CSI Chicago Final Report

Let's make less crimes happen

Friedrich Amouzou
University of Colorado Boulder
Boulder, Colorado
Friedrich.Amouzou@colorado.edu

Maria Knigge
University of Colorado Boulder
Boulder, Colorado
Maria.Knigge@colorado.edu

Maria Pazos[*]
University of Colorado Boulder
Boulder, Colorado
Maria.Pazos@colorado.edu

Brandon Stone
University of Colorado Boulder
Boulder, Colorado
Brandon.n.Stone@colorado.edu

ABSTRACT

The city of Chicago, Illinois has a significantly higher crime rate than the rest of the United States [3]. Researchers have investigated crime in Chicago for decades but the majority of studies exclusively focus on violent crime [4]. The purpose of this project is to find interesting relationships between the various types of crime that occur in Chicago, the date they occur, and their location. By data mining the Chicago Police Department's CLEAR database of reported instances of crime we could potentially find unintuitive information that can aid in the process of crime prevention and reduction. This project intends to provide a more holistic view of the geo-temporal relationships of crime in Chicago.

CCS CONCEPTS

• **Information systems → Data analytics**; Data mining; • **Mathematics of computing →** *Exploratory data analysis*;

KEYWORDS

Chicago, data mining, crime, research, CLEAR

ACM Reference Format:
Friedrich Amouzou, Maria Knigge, Maria Pazos, and Brandon Stone. 2018. CSI Chicago Final Report: Let's make less crimes happen. In *Proceedings of ACM Never Happened (FakeConference).* ACM, New York, NY, USA, 11 pages. https://doi.org/12.345/123_4

1 PROBLEM STATEMENT/MOTIVATION

Prior research on crime in Chicago has been done but the studies tend to be longterm or specific. Previous studies focused exclusively on gun violence [4]; or were completely comprehensive, but took 48 years for data to be gathered and analyzed [3]. We want to analyze available data on crime cross referenced with its geographical position to get a robust source to visualize and anticipate crime. We want to cross reference all crimes with their coordinate location to get a comprehensive view of all crime throughout the city and receive results sooner than 48 years.

To avoid simply highlighting dangerous neighborhoods, we are utilizing the types of crimes committed to form specific hot spots for specific crime. We want a comprehensive look at all crimes to educate and predict when and where these specific crimes tend to happen. We believe we can use the massive amount of location and time data to make a sketch of high-crime areas in the city, as well as use these attributes to predict future crime and, or arrests. We believe that past trends may

[*] goes by Sol

FakeConference, March 2018, Boulder, Colorado USA

ACM ISBN 123-4567-89-012/34/56.
https://doi.org/12.345/123_4

be accurate in defining a general geo-temporal profile, which will allow us to predict whether or not an arrest has occurred and potentially predict criminal incidents in the future. By analyzing the data deeper for connections we aim to build a relevant, useful, and accurate predictor of crime at a place and time.

2 LITERATURE SURVEY

Multiple universities have conducted studies on crime in Chicago in the past. These studies tend to take a long time to gather data or are specifically focused on a certain aspect of crime. Three major studies stood out when looking for previous work done in this field. There is little indication that any university is using gathered data to form visualizations that can help predict future crime. Many are analysis into social problems in relation to crime. Many of these studies are also city wide meaning the reader does not get the granular detail of crime by neighborhood.

2.1 University of Chicago

The University of Chicago does ongoing research with their Crime Lab which studies different social programs and policies to reduce crime and violence [1]. The Crime Lab is part of the University's Urban labs and was founded in 2008. The lab focuses on the cause of crime and developing and evaluating new ways to combat crime. The University of Chicago performs ongoing research on crime in Chicago. Although the University of Chicago can do ongoing research with accurate and relevant data, they are using it primarily to form help groups that aim to prevent and/or lower certain crimes. This is much more of a social project that is built on top of gathered data, there is no indication of visualizations of crime tied to location and time. There is also no indication of using gathered data on crime to predict future crimes in an area with any specificity past "crime was committed".

2.2 University of Wisconsin - Madison

Assistant professor Robert Vargas is focusing on the relations between gangs and student achievement [2]. His work is centers around the fact that, "You can build great schools and great community programs, but if these kids don't feel safe walking around in their own neighborhood, it's kind of a moot point." [2] This study was useful in showing general trends of crime. This document details how crime affects student achievement and vice versa. There was good analysis based on gathered data however it appeared to limit itself to specific districts in the city. This paper was based on student achievement and therefore did not cover all crimes in all neighborhoods of the city. It instead focused on violent crime within a certain distance to the nearest school. Although this data can help in forming visualizations of crime in these specific pockets, it will not scale well for a city wide application. There is also no attempt at future predictions of crime, something we are very interested in.

