Analyze Crime Rates to Help Mitigate Crimes in Chicago

Jimit Bhalavat
Colorado State University
jimit@rams.colostate.edu

Logan Cuccia
Colorado State University
lcuccia@rams.colostate.edu

Kyle Cummings
Colorado State University
kc7@rams.colostate.edu

Mike Peyton
Colorado State University
mikep99@rams.colostate.edu

Introduction

Chicago was once a pillar of prosperity and trade where companies housed their headquarters, where business partners formed coalitions, and where impactful business decisions were made. Today, it is still a major capital for trade, and some areas are very prosperous. Other sectors are rampant with crime, murder, theft, break-ins, and kidnappings. Although there are countless factors contributing to the growing crime rate and law enforcement procedures, crimes are overwhelmingly complex and have to do with politics, state, and local government actions. The question arises how has crime changed over the years? Is it possible to predict where or when a crime will be committed? Which areas of the city have evolved over this time span?

The goal of this project is to tackle how we can use data from these crimes to potentially stop them from becoming so frequent. It will also provide data summarizations such as the location and frequency of the crimes, and the nature of the crimes which can ultimately help mitigate crimes and help law enforcement agencies pinpoint where these crimes are more likely to occur, and which districts should have a higher abundance of police officers. The dataset contains attributes such as date, coordinates of the incident, district, the nature of the crime, etc. We will attempt to use these attributes to generate summarizations that will assist in the fight against crime as well as assist in the city's economic future. This is an interesting Big Data problem because "With the fast development of positioning technology and prevalence of mobile devices, a large amount of modern urban data have been collected and such big data can provide new perspectives for understanding crime" (Wang, Hongjian).

Methodology to Solve the Problem

In order to solve this problem and help aid law enforcement agencies in Chicago, we will aim to analyze this dataset and run a variety of different processing techniques and models in order to interpret results and answer the questions mentioned above. This dataset is large and contains various attributes, so it is complex to analyze through normal methods. In order to process this data effectively and efficiently, we will use MapReduce and Hadoop to process the data and provide numerical summarizations such as the accuracy rates of prediction through

Random Forests, K-Nearest Neighbors (KNN), and K-Means Clustering. Other numerical summarizations will include the proportions of arrests made by the crimes committed, whether the trends in crimes have increased over the years, and if yes, what crimes are more likely to be committed.

In this project, we aim to analyze the relationship between different types of crimes and the location they happened. In our research, we also aim to analyze the statistics of a few specific crimes: theft, narcotics, and criminal damage, and whether these crimes have declined or increased over the years. We aim to begin with K-Nearest Neighbors and try to predict the relationship between the types of crimes and the location they occurred. We also plan to use Random Forest Regressor and K-Means Clustering to try and predict the accuracy of our model and aim to predict if the crimes have changed over the years. After selecting the best model, we plan to provide data visualizations so that it is easily readable and interpretable. We aim to use a few of the packages available in Python3 such as Pandas (Dataframe), Numpy (Math), Seaborn and Matplotlib (Data Visualization), and Sklearn (Algorithms). The framework we plan to use is Hadoop and Spark.

We begin our project by first processing the dataset. The data has a lot of null values which hinders our predictions and goals of the project. We remove the rows with the null values, since these rows and values are of little importance in the prediction. We then use MapReduce in Hadoop to provide summarizations of the dataset, such as what types of crimes had the highest amount of arrests, what wards and districts were prone to what crimes. We also provide some visualizations to observe the trends of crimes in Chicago over the years. Next, we move to KNN in order to find a correlation between crimes and the location of the crimes. We use KNN to provide a correlation matrix between the features in order to analyze what crime types can occur at a particular location. Then we move to K-Means Clustering where we experiment with different numbers of clusters and the clusters include different features. These clusters then run through the model and then try to produce results such as which Districts are prone to what sort of attacks. Finally, we try to answer the question of whether the crimes, at a particular location, can be predicted. We used Random Forests in order to predict a crime at a particular location along with the probability of other crimes at that particular location. We try to run Random Forests models multiple times with different features.

Dataset

The dataset comes from Chicago Data Portal and reflects the reported incidents of crime in the City of Chicago from 2001 to Present. This dataset is withdrawn from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting System. The entire dataset spans around 2GB. The entries of the file have attributes such as a unique identifier, block where the incident occurred, description of the location where the incident occurred, whether the incident was domestic-related, police district where the incident occurred, latitude and longitude of the incident, to name a few. To protect the privacy of victims, the exact addresses of the crimes are not shared, but the location data included will allow us to pinpoint the general location of the crime. We can use the description attributes to categorize what type of crime was committed, and also filter based on other attributes such as whether an arrest was made, or whether the case was domestic.

Discussion and Analysis

We ran the Chicago Crime Dataset through multiple machine learning algorithms in order to predict the crimes and the location with multiple features to achieve higher accuracy. Map reduce is a method to distribute data across a cluster and run some sort of algorithm on that data to get a result. This is useful particularly when we have to deal with large amounts of data. It can significantly decrease the running time to extract or manipulate data in a large dataset. This was developed by google and this framework is still very popular today. There are many applications and variations that can be used to solve more particular problems.

