to determine which places experience higher violent crime rates
- Use classification to determine which days and times have higher violent crime rates
- Use clustering to generate a model that produces profiles for violent crimes

Success Criteria - Slide 4

The desired outcome from this research analysis is to identify factors that predict violent crime effectively and use that information to prevent it from happening. To evaluate these factors, the analysis will be used to

- Predict the days and times that experience the highest levels of violent crime
- Identify the places with highest violent crime rates
- Determine common profiles of violent crimes

Data Understanding - Slide 5

The Chicago crime dataset is comprised of public crime data occurring between January 19, 2015 to February 6, 2015. Promising variables in this dataset are Primary Type, IUCR codes, Police District, DateTime, Domestic Violence, and Arrest Rate. The Illinois Uniform Crime Reporting (IUCR) codes will be used to create the flag target variable, which will classify crimes as violent or non-violent. Primary Type, renamed Crime Type, is the type of crime and will be useful for identifying crimes by their IUCR codes.

Of the potential location variables, District was selected because it aligns with the organization of the police department and the results can be easily applied to departmental solutions. The DateTime variable provides the temporal dimension to the crimes which will directly contribute to predicting when a crime will occur. Both domestic violence, a separate category from violent crime, and arrest rates may add a useful dimension to analysis.

Due to the fact that the dataset only captures a few weeks worth of crime data, there is not enough information to yield accurate predictions. Although the scope of the analysis is limited, we still may uncover some useful information.

Data Preparation - Slides 6-8

Variable Pre-Processing - Slide 6

The initial raw Chicago crime dataset came with variables that as-is would not be useful for analysis. The data audit revealed that the dataset contained missing data, unformatted data, and fields that need some type of transformation.

- DateTime: The DateTime field provides two variables that can be separated out into Date and Time. A Weekday variable indicating the weekday the crime occurred was derived using date. A Time Intervals variable was derived into eight 2-hour blocks of time.
- District: The District field was missing 157 values with 98.618% of the field complete. Additionally, the District values imported contained leading zeroes and were formatted to show just the integer. This was helpful when matching District values on an imported map.
- Location Description: Looking at the Location Description field, some values were too granular and were consolidated, such as Airport-related values, CTA-related values, and residence-related values. Additionally, Location Description was missing five values, with 99.956% complete. Blank values in this field were classified as null, since modeler did not detect the missing values.
- IUCR: The Illinois Uniform Crime Reporting codes are used to classify criminal incidents. These codes were used to classify crimes as either violent or non-violent, see Appendix A for more information on violent crime classification. Although domestic violence is considered a violent act, the definition of a violent crime provided by the city of Chicago did not use this classification to identify a crime as violent, and, will not be used as part of the definition. Please see Appendix B for more details.

Distribution of Violent Crime - Slide 7

A brief look at the newly derived variable, violent crime, shows that violent crime tends to occur in the south and west of Chicago. 9.7% of violent crimes occurred in District 11, followed by District 6 with 8.4%. In Districts 16 and 20, only 1.4% and 0.9% of violent crimes occurred.

Feature Selection and Anomaly Detection - Slide 8

To optimize the model, Feature Selection and Anomaly Detection were applied to the dataset. The Feature Selection node revealed that Location Description, District, Arrest, and Time Intervals were all important variables to the model. Both Domestic and Weekday, however, were ranked as unimportant and consequently excluded from the model. The Anomaly Detection node produced four peer groups, with a total of 20 anomalous records, mostly related to the five blank fields in Location Description. Removing the blank fields reduced the number of anomalies by about half. With only 5 records missing data having a large impact on the anomaly presence, they were excluded from the dataset. The null values in District did not have nearly the same effect and will be imputed for modeling.

Modeling & Evaluation - Slide 9-13

Boosting - Slide 9

Two models will be used to evaluate the data for predicting violent crime, a decision tree classification algorithm and clustering. The distribution of the target variable, Violent Crime IUCR, was unbalanced, with only 8% marked as violent crimes. This major imbalance leads to overfitting of the data during training, producing a less reliable model. To address this imbalance prior to modeling, violent crimes marked as ‘true’ were boosted, resulting in an even 50/50 distribution split for true/false.

Classification - Slide 9 - Slide 10

The first classification models were generated using the auto-classifier to determine which decision tree model yielded the best results. The available inputs for the model were Location Description, Arrest, Time Interval, District,

and Anomaly, with Violent Crime IUCR designated as the target variable. An initial run of the auto-classification node revealed the Chi-square Automatic Interaction Detector, classification model, or CHAID model is the best candidate for modeling. The C5.0 model produced a great amount of overfitting, with almost 20% greater accuracy on the training set. Since we're interested in predicting Violent crime accurately, a look at the performance evaluation shows CHAID with 65% true positive outperformed the C&R Tree model with 59% true positive rate.

