CrimeIntel

A data-mining approach to public safety

Max Smith
maxs@colorado.edu

Aaron Shyuu
aaron.shyuu@colorado.edu

Thomas Dolan
thomas.j.dolan@colorado.edu

ABSTRACT

Multiple data mining techniques were applied to the popular *Chicago Crime Dataset*, which is a dataset maintained by the city of Chicago and provides crime data within the city for the period of 2001-2020.

Exploratory data analysis (EDA) indicated two distinct periods in our crime data: 2003-2015, where annual crime decreased steadily, and 2016-2019, where annual crime plateaued. Linear regression techniques were used to determine that most of the plateau was driven by three crimes: theft, battery and narcotics.

Additionally, outlier analysis revealed that March-July 2020 showed significant decreases in total crime. Domestic crimes as a share of total crime increased significantly during this period, indicating that this might be a direct result of COVID-19 stay-at-home orders issued by the city around this time..

We also wanted to look at what we could learn about where and when crime was being committed in the city based upon the many categorical attributes in the crime data. Furthermore, we wanted to see if by applying additional crime type categorizations derived from the FBI's uniform crime reporting program, whether that would be useful in providing additional insights. Doing so allowed us to verify at least three three major crime hotspots as well as provided additional details on the type of crime being committed in each and in some cases when. We gained several entry points for additional analysis based on these results and proved that by adding the additional categorizations some new insights could be gleaned, making this information useful both in current and future analysis.

While we were exploring the number of arrests that were made for crime from 2002 to 2019, it became evident that there are significantly more arrests than non-arrests when we visualize with a pie chart in our EDA. This brought up an interesting question on the factors that contribute to the arrest of a criminal and whether we could effectively predict if an arrest would be made for a given crime. Decision Trees, Naive Bayes, Logistic Regression and K-Nearest Neighbor are data mining techniques used to compare the different models based on evaluation metrics including accuracy, precision and recall.

After classifying *arrest* based on a selected number of crime attributes, the model performances were improved with parameter tuning and feature selection. The results from the classification problem indicate that domestic crime, crime against society, crime against property, and index crimes are good predictors for the arrest (or non-arrest) of criminals. Given these predictors, the Decision Tree, Naive Bayes and Logistic Regression models also predicted arrest better than the KNN model in terms of evaluation metrics and computation time.

INTRODUCTION

Our first question was "what is causing the crime plateau between 2016-2019, and how might it be addressed?". This question is important because residents and public servants of Chicago may have become acclimated to decreasing crime over the period 2003-2015. The sudden stop in crime reduction may have effects on the quality of life for the average Chicago citizen. Further, it is likely that

many crimes continued decreasing at pre-2016 rates, leading us to believe that perhaps a handful of crimes may be increasing significantly and thus causing the plateau.

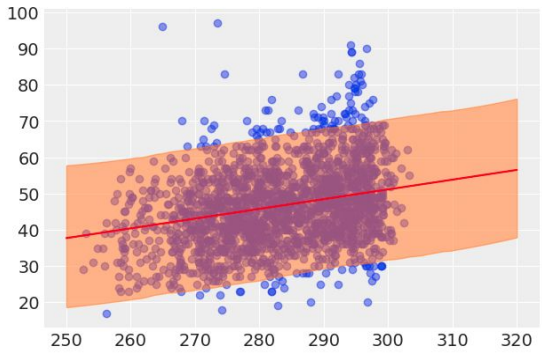
Our second question was "does economic activity correlate with crime?". This is important because if economic measurements are shown to be leading indicators for crime, decision makers in Chicago could use them to proactively allocate law enforcement resources.

Our third question was: What can we learn about where and when specific kinds of crime are happening by analyzing associations between crime type categorizations and their location and time data? Does adding more crime categorization information reveal any additional insights to these results? The first question is important because not only does it help organizations know how much of their resources to apply to a particular problem, but it may help show what resources in particular are required and how specifically they might be deployed. It may also reveal where certain solutions to the problem will fall short, and where more creative or progressive solutions could benefit from added support. Perhaps more importantly, it helps us to understand the nature of the problem, which given the difficulty of solving crime in Chicago and elsewhere and the politics surrounding it, that is the most important and realistic of our goals for this project. In the same vein, by adding new crime categorizations there is the possibility that completely new insights might be gained by revealing new angles from which to look at the problem, which is invaluable in terms of framing the problem in new ways which may lead to new solutions by challenging the status quo. Proving the veracity of this claim provides additional proof that implementing updated uniform classification systems such as the FBI's NIBRS system can lead to real improvements in how crime is understood and addressed in not just Chicago, but all cities and communities who choose to adopt it.