2.3 Yale University

The study conducted by Andrew Papachristos, Ph.D aims to be a comprehensive look at the trend of crime in Chicago. The study encompasses many attributes of crime around the city and took 48 years to gather and analyze data. Trends in crime over time were heavily analyzed in this study however not a lot of attention was given to the relation between geographical location and crime [3]. This was a very important study to utilize for our project. This is a major, comprehensive study on city wide crime in Chicago. This information is extremely helpful in making city wide visualizations. This study also was conducted over the course of 48 years. We trust the data is accurate due to the source and sheer amount of time to conduct. The downside again is the lack of connection between crime and time and place in the city. This shows a strong general trend of all crime in the city. This information will help us build our visualizations as well as check some aspects of our own study by comparing results with the data from this study. We can even use this data for testing our own code.

3 DATA SET

The principal dataset, found here:

> https://catalog.data.gov/dataset/crimes-2001-to-present-398a4

is the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) database. It's a relational table containing over six million reported instances of crime in Chicago starting from January

1st, 2001 to a week before the present. Each instance of crime has twenty-two attributes associated with it:

3.1 Crime Identification

Two of the attributes, *ID* and *Case Number* are exclusively used to identify the particular instance of crime. There's also the *Updated On* attribute which indicates when the crime was last modified in the database. These attributes, along with the geo-temporal attributes identify which individual incidents are part of the same overall crime.

3.2 Types of Crime

Nearly a third of the data's attributes describe the nature of the crime. This includes nominal data like the literal descriptions of the crime in the *Primary Type* and *Description* attributes which categorize the crime and its severity in written word. Additionally, there are the *IUCR* and *FBI Code* attributes which provide the Illinois Uniform Crime Reporting and Federal Bureau of Investigation codes respectively: standardized government categories for the crimes. There are also two binary attributes *Arrest* and *Domestic* which indicate whether an arrest was made and whether the crime is classified as domestic.

3.3 Geo-temporal

The majority of the attributes indicate when and where the crime occurred. The numeric interval attributes *Date* and *Year* describe the time the crime occurred down to the nearest second[1]. The *X Coordinate,Y Coordinate*, *Latitude*, *Longitude*, and *Location* (which is a combination of the two former) describe the precise location of where the crime occurred. These attributes will be especially useful for mapping out our results. The rest of these attributes are nominal descriptions and categorizations of the location: *Block, Beat, District, Ward, Community Area, Location Description.* All together the geo-temporal attributes of this dataset provide a in-depth account of where these crimes transpired.

4 WORK DONE

4.1 Preprocessing

Because the dataset contains over six million entries, we will be removing any incomplete police reports (i.e., we will remove data containing NULL values). Rows that include missing data did not end in any reprimand, so they will be ignored.

Additionally, we will need to limit the dimensionality of our dataset. To accomplish this, we will eliminating uninformative and redundant features. Whether or not a feature is informative will depend on the type of analysis we are looking at. This will give us a rich view at the many aspects of crime while also eliminating false positives; false or incomplete reports, etc. Once cleaned, the data will process faster and output clearer results.

For predictive modeling, we will be utilizing the following features *Month, Day, Time, Year, IUCR, Domestic, Location Description, Latitude* and *Longitude* to predict the response feature *Arrest.* Only Latitude and Longitude are continuous, nominal values–the rest are categorical data types–therefore, additional preprocessing steps must be taken. This includes encoding categorical data and the implementation of dimensional reduction using Linear Discriminant Analysis or Principal Component Analysis. We will look at the application of each of these methods to our dataset.

4.1.1 One Hot Encoding. In order to generate models, we must first transform categorical variables into a usable form. One Hot Encoding first translates categorical data into numeric values, then it creates individual binary categories for these numeric values. Essentially, One Hot Encoding creates a sparse representation of categorical features. For example, if we were looking at the hour at which a crime occurred, One Hot Encoding would first assign each hour a numeric value (i.e., 12 am maps to 0, 1 am maps to 1, ..., 11 pm to 23). Next it creates features corresponding to the categories (Hour0, Hour1, ..., Hour23), where for each tuple, only one of the hours will contain a '1', indicating that this was the hour of interest. This increases dimensionality substantially, since for every one initial category, it is replaced by the number of unique identifiers in the original category.

4.1.2 Linear Discriminant Analysis. Linear Discriminant Analysis (LDA) can be implemented both as a classifier, as well as a dimensional reduction technique. The

[1] Although this is seemingly only accurate to the nearest hour

goal of LDA is to maintain discriminant information and maximize the separation between classes by projecting the feature-space onto a subspace of the feature-space.

4.1.3 Principal Component Analysis. Principal Component Analysis (PCA) is another linear transformation technique, but it is unsupervised. The goal of PCA is to maximize the variance of the data. PCA utilizes orthogonal transformations to project the covariances of the dataset into a new feature-space for which the components are linearly independent. The first principal component is the element with the greatest variance, and each of the subsequent principal components corresponds to the next greatest variance. PCA works best when the variances of the projected space taper off rapidly because these variances directly correspond to the amount of variation in the original model and we want fewer components to explain the variation.