CHAID Evaluation- Slide 11

Attempts to improve the CHAID model with boosting to enhance accuracy or bagging to enhance stability only increased overfitting rather than model accuracy. Of available inputs, the model ranked Arrest, District, and Location Description as the best predictors, whereas Time Intervals played less of a role. The model when run, still had some overfitting of the data, with 71% accuracy on the training set compared with 65% accuracy on the testing set. The ensemble model, which combines models and uses voting to determine the best model, outperformed the standard reference and naive models.

Looking at the decision tree, violent crime was predicted to occur in Districts 22 and 3 when the location was an apartment, building, private space, or vehicle. Crimes reported at Chicago Housing Authority spaces and public spaces were predicted to experience violent crime in all districts except for 12 and 16. In Districts 17, 18, and 25 the predicted violent crime in the late morning and early evening hours. The Chicago Transit Authority also was predicted to experience violent crime. Health care facilities and parking lots were generally not predicted to experience violent crimes. Residences and stores in Districts 10 and 6 were predicted to experience violent crimes at any time of the day. Districts 19, 2, and 7 were predicted to experience violent crimes in the late night and early morning hours.

Clustering Modeling - Slide 12

The clustering model similarly performed better without domestic or weekday variables due to their low predictor importance. The optimal model generated four clusters from three inputs, Time Intervals, Arrest, and Location Description, with a fair silhouette of 0.4. With fair cluster quality at 0.4 and

strong predictor performance, there is some possibility the clusters variable may improve the model. The clusters can be reclassified as Arrests with 28% of data points, Night Crimes with 40% of data points, Midafternoon Thefts with 13% of data points, and Early Afternoon Apartment Crimes with 18% of data points.

Clustering Evaluation - Slide 13

- Arrests: The first cluster is characterized by all crimes resulting in an arrest. Both Location Description and Time Intervals followed the general distribution of these fields. A noteworthy observation is that 37% of arrests involved Narcotics and only 11% involved theft. Considering that theft is responsible for 20% of all reported crimes in the dataset, measures may need to be taken to address this issue.
- Night Crimes: The second cluster and largest cluster is defined by no arrests, crimes occurring at night between 6 pm and noon primarily at public places and residences. 12% of incidents in this cluster were characterized as violent and contained over 3 times more violent crimes than any other cluster.
- Midafternoon Thefts: The third cluster is the smallest, characterized by crimes with no arrests occurring between 3 and 6 pm. Almost one third of incidents in this category were thefts.
- Early Afternoon Apartment Crimes: The fourth cluster is characterized by no arrests, occurring predominantly at apartments between noon and 3 pm. One feature of this cluster that stands out is that 35% of these incidents involved battery and assault. The high population-density of apartments may be partly responsible for this statistic, though.

The feature selection node ranked the new clusters variable as important reinforcing the possibility that it may improve the model. Upon running the CHAID model including the Clusters variable initially showed an improvement in the accuracy of the model, but further analysis revealed the improvement in the model was limited to the training set and the model performed slightly worse on the testing set in compared with excluding the Clusters variable. Further, the CHAID model overfit the training set by 10%, producing a less favorable result.

Deployment - Slide 14

With this information, some new strategies can be put in place to address issues surrounding violent crimes. Chicago Housing Authority and Chicago Transit Authority spaces may benefit from overall increased police presence. Due to the rise in crime rates in the evening and during the night, shifts during these times should be staffed accordingly.

Apartments in particular appear to experience crimes in the early afternoon with a significant portion involving battery and assault. These spaces may benefit from an increase police presence to deter some of the violent crime.

Theft accounted for 20% of crimes while only 10% of thefts result in arrest. Although theft was prevalent in all districts, 18% of thefts occurred in Districts 1 and 18, so these districts might be targets for reducing theft first. Additionally, 40% of thefts occurred in the late afternoon and early evening, supporting increased staffing of shifts during these times.

Although narcotics have a highly efficient arrest rate, the prevalence rate of 11% itself is high. The city of Chicago may consider investing in more social services to provide support for these areas and the police may want engage in a drug-free campaign with local citizens and schools.

Conclusion - Slide 15

To address high violent crime rates, this analysis explored violent crime in the city of Chicago. Questions asked were about when and where violent crime happened. The analysis revealed that violent crime generally occurred between 6 pm and noon. Unfortunately, only 13% of violent crimes resulted in arrest. Additionally, it was discovered apartments experienced higher violent crime rates in the early afternoon, and thefts generally occur in the late afternoon. Because the variable week day did not effectively contribute to the model, we were not able to predict what days experience more violent crime.