Our last question was "can we predict whether an arrest will be made for a commited crime?" This is a critical question since law enforcement are constantly looking to take preemptive measures to minimize loss and allocate resources efficiently. By understanding the likelihood of arrests and how certain properties of crimes make arrests more difficult, criminals with a given set of descriptions could be targeted first. This also makes the data mining model of our choice imperative for better predictions. After predicting arrest with our classification data mining models, how can we improve the evaluation metrics of each model with parameter tuning and feature selection? An accurate and effective model that can predict the possibility of arrest can potentially guide law enforcement resources and distribution of effort.

RELATED WORK

Many analyses relating to Chicago crime have been done including those that have used our same dataset. Some of these projects exist as notebooks accessible on Kaggle's website, while some others are hosted on personal blogs or sites. Four of these in particular provided inspiration for our project, two of which utilize data mining / KDD methods. The project "Chicago Crime on Christmas" helped us realize the benefits of integrating additional data types to yield novel insights, and the frequency of the data we would need in order to do so successfully [3]. A Python project by Cjango provided a vis heavy approach that helped us get a better sense for the data we would be working with and what kind of questions we could be asking [4]. Two separate data mining projects using different tooling than us were especially useful for visualizing how the data mining methodologies we have been learning in class could be applied to our dataset [5, 6]. In both cases we felt the results were somewhat inhibited by being limited to only the categories provided in the original dataset, which motivated us to work towards including data from additional datasets to improve our findings. Doing so allowed us to derive additional attributes, some of which improved our data research by providing effective hierarchies within the data.



Excerpt from Mikko Karkkainen's "Chicago Crime on Christmas" [3], showing a correlation between temperature and incidents of crime.

In addition to these projects, many academic and civic studies have also been done. A data driven article published by Northwestern helped give us context surrounding the problem, potential causes, and possible solutions [1]. This guided some of our questions and we were happy to find that many of our results validated the information we learned in this article. An academic article by Julia Burdick-Will on the relation between violent crime and academic achievement in Chicago schools helped to illuminate the challenges associated with identifying causality and prescribing solutions to crime given its endemic nature and the role that data plays in this. This contributed greatly to our recognizing the inherent biases in our data during the data mining process, helping to understand the limitations of our approach.

DATA SET

Our primary data set is the Chicago Crime dataset, hosted by Google's Big Query. A link to the dataset and description can be found on Kaggle, at https://www.kaggle.com/chicago/chicago-crime [9]. The data is generated and owned by the Chicago City Police. Each record represents a reported crime and includes information regarding its time, location, and various categorizations describing the crime. The dataset contains 7,213,094 entries from 2001 to 2020.

Our secondary data set is the CFNAI Historical (Real-time) data set, created and maintained by the Federal Reserve Bank of Chicago. The CFNAI metric is a monthly average of 85 series of economic data representing a wide range of economic indicators, each of which has been weighted, normalized, and adjusted for inflation. Each record represents the CFNAI metric for a given month, and includes four similar metrics each of which is derived from a subset of the 85 series whose indicators fall into one of four categories, 1) production & income, 2) employment, unemployment & hours, 3) personal consumption & housing, and 4) sales, orders & inventories. It is available for download at https://www.chicagofed.org/research/data/cfnai/historical-data.

MAIN TECHNIQUES APPLIED

Our project was broken down into three distinct sections and their corresponding notebooks. Data preprocessing occurred in our ETLPipeline notebook. Here, we used SQL to query the public dataset hosted by Google. The results were put into a Pandas dataframe and subsequently cleaned and integrated. Before performing any Exploratory Data Analysis or data mining techniques, we ensured that our raw data are transformed into useful and efficient formats. The Chicago Crime dataset is a huge dataset consisting of millions of rows. This means that we started data cleaning by verifying any missing/null values and addressed them appropriately. Through some initial filtering, it became clear that the majority of our missing values occur in the location and coordinate attributes. Due to the size of our dataset and the relatively small number of rows that contain at least a column with a missing value (< 0.1% of our dataset), we safely removed them with minimal loss. Next, we checked if there was any redundancy in our data and found no duplicate rows that take up unnecessary space. To validate the quality of our data, we reviewed certain attributes that could be more prone to input errors, such as the FBI attributes and made corrections as needed.

In terms of data reduction, we removed attributes that are not beneficial to EDA or data mining including case number and update time. Moreover, ID attribute columns were dropped as they do not contribute to the data mining process.