4.2 Data Analysis

4.2.1 Heat Maps. After cleaning our data, we were able to make initial visualizations. One of our focuses for visualizing our data was with heat maps. We felt it would be the best way to visually show amount of crime in a given area at a given time. The primary traits we wanted to visualize were Total Crime (fig.3), Total Arrests (fig.4), and Total Domestic Cases (fig.5). This was our primary step in forming an understandable view of crime. Latitude and longitude data points were taken from our dataset and matched with appropriate police beats.

4.2.2 Training and Testing Data. In order to obtain an accurate model, we must split our dataset into training and testing subsets–we are using a 75-25 split. We will train our model then use the training set to determine the accuracy of our model.

4.2.3 Logistic Regression. Using logistic regression, we are able to create a model that can classify binary or boolean responses based on one or more covariates (features).

4.2.4 Support Vector Classifiers. Support Vector Classifiers (SVCs) are a generalization of the maximal margin classifier–SVCs create a separating hyperplane that maximizes the margin between classes. We will implement both linear and radial Support Vector Machines (SVMs) and compare to determine which provides the most accurate prediction of arrests. Linear SVCs can be applied when the data is linearly separable (i.e., the boundary between two classes is linear). SVMs, on the other hand, relax the linearity assumption and allow for boundaries to exist as nonlinear hyperplanes.

5 EVALUATION METHODS

We will evaluate the efficacy of our method to predict whether an arrest occurs by verifying that test data will generate an accurate prediction with 85-95% accuracy. A larger general range will be accepted as the number of arrests made is significantly fewer than the number of total reports (random sub-sampling has revealed that arrests occur in ~25% of the reports). We also seek to maximize the sensitivity (percent of true positives) and specificity (percent of true negatives) of the models. A major success for us would be to properly generate visualizations of crime city-wide; however, this is easier said than done.

6 TOOLS

We will be using Python and Jupyter Notebooks for this project, relying heavily on Python as our main tool. We will utilize Python's libraries, including numpy, pandas, scipy.stats, sklearn, etcetera. Sklearn is essential for the implementation of logistic regression and support vector machines.

7 MAJOR MILESTONES

- We expected the data processing to be the most difficult and time consuming part of this project. This portion of the project was to be completed by March 20th. In processing our data, we had to remove invalid and incomplete entries. This step also involves finding a map to overlay data.
- By April 3rd we have completed and debugged our code, and we should be producing results and graphs. Visualizations were important to us in showing specificity in crime with time and location. Completion of code involves having data overlay onto at least one map for visualization purposes.
- The rough draft of our paper was completed by April 19th. This paper is a formal summary of our process as well as our findings. Drafting can be sent to other groups working in similar projects.

This is to corroborate any findings and share data to build a more accurate view of crime in Chicago.

- Contact with other groups building similar projects by April 24th to see if we can share any relevant data. If there is any relevant data between our groups, we will include it in a "Similar Projects" section of our final paper and our final presentation.
- Final drafting of our paper and presentation will be completed by April 26th, in order to have enough time for last minute corrections and such before the due date, May 1st.

7.1 Milestones Completed

- Data Preprocessing and Processing: Data has been cleaned somewhat sufficiently, enough to run models against. We are able to take data and produce graphs and models that show rough trends. In total, 11 attributes were removed due to redundancy or lack of relevance. Those that remain are: Case Number, Date, Primary Type, Description, Location Description, Arrest, Domestic, Beat, FBI Code, Latitude, and Longitude. Additionally, any tuples with missing location data, i.e, Latitude, Longitude, and Location Description, were removed to allow proper geographic mapping. Lastly, a handful of tuples were far outside the boundaries of the city of Chicago and thus were removed as well.
- Dataset Description: Dataset columns have been researched and explained. Initial descriptions have been explained and can be found in "CrimeNotebook.ipynb" and uploaded to team Github. Descriptions were taken from dataset source.
- Code Debugging Round 1: Framework and general design for code has been set. We have discussed initial criteria that needs to be taken care of.
- Code Debugging Round 2: Framework has been coded and debugged for initial testing with dataset. We are able to draw models that display some simple analysis of the data. Reported occurrences vs. time has been graphed, latitude/longitude range and outliers have been graphed.
- Further Cleaning: Further removal of data. We need to further analyze the outliers with geographical data on Chicago. Whether these are mishaps in GPS data, an officer responded to a particularly violent crime, and other factors could account for these outliers.
- Displaying Results: We are able to generate a fair amount of visualizations with the cleaned dataset. This includes a visualization overlay on a map of Chicago. We are able to correctly overlay multiple crime instances onto the map.
- Finalization of Data Presentation: We are able to visually display our cleaned data as well as apply some amount to predict future crime. Data and findings will be added to our final paper and presentation.