Effectively staffing and deploying the police force in communities based on these recommendations will hopefully lead to higher arrest rates and reduced crimes. Moreover, these measures will allow the department to make cost-

effective decisions for staffing the department by optimizing the shifts during high crime and low crime times. That said, it is important to keep in mind the limitations of this analysis due to the small dataset and few variables available for analysis. Lastly, a byproduct of this analysis revealed a drug problem in the city of Chicago. The police may consider launching a drug-free initiative to raise awareness of this problem and work with communities to find solutions.

Appendix A - Classification of Violent Crimes - Slide 16

Violent crimes were classified using IUCR codes and according to the definition by the City of Chicago, as specified at the Chicago ClearMap Crime Summary website:
http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html#N03

The following codes were classified as Violent Crimes:

01A Homicide
0110, 0130

02 Criminal sexual assault
0261, 0262, 0263, 0264, 0265, 0266, 0271, 0272, 0273, 0274, 0275, 0281, 0291, 1753, 1754

03 Robbery
0312, 0313, 031A, 031B, 0320, 0325, 0326, 0330, 0331, 0334, 0337, 033A, 033B, 0340

04A Aggravated Assault
051A, 051B, 0520, 0530, 0550, 0551, 0552, 0553, 0555, 0556, 0557, 0558

04B Aggravated Battery
041A, 041B, 0420, 0430, 0450, 0451, 0452, 0453, 0461, 0462, 0479, 0480, 0481, 0482, 0483, 0485, 0488, 0489, 0490, 0491, 0492, 0493, 0495, 0496, 0497, 0498, 0510

All other crimes were classified as non-violent.

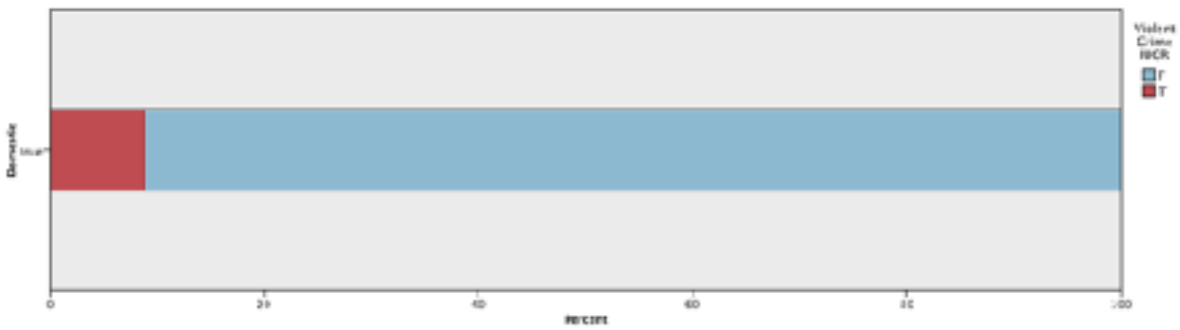
Appendix B - Slide 17

Although domestic violence is included as a variable in the dataset, its definition is based on the Illinois Domestic Violence Act and did not completely coincide with the definition of a violent crime as determined by the IUCR codes (4,5). Only 9% of crimes classified as domestic violence met the criteria of violent crime, as defined by IUCR. Due to the delicate nature of domestic violence and its special impact on families and communities, it deserves a full assessment which is outside the scope of this analysis.

Domestic Crimes Violent and Non-violent Classification

Violent Crimes
Count = 147
Percent = 8.914

Non-violent Crimes
Count = 1502
Percent = 91.086



References - Slide 18

1. Gorner, J. (2017, March 03). Few answers as Chicago hit with worst violence in nearly 20 years. Retrieved April 23, 2017, from http://www.chicagotribune.com/news/local/breaking/ct-chicago-violence-2016-met-20161229-story.html
2. Sanburn, J., & Johnson, D. (2017, January 30). Violent Crime Is On the Rise in U.S. Cities. Retrieved April 23, 2017, from http://time.com/4651122/homicides-increase-cities-2016/
3. Chicago Police GIS. Retrieved April 23, 2017, from http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html#N04B
4. Illinois State Agency Links. Retrieved April 23, 2017, from http://www.isp.state.il.us/crime/domesticviol.cfm
5. Chicago Data Portal. Retrieved April 23, 2017, from https://data.cityofchicago.org/view/5cd6-ry5g