To find correlations between the Chicago Crime dataset and economic dataset, we had to join the two datasets through the date attribute. This gave us a combined table where we are able to see the values of crime attributes and economic attributes for a given date. (represented in an object or row of the table). In order to find correlations and associations, we also made some transformations which included rearranging our data and grouping by crime types. We also derived nominal attributes—crimes against person, crimes against property, and crimes against society— which are useful for multilevel pattern analyses and our data mining models. Furthermore, new attributes such as fbi code descriptions and weekend (a binary yes or no) was created to support classification and model construction. In essence, we derived many additional nominal attributes to aid our data mining process.

For further data preprocessing, we also converted a lot of data (with many different data types) to numbers in separate notebooks so we can use them for data mining techniques. Firstly, we converted the different attribute values to numbers. For example, we converted the arrest attribute to binary (1s and 0s) and the description attribute to the respective encoded numeric values. When we performed different conversions, we were attentive to use the respectively correct methods.

Not only did our data cleaning include dropping null values and unnecessary attributes, we also converted various data types like latitude and longitude to geopandas objects. The cleaned dataframe was then exported to a .CSV file for subsequent analysis.

From there, we completed an exploratory analysis of our data (our EDA notebook). This mainly involved generating descriptive statistics and visualizing relationships between variables. This analysis resulted in the formation of our main data mining questions (see *Introduction*). Our questions were then moved to individual notebooks as they required specific libraries. (however, question 1 and 2 are in a single notebook).

To answer our first question (What is causing the crime plateau between 2016-2019, and how might it be addressed?), we modeled the trends of the top 20 crimes using linear regression for a) the decline years (using the subset of 2012-2015) and for b) the plateau years (2016-2019). The slopes were then compared to determine the main drivers of the plateau.

To answer our second question (Does economic activity correlate with crime?), we compared the various economic indicators with total crime using scatterplots. Although no correlation was shown, a global outlier was found in our economic data during this process. We then used contextual outlier analysis (filtering crime data and calculating z-scores) to see if this corresponded to an outlier in our crime data. Although no outlier was uncovered, we noticed a clear substantial drop in crime occurring during the “outlier” period from our economic dataset.

To answer our third question (What can we learn about where and when specific kinds of crime are happening by analyzing associations between crime type categorizations and their location and time data?), we used frequent itemset mining to create association rules based on the full dataset, which were then evaluated for interesting insights. This involved two parts, the first of which was laying the infrastructure required to perform the frequent itemset mining and the second of which was the actual data mining itself.

In order to frequent itemset mine, we chose to use the apriori algorithm despite some of its shortcomings in terms of speed. It was the easiest to implement and understand the results of for a beginning data miner. In order to perform the mining, we needed to remove all of the non-categorical attributes we would not be using and then convert them to integer values representing each unique value in the dataset. We then needed to

reduce the data further into a variety of subsets for the actual frequent itemset mining in order to ensure doable runtimes and interesting results. Of the 1,048,575 possible, we selected 18 based upon how the number of unique values in each attribute related and with the intention of achieving a healthy mix of attributes falling into either the what, where, or when of each crime. We defined functions to create these subsets, convert the data, perform the mining, and finally to lookup the original values and display them as association rules in a human readable format.

With the infrastructure in place, we performed the apriori data mining on each of the subsets over many iterations. The frequent itemset mining is a very dynamic process, requiring the minimum support, minimum confidence, and minimum lift to be tweaked between each run in order to maximize the interesting results while not producing too many results that take too long to mine. Each set of results had to be analyzed by a person for validity, as in each case there are always association results that are either so obvious they are wholly uninteresting, or worse the result of overopular values in the data obscuring the more interesting results and even biases in how the data was recorded leading to high associations that are more or less fabricated by the context.

To answer our last question (Can we predict whether an arrest will be made for a commited crime?), we used classification data mining techniques to predict arrest (or non-arrest) based on a selected set of crime attributes. Decision trees, Naive Bayes, Logistic Regression and K-Nearest Neighbor are models applied and compared with evaluation metrics to determine the most suitable model for this predictive task.

Before constructing our classification models, we also needed to choose the categorical attributes that will serve as inputs to our models. After selecting those crime attributes, we preprocessed their values into appropriate form that is acceptable to our model. For example, the domestic attribute, indicating whether a crime is domestic or not, is converted to 1 or 0 respectively.

For all of our classification models, we chose to split our dataset to a train and test set. To be precise, 75% of our dataset was used for training and the rest of the data was used for testing. We chose to use a train and test split method even though there are other validation options such as train/test on the same data and k-fold cross validation. This is because training and testing on the same dataset can lead to overfitting and k-fold cross validation is computationally slow.