7.2 Milestones To Do

- Final Drafting: Visit other groups studying crime to see if there is space to share data. Merging shared data into final drafts of paper. We can compare results as well as processes to see if there is room for improvement.
- Further Data Analysis: If time allows we can think of more creative links within the dataset. See how many ways we can model crime to form an accurate big picture of crime.
- Presentation: Present our findings to the class.

8 KEY RESULTS

So far the majority of the work done has been data preprocessing and exploratory data analysis. No new patterns where mined for but we developed a basic understanding of the data.

8.1 Dates

Overall there has been a downward trend in crime. There's a consistent yearly pattern of peaks and lows in crime. During the summer months, crime invariably reaches its yearly maximum and the same applies to winter yearly minimums.

8.2 Location

Naturally, the reported locations of crime spanned the entire city of Chicago. However, while the latitude is evenly distributed, the longitude is heavily biased eastward. This is also made apparent in the geographical plots: maps of Chicago segregated by police beats[2] and

[2] The beat map is of the most current police beats but the crime data plotted against it ranges from the present back to 2001

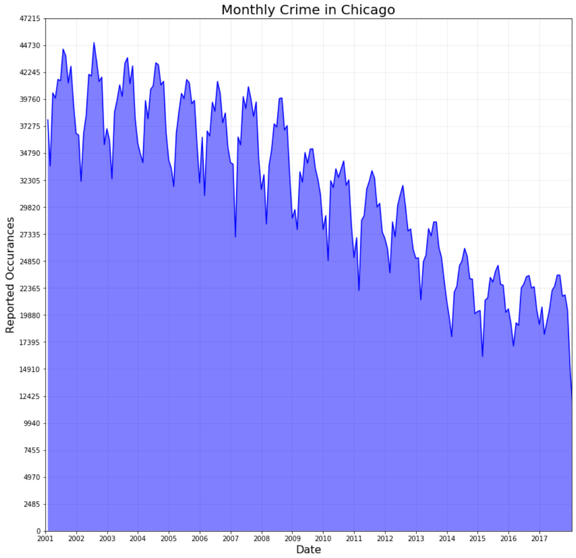

Figure 1: Crime frequency per month

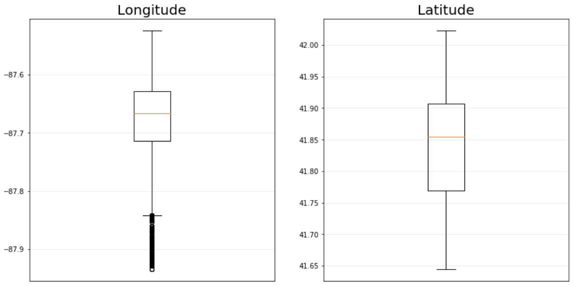

Figure 2: Latitude and Longitude distribution

color-mapped proportional to the variable of interest.[3] A beat is the smallest police geographic area; each beat has a dedicated police beat car. All three plots share similar hot-spots for crime. What wasn't apparent in the box-plots, however, is that crime is focused southward.

8.2.1 Total Crime. The total crime across police beats is poorly distributed. Beats with high concentrations of crime are directly adjacent to beats with low concentrations of crime. Generally, areas of high crime are distributed into multiple beats to help combat the increased crime rates, this means that low-crime beats are generally larger than high-crime beats.

8.2.2 Total Arrests. The map of total arrests across police beats is naturally similar to the map of total crime, however, the arrest map is much more evenly distributed. Additionally, the southern part of the city has lower relative arrest rates than its counterpart in total crime.

8.2.3 Total Domestic Crimes. The map of total domestic cases across police beats shares a lot with the map of total crime but there are a handful of differences. The northeastern part of the city has significantly lower relative rates of domestic crime. Also, it is more poorly distributed than total crime map, this is especially apparent at the hot-spots.

8.3 Types

Although a large proportion of reported crimes in Chicago are violent, the majority are non-violent. Stealing: theft, robbery, etc..., is the prevalent class of crime. There is also large amount of drug-related crime as well as trespassing and damaging property.

8.4 Predicting Arrests

In order to produce our models, we needed to significantly downsize our dataset. Ideally, we would be able to use Random Gaussian Sparse Matrix Projection, as it preserves the properties of the data while reducing dimensionality and/or the number of tuples in the dataset, but this was not possible using one computer and the sci-kit learn packages in Python. So we reduced the dataset by sampling approximately 1 million random rows from the dataset. The following tables contain our results:

Table 1: Logistic Regression

	Logistic Regression	**Logistic + PCA**	**Logistic + LDA**
Accuracy	0.8900	0.8886	0.8906
Sensitivity	0.5822	0.5850	0.5952
False Positive Rate	0.0261	0.0286	0.0289
Specificity	0.9739	0.9714	0.9711

[3] Black indicates lack of data

[4] The "OTHER" category is the cumulation of categories too small to be displayed on the chart. The "OTHER OFFENSE" category are crimes label as such by the CLEAR, IUCR system

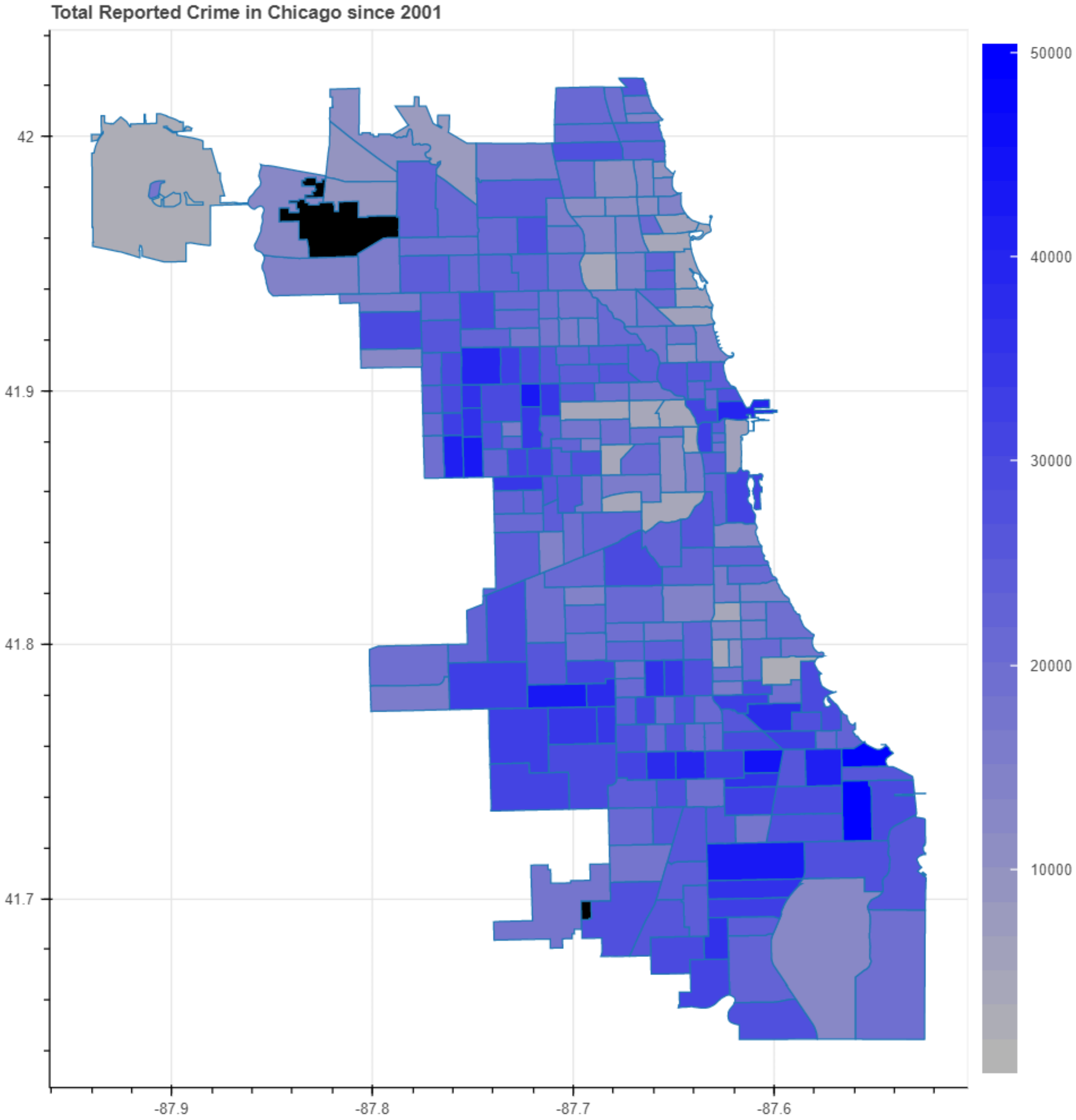

Figure 3: Total crime in police beats

Table 2: Support Vector Machines

	Linear SVM + PCA	Linear SVM + LDA	Radial SVM + PCA	Radial SVM + LDA
Accuracy	0.8661	0.8916	0.8915	0.892
Sensitivity	0.5426	0.5705	0.5700	0.5686
False Positive Rate	0.0329	0.0209	0.0207	0.0199
Specificity	0.9671	0.9791	0.9793	0.9801

Much to our delight, we were able to produce arrest predictions that fall within our desired range of accuracy. However, due to the small ratio of arrests to total records, we see significantly reduced sensitivity (or true