The first data mining technique we used for predicting arrest was logistic regression. Since logistic regression is an effective model for binary classification (our arrest attribute has classes 0 or 1), we expected very good results from our evaluation metrics. Knowing that our arrest classes have a strong bias towards non-arrests (ie. there are many more 0s than 1s in our crime dataset for arrest), we had to make the class weight balanced. This essentially replicated the smaller arrest class until there are as many samples there as in the larger class, increasing the balance of the two classes. After training the data, we applied our test set (25% of our dataset) for prediction to analyze the performance of our logistic regression model.

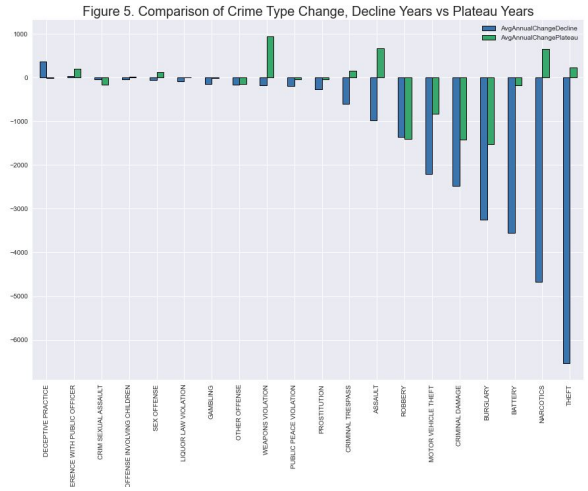
Similarly, we predicted arrest with a decision tree which uses entropy as the attribute selection measure. We were curious about the impacts of attribute selection measures on our evaluation metrics so we also created a decision tree with Gini impurity to compare the two results. Furthermore, we were able to plot the decision tree and predictions can be made by traversing the paths of the tree. With a similar workflow, we tested our prediction with a Naive Bayes model which essentially predicted the class membership probabilities.

Lastly, we utilized K-Nearest Neighbor to predict arrest. To deal with the imbalance in the arrest classes which would impact our results, we had to down sample our non-arrest class so that the number of arrest objects and non-arrest objects are

equal. Using Euclidean distance as our distance function and selecting an appropriate k, we were able to train and predict arrest at a slower rate than other models. We concluded our analysis with model performance and insights (see Key Results below).

KEY RESULTS

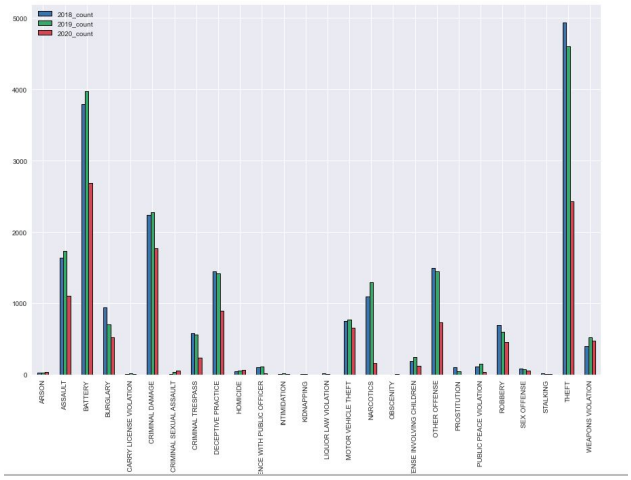
The key takeaway from question 1 (What is causing the crime plateau between 2016-2019, and how might it be addressed?) was that approximately 70% of the plateau was being driven by three crimes: theft, narcotics and battery.



Modelling the trends in various crimes during the decline years (in blue) and the plateau years (green) revealed that while theft, narcotics, and battery were decreasing by approximately 6,000, 5,000, and 4,000 instances per year (respectively), those annual reductions stopped after 2015.

This is not particularly surprising given that these three crimes reside in the top five crimes from the period 2002-2020. However, this insight can be used to guide policing in Chicago in an effort to get back to pre-2016 crime trends (see "Applications").

Question 2 ("Does economic activity correlate with crime?") first appeared like it would not turn up any useful patterns using traditional correlation techniques. However, when outlier analysis was applied to our data, it revealed an interesting pattern in 2020. The outlier in economic data pointed us toward a stunning reduction in crime starting in March 2020. Controlling for seasonal cycles and time periods (plateau vs decline), we found that monthly crime has been approximately two standard deviations below the mean since March of 2020.