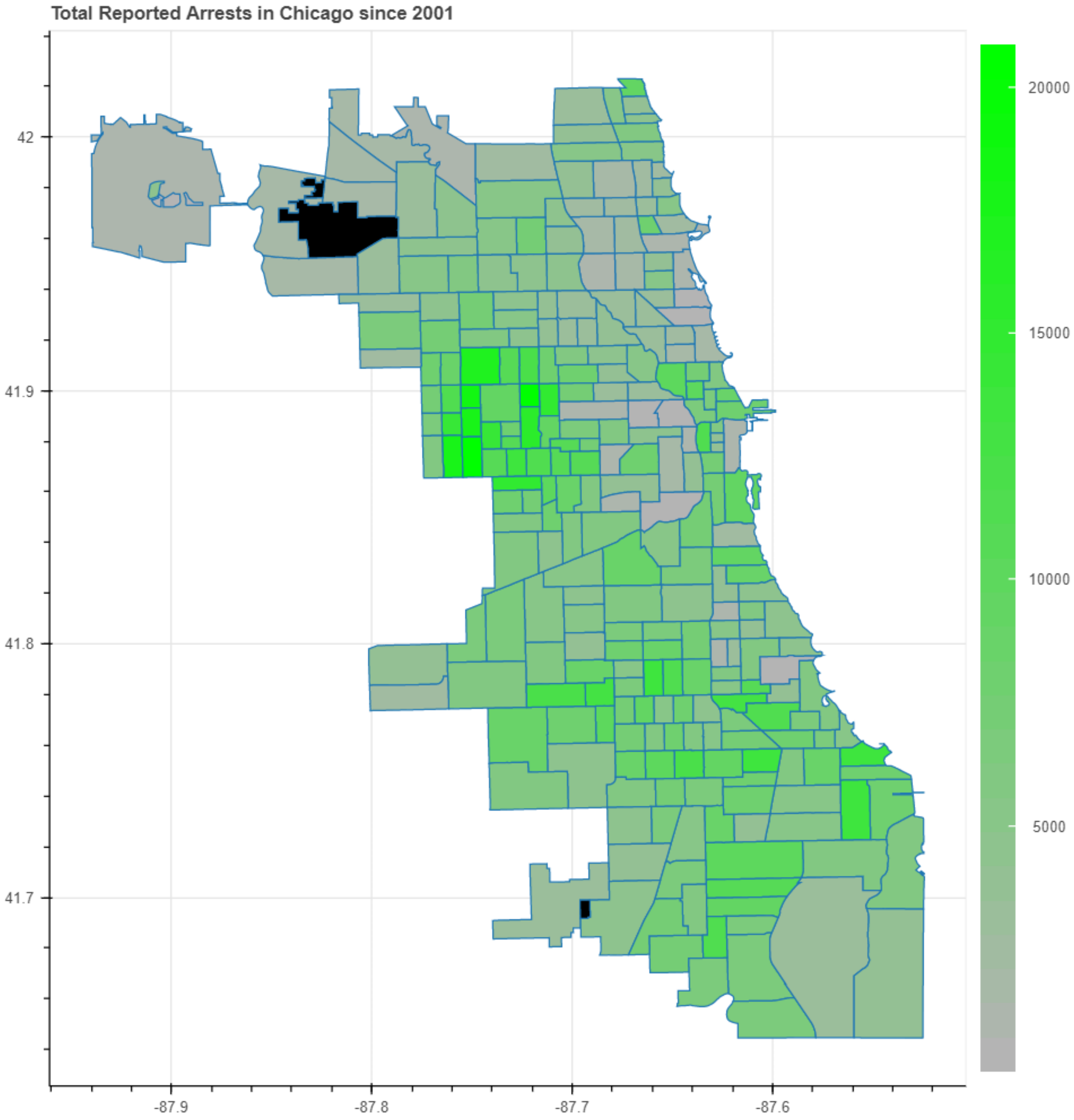

Figure 4: Total arrests in police beats

positive rate). This may also be in part to the downsampling that was necessary to generate our models and classifiers.

The full logistic regression model (without any reduction to dimensionality) performed almost as well (with an accuracy score of only 0.06% less) as the logistic model using LDA. While the model using LDA has the highest false positive rate (2.89% compared to 2.61% for the full logistic model), it has the greatest sensitivity (59.52%). Because the difference is minor, we can assume that Logistic regression with LDA is the

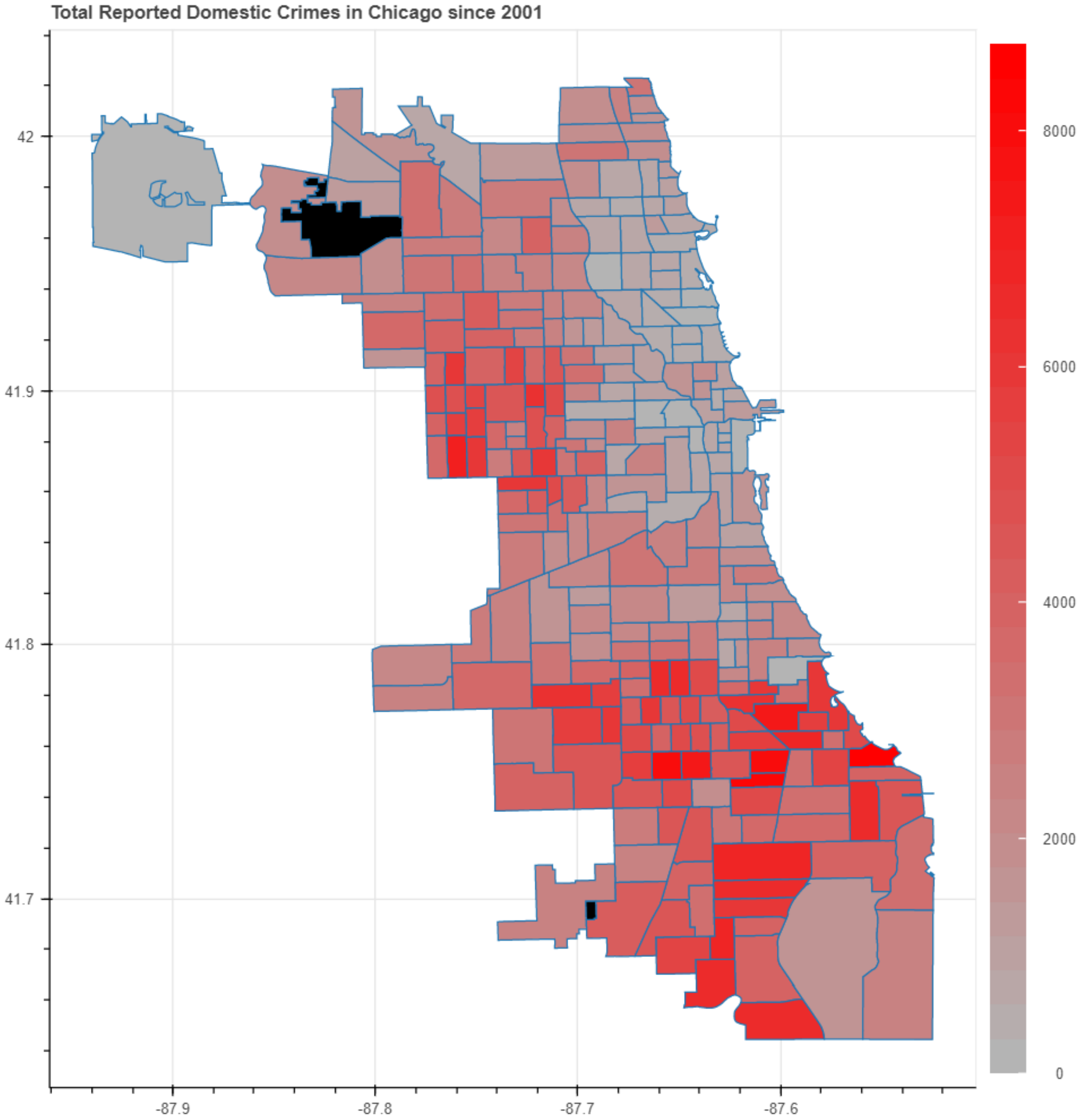

Figure 5: Total domestic cases in police beats

best option when using logistic regression to infer new data.

Looking at Table 2, we see that the linear and radial SVMs perform with roughly the same accuracy (except the Linear SVM with PCA, which failed on all fronts). The Linear SVM with LDA has the highest sensitivity of the support vector machines (57.05%), but it's error rate is the second greatest across the support vector classifiers. The radial SVM with LDA reduces the false positive rate most effectively (bringing it down to 1.99%) and predicts with the greatest accuracy (89.2%), but at the cost of sensitivity. If we compare the logistic model