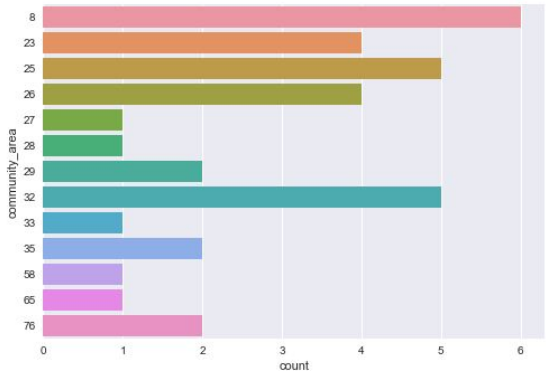
April 2020, when compared to April 2018 and 2019, showed a marked decrease in most crimes.

This drop in crime correlates to the COVID-19 restrictions seen in Chicago around that time [11]. When we looked at the ratio of domestic crimes to total crime, we found that domestic crimes as a share of total crime increased during this period, indicating that this drop in crime was likely due to COVID-19 stay-at-home orders. The rationale behind this assumption is that more people will be at home than in public, resulting in an increase in domestic crime percentage.

Question 3 was "What can we learn about where and when specific kinds of crime are happening by analyzing associations between crime type categorizations and their location and time data? Does adding more crime categorization information reveal any additional insights to these results? While our original exploratory data analysis revealed that larceny, battery, criminal damages, and drug abuse were the four most prevalent crimes, by employing frequent data mining we learned that drug abuse and larceny were focalized in several regions. Drug abuse is associated with community areas 23, 25, 26, and 29, or Humboldt Park, Austin, West Garfield Park, and North Lawndale, respectively. Drug abuse in these areas is most

heavily associated with 2003, 2006, and 2007. Historically these communities have a history of violence and gang activity, which the prevalence of drug abuse indicates. If there is a larceny, there is a good chance it is in community area 32, or "The Loop", Chicago's central business district and heart of downtown. Perhaps more interesting, we learned that crimes in this area are most associated with the hours between noon and 5pm. Just next door, in the Near North Side where the Navy Pier and many other tourist attractions exist, pickpocketing, credit card fraud, retail theft, and fraud are also heavily associated with the area. This is especially interesting when we consider the new information that besides the typical department stores and small business stores, a lot of the crimes in these areas are occuring in bars or taverns, restaurants, and hotels, indicating that a lot of this crime is linked to the tourists in the area.

Departing from these favorite areas of major crime, we also identified the primary sources of several non-major crimes. In some cases, these associations turned out to be in large part due to biases in the data, but not all. Where there is a crime for prostitution, we have an association rule saying that it will be in community area 28, or the Near West Side. This area is large and lies between the previously mentioned trouble area of 23, 25, 26, and 29 and The Loop, or downtown Chicago. The area has a long history of crime dating back at least to the 70's, but has undergone a lot of redevelopment since the early 2000's, so it is an outstanding question what the cause of the ties to prostitution there are, or whether it is an artifact from the earlier years in our data. That answer would require further discovery. Perhaps the strongest association we found anywhere in the data set was the link between "Other Weapon Violation" and community area 76. As it turns out, the community area is O'Hare, where the airport is located. We speculated that this could be the result of people attempting to carry disallowed items onto planes, which required its own description given that not all of those items would be considered weapons in a different context and likewise would not necessarily be a crime anywhere else, leading to an extremely high correlation between this description and location. This was more or less confirmed by the later association rule telling us that, if there is a crime in O'Hare, it is extremely



likely that it occured in either the Upper or Lower Terminal of the airport. Equally biased was a similar association found between ‘theft by lessee’ and the O'Hare area, which obviously points to the fact that this is where the majority of car rental locations are, and so is where the crimes are reported. .

Aside from associations relating to specific community areas, we also found some potentially interesting associations between location descriptions and various types of crime, as well as other location types. Most obvious was the prevalence of larceny taking place in department stores, in particular the Loop area, but also throughout Chicago in which case it was also heavily associated with grocery stores. Perhaps just as obvious was the high association with drug abuse occuring on sidewalks. This becomes a bit more interesting when taken in conjunction with the finding that drug abuse on the sidewalk is heavily associated with community area 25 and that sidewalk crime alone is heavily associated with the West Ferdinand St block, which tracks laterally through the Humboldt and West Garfield park areas. Taking these rules together begins to paint a picture of the drug abuse crime and homelessness issues in the area and tells us where exactly the drug abuse on sidewalks is most prevalent. We also found that if a crime is committed in a residence garage, it is most likely to be a burglary. We also found that, of all crime committed in garages, most were described as