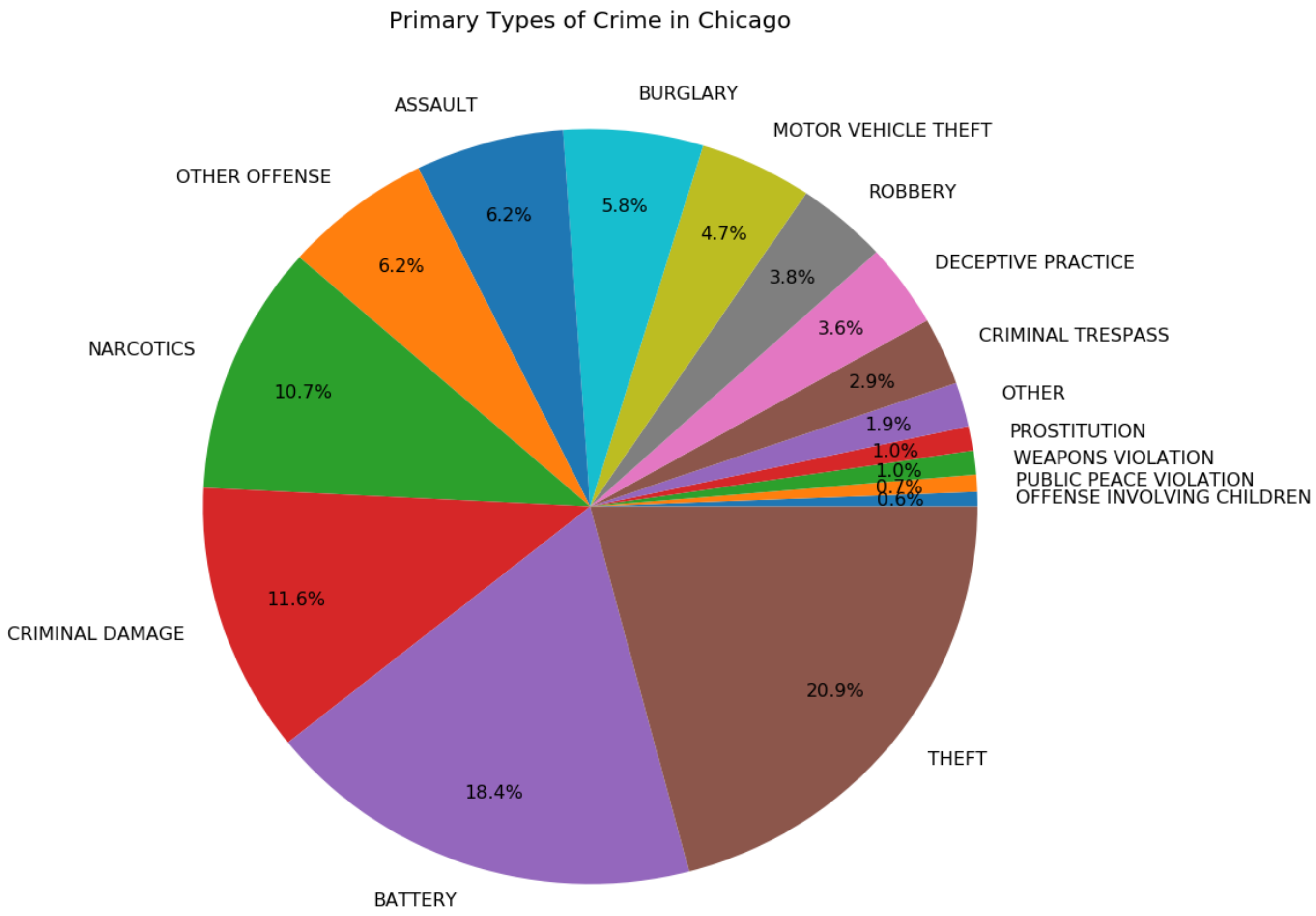

Figure 6: Prevalent types of crime[4]

using LDA with the radial SVM using LDA, we see that the radial SVM is more accurate by 0.14%, predicts fewer false positives by 0.9%, but predicts fewer true positives by 2.66%. There is a definite tradeoff when it comes to determining which model performs better. Ultimately, it comes down to whether we want a greater true positive rate (sensitivity) or greater predictive accuracy over all.

9 APPLICATIONS

9.1 Current Applications

The visualizations can be used to form a general idea of crime and arrests for individual police beats. We wanted to provide a number of visualizations so we could clearly show simple, and clear views of the relative violence throughout the city. As mentioned before, beats are drawn based on relative crime rates in the area so that more police are concentrated in areas with higher crime rates.

9.2 Future Applications

Police beats can potentially use this data to redraw beat districts to better apply their police force to areas that show increases in crime over time. Our predictions will show crime trends that we expect which will help to form a fully comprehensive view of crime. The visualization of total arrests in police beats can be used in conjuncture with the visualization of total crime in police beats to roughly but comprehensively demonstrate arrest rates per beat.

In the data we have collected, a reported crime does not necessarily imply an arrest was made in connection with that crime. Each case indicates whether an arrest was made, so one can easily tell which beats are more successful in finding suspects. This data can help identify beats that have unusually low arrest rates, so that officers in those beats can be evaluated for incompetence or purposeful inactivity. In addition to having the total crime and arrest data, one can easily also evaluate the arrest data on a by-crime basis. Low numbers of arrests in cases of sexual assault, for example, can reveal that police officers have not been properly trained to handle those types of investigations.

ACKNOWLEDGMENTS

The authors would like to thank their mothers and our wonderful professor, Elizabeth Sugar Boese.

REFERENCES

[1] [n. d.]. Urban Labs Crime Lab. ([n. d.]). https://urbanlabs.uchicago.edu/labs/crime

[2] Jim Dayton. 2015. Professor studies impact of Chicago gang violence. (Feb 2015). https://news.wisc.edu/professor-studies-impact-of-chicago-gang-violence/

[3] Andrew V Papachristos. 1970. 48 Years of Crime in Chicago: A Descriptive Analysis of Serious Crime Trends from 1965 to 2013. (Jan 1970). https://isps.yale.edu/research/publications/isps13-023

[4] Mark Peters. 2017. UChicago Crime Lab releases 2016 gun violence report. (Jan 2017). https://news.uchicago.edu/article/2017/01/19/uchicago-crime-lab-releases-2016-gun-violence-report