"forced entries". At face value this may not mean much, but again helps to paint a picture of how and where these crimes are being committed that could be useful not only to law enforcement, but individuals or security companies looking how best to protect homes, who now know that garages are the most vulnerable places to burglary via break-ins. Finally, there were what we considered a surprising amount of crimes associated with hallways, stairwells, elevators, and apartments occuring in the community area 28. This is the same Near West Side where prostitution is most likely to take place. In general, crime appears to be most common on the street in Chicago, except where larceny is taking place in department stores and grocery stores, however here it seems that there are far more crimes being reported off the street and inside buildings. Given the history of crime in projects in this area, it is reasonable to suspect that many of the apartment complexes in the area are a root cause for this. Speaking of apartment crime, we also found an alarming set of almost identical crime associations with apartments along the st blocks of E 68th St, E 70th St, E 78th St, E 80th St, E 81st St, E 82nd St, and S South Shore Dr. All of these blocks run through the South Chicago and Avalon Park neighborhoods, which until this point have not shown up in our frequent itemset mining. We believe this would be another great entrypoint for additional discovery.

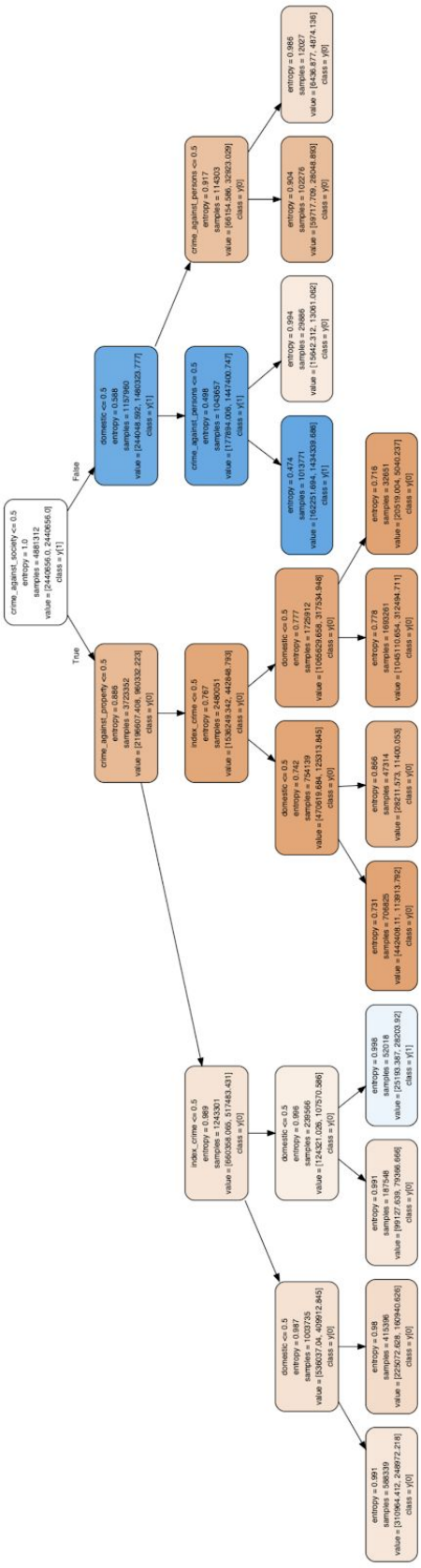
Finally, besides mining the more specific attributes for insights, we also took a look at the broader categorical attributes in an attempt to gain broad stroke information that might lead to further analysis. Mining these attributes is particularly difficult because of the need to significantly lower support and raise lift while fine tuning the confidence inputs in order to filter out the high number of uninteresting matches, however we did find a few useful pieces of information. We found that in ward 28, the drug abuse crimes were more likely to occur at 7pm and 4am. We also found that larceny in ward 42 is especially associated with years 2017, 2018, and 2019, which is significant given the large role larceny has proven to show in the plateau of crime in the last four years. We also found that in community areas 25 and 23, non-index, non-domestic crime is likely to lead to an arrest, which is very surprising because it apparently contradicts associations made with domestic crime in general. Generally speaking, if a domestic crime is non-index it is strongly associated with a non-arrest, and if it is an index crime (more serious crime category), then it is *very* strongly associated with an arrest. This means whether or not a crime is index or not tends to be the determining factor whether there will be an arrest for index crime. However, in these communities domestic crime that is non-index is still being associated with an arrest, telling us that law enforcement in these areas is more strict at least when it comes to domestic crime. Another interesting piece of information we found that we did not expect was that while a crime against property or society could be expected to occur on a weekday, which we would assume is the case given there are more weekdays than weekend days, if it is a crime against a person we have an association rule with fairly high lift telling us it will be on the weekend. Crimes against persons include homicide, involuntary manslaught, criminal sexual assault, aggravated assault and battery, simple assault and battery, criminal sexual abuse, and offenses against family. Crimes against persons have the largest share of both index (more serious) and violent crime. Given this rule, it is safe to say that the weekend is typically more dangerous for people than the weekday.

The key takeaway from our last question (“Can we predict whether an arrest will be made for a commited crime?”) is that domestic crime, crime against society, crime against property and index crimes are good predictors for the arrest or (non-arrest) of criminals. Given that these predictors are inputs to our data mining models, Decision Tree, Naive Bayes and Logistic Regression predict arrest better than K-Nearest Neighbor based on evaluation metrics and computation time.

In terms of model comparison with evaluation metrics, the Decision Tree model gave the highest accuracy out of all of the models with

83.5% followed by Naive Bayes and Logistic Regression models at 82%. These 3 classification models are all relatively decent for predicting arrest based on a set of attributes including "domestic", "index_crime", "crime_against_property" and "crime_against_society". However, accuracy metric for these models can still differ by a few percent based on a difference of one crime attribute. For example, Decision Tree obtains the highest accuracy with "domestic", "index_crime", "crime_against_persons", "crime_against_property" and "crime_against_society" attributes but Naive Bayes performs the best with "crime_against_persons" substituted with "property_crime". Looking at the precision and recall of the different models, we found that the decision tree has an impressive average precision of 0.83 and recall of 0.84 with logistic regression as a close second at 0.81 and 0.82 respectively. Knowing that precision measures exactness and recall measures completeness, this is a good sign for both models as they are both close to the perfect score of 1.0. Lastly, we analyzed the result even more with a confusion matrix. This is extremely beneficial since the True Positive, True Negative, False Positive and False Negative counts tell us the exact number for successful and failing predictions. For example, the decision tree confusion matrix indicated that 1,359,192 arrest values are predicted correctly (sum of TP and TN) while only 267,912 are predicted incorrectly. This is a nice improvement from our logistic regression model.

K-Nearest Neighbor can be improved with increasing input data to the model but performs the worst in terms of evaluation metrics and computation time. It is clear that the decision tree model is great for interpreting classification decisions. The tree structure is extremely intuitive and easy to traverse as we split on attributes until we reach a prediction class. As we can see from our decision tree, the classes give 1 or 0 for arrest and not arrested respectively. This allows us to make case by case predictions or evaluate predictions based on a test set (which we did)



The Naive Bayes model supports incremental data and runs extremely fast in comparison to the other models we tested. On the other hand, the speed of the KNN model makes it great for smaller datasets but not for this large crime dataset. For k in KNN, there was not a big shift in accuracy with the tuning of the k parameter. Since KNN is lazy learning, it also doesn't generalize our data in advance and only starts learning when new instances need classifying.

In order to further improve our model accuracy and other evaluation measures, we used two methods for feature selection. (Please see the progression of accuracy in ClassificationDataMining notebook) We plotted the 5 most important features using a randomized decision tree classifier (ExtraTreesClassifier) which samples a random subset of the feature space to give feature priority. We then used univariate selection to look at each feature individually. With SelectKBest, each feature's relative importance in predicting the output (true or false for arrest) is ranked. We used the ExtraTreesClassifier to select features for our decision tree models since it is essentially making a quick random version of a decision tree to test out the attributes. Furthermore, we took univariate selection into account for the building of our other classification models. Our results indicated that the feature selection process was crucial for improving our model performances.

APPLICATIONS

Our analysis into question one indicates that additional resources should be allocated toward law enforcement to help combat the current trend in crime. Additionally, any additional resources should go toward combating theft, narcotics, and battery.

Our analysis into question two indicates that the decrease in crime seen between the March 2020-July 2020 is likely due to COVID-19 restrictions and will thus be temporary. Our recommendation to decision makers in Chicago is not to make any structural changes to policing based on recent drops in crime, as they will likely revert toward previous trends once COVID-19 restrictions are eliminated in the coming years.

Our analysis surrounding question 3 gave some insights that could be applied immediately, but more than anything raised additional questions that would be worth exploring either by continuing to apply frequent itemset mining but on additional variations of subsets of the data, or by applying other data mining techniques or more traditional statistical analysis to questions raised. With our frequent itemset mining results, organizations looking to curb crime or enact positive change in the community can gain insights on how best to allocate their resources. Sidewalk crime tied to drug abuse in the Austin community area, particularly along West Ferdinand St, as well as the surrounding communities of West Garfield, Humboldt, and North Lawndale, indicates a strong need for public services addressing homelessness and addiction, as well as the curtailing of gang activities. The high rate of larceny in The Loop shows the ever presence of retail theft wherever department stores and grocery stores are central, which could be assisted in the knowledge that most retail theft occurs between noon and 5pm. The high association between crimes such as pickpocketing and credit card fraud as well as the increased association with crime occurring in bars, restaurants, and hotels in the Near North Side show the usefulness something like an ad campaign targeted at tourists might have. Generally speaking, the weekends are most dangerous. While not entirely useful in and of itself, this information is revealing given that crime overall does not increase over the weekend, which means the kind of crime being committed is changing somehow over the weekend and that should be explored further to better understand why. For people looking to protect their homes, securing their garages from break-ins and educating themselves on phone-based threats and fraud address the areas most vulnerable to crime from an external source. Finally, entities like community organizers and child protective services in addition to law enforcement should be aware of the high association between crime and apartments

in the South Chicago and Avalon areas, specifically on the blocks of E 68th St, E 70th St, E 78th St, E 80th St, E 81st St, E 82nd St, and S South Shore Dr. Enabling citizens in these areas with better resources and oversight could be an effective means of reducing crime at its source.

For those looking to perform additional data mining analysis using frequent itemset mining, we have shown that adding additional classification categories is an effective method to gain more insights. This is important information given that the Chicago crime dataset still uses the old SEDS crime FBI crime categorizations, which have less detail than the new NIBRS codes. This is an incentive for cities with old or non-existent uniform crime reporting to upgrade their records as soon as possible if they hope to understand crime with nuance and context. The results we found could also be used as a basis for reducing the dataset to more specific subsets in hopes of gaining additional insights that are not cluttered by the most popular crimes are biased by data that is no longer current. For example, larceny is heavily featured in our dataset, but tends to obscure some of the less common, but more significant, violent crime that is occuring in the city. Or, constraining the results to the communities in and around south chicago and Avalon may lead to more information regarding the type of crime occurring in the many apartments in this area.

Our analysis for our last question indicates that a good predictive model with feature selection can help law enforcement allocate resources by understanding the likelihood of arrests and how certain properties of crimes could make arrests more difficult. Knowing that "domestic", "crime_against_property", "index_crime", and "crime_against_society" are attributes that can predict arrests at a high accuracy, law enforcement could also find these attributes useful for predicting crime. Moreover, law enforcement can spend more time on analyzing these attributes to find their strengths and weaknesses in making arrests. For example, if a crime that is domestic and against property (for example, robbery) has a low arrest rate, this could be an issue of high priority that law enforcement can look at and dedicate more resources to improve this area. Another key application is that using the features of criminal incidents to predict arrests, law enforcement would be able to manage the prison population more effectively.

REFERENCES

[1] Mary Pattilo. 2018. Crime in Chicago: What Does the Research Tell Us? Retrieved October 2020 from https://www.ipr.northwestern.edu/news/2018/crime-in-chicago-research.html.

[2] Matt Ford. 2017. What's Causing Chicago's Crime Spike? *The Atlantic*. Retrieved from https://www.theatlantic.com/politics/archive/2017/01/chicago-homicide-spike-2016/514331/.

[3] Kaggle. 2019. Chicago Crime on Christmas. Retrieved October 2020 from https://www.kaggle.com/mkrkkinen/chicago-crime-on-christmas.

[4] Cjango. Chicago Crime Data Analysis (Python Project). Retrieved October 2020 from https://cjango.wordpress.com/portfolio/chicago-crime-data-analysis-python-project/.

[5] Kaggle. 2019. Chicago Codes - Crime Map and BigQuery in R. Retrieved October 2020 from https://www.kaggle.com/alkadri/chicago-codes-crime-map-and-bigquery-in-r.

[6] Sadaf Tafazoli. 2018. My Notes on Chicago Crime Data Analysis. Retrieved October 2020 from https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis-ed66915dbb20.

[7] Julia Burdick-Will. 2013. School Violent Crime and Academic Achievement in Chicago. *Sociology of Education*. PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831577/.

[8] Chicago Police. Crime Type Categories: Definition and Description. Retrieved from http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.

[9] Kaggle, Chicago City Police. 2018. Chicago Crime (BigQuery Dataset). Retrieved October 2020 from https://www.kaggle.com/chicago/chicago-crime/metadata.

[10] Federal Reserve Bank of Chicago. 2020. CFNAI Historical (Real-Time) Data. Retrieved October 2020 from https://www.chicagofed.org/research/data/cfnai/historical-data.

[11] City of Chicago, 2020. Coronavirus Response Center. Retrieved December 2020 from https://www.chicago.gov/city/en/sites/covid-19/home.html