Because this was such a large dataset, getting metrics and information about the data required us to use something like map reduce. In this experiment, we used map reduce to sort through the data and combine the data into a much smaller csv file which could then be used to graph visualizations of what the dataset was showing. For example, we run mapreduce on the dataset to get a simplified dataset with just the wards and the percentage of time there was an arrest. We also got the total counts for all types in each ward. This allows us to compute the total number of incidents in each ward for this dataset. Finally, we also got the percentage of times there was and wasn't an arrest for each primary type. This was important in helping us understand the data distribution of the primary types. All this data we were able to extract from this dataset was useful for us with understanding what was actually happening within the large dataset. We were able to pull out important information which was hidden so that we could learn more about what this dataset encapsulated. Because we used map reduce for the majority of our data manipulation we saved a lot of time.

In order to understand the variation of crime among different areas in the city of Chicago, we want to be able to predict what type of crime will occur at a given location. We felt that using a simple classification algorithm would be a good starting point. According to Leif E. Peterson with the Center for Biostatistics at the Methodist Hospital Research Institute, “K-nearest-neighbor (kNN) classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data” (Peterson). Since kNN is known as a great starting point for classification, we decided to use this algorithm as our first attempt at making predictions.

We are predicting the type of crime by location, therefore, we knew that our target value would be “Primary Type”, which describes the type of crime committed such as “Theft” and “Assault”. There are 37 different primary types in this dataset. Since there are many different location identifiers included in this dataset we had to find a way to narrow down which ones to use. In order to choose which features would be used, we found the correlation between every feature, and then plotted it using a heatmap from the Seaborn library. We added a new column to the dataset where we refactored “Primary Type” from strings to integers in order to include primary type in our correlation plot. This correlation method proved to be unsuccessful because when two features with high correlation were used in our kNN model, it would try to train and predict for hours without ever finishing. We think this could be due to some corrupted values in the dataset that the kNN model could not read properly. Once we tried features with lower correlation such as “Latitude” and “Longitude” the model was able to train and predict in a reasonable timeframe.

Next, we had to split the dataset into a training and testing set. We decided to use a 70/30 split, where 70% of the dataset would be used for training the model and 30% of the dataset would be used for testing the model’s accuracy. In order to do this, we used the train_test_split function from the sklearn python library. Then, we used the KNeighborsClassifier, also from the sklearn python library, for our kNN classification. This is a very well known and commonly used python library with many different machine learning tools. For our kNN model, we decided to use a k value of 15 because that was the optimal value for this dataset. A k value greater than 15 did not give us any more accuracy and a k value less than 15 gave us a lower accuracy.

To better assist law enforcement within the Chicago community, K means clustering can help predict where some crimes are more likely to occur using police Wards, IUCR codes and districts to approximate a location. We can also predict when some of these crimes are more likely to occur using the time of day and IUCR codes for location. We can then plot the predictions to get a good view of the clusters. While the dataset at hand is very large, the features are largely uncorrelated with each other and makes the classification much more difficult to succeed. It also does not include over 70% of crime that takes place due to those crimes being unsolved and or having suspicious circumstances. We can still try to cluster the data to get districts where these crimes are more likely to occur. Using the elbow curve rule as with other K means clustering problems, the optimal value for n_clusters can be found. Running K means clustering with k = 1 through 10, the optimal value was found to be between 3 and 4. A single K-means classifier with 4 n_clusters is shown in figure 6.

K means clustering was also performed on the time of day of crimes as well as the IUCR codes since it is the best feature for location. To not make the IUCR codes so apparent in the classification, data normalization was performed before the prediction. Running this K means cluster with n_clusters = 3, the clusters saw minimal change. Using this sub data and the predictions previously calculated, we can get a decent prediction for which crimes occur at what location. Results from creating this diagram are shown in figure 5. K means clustering provides some insight to this problem, but has a weak capability of predicting exact times and locations of crimes about to be carried out. This is not only due to the data, but the nature of the problem. Although these features can correlate to the crimes committed in the province of Chicago, the nature of the problem and the absence of other crucial data mitigates the classifiers ability to perform. The groupings that resulted from the clustering show that more western parts of

Chicago are far more dangerous than the south, for example. Although there are some areas that have far lesser crime rates than the west side, Chicago is undeniably still prone to violent crime. These graphs clearly indicate the western districts of Chicago have extremely high crime rates compared to the national average. We can compare this data further in the Random Forest Trial.

To have a better response and understanding of criminal activity, it is critical that one should understand the patterns in crime. We analyze these patterns by taking the Chicago Crime Dataset. The major aim of this project is to predict which category of crime is most likely to take place in Chicago. Random Forest algorithm is used to solve regression and classification problems. We decided to use Random Forest over other machine learning algorithms because it is more efficient than Decision Trees and it can produce a reasonable prediction without hyper-parameter tuning. In every random forest tree, a subset of features is selected randomly at the node's splitting point. "Random forests classification employs an ensemble methodology to attain the outcome where the training data is fed to train various decision trees and features that will be selected randomly during the splitting of nodes" (Yuki, Jesia Quader).

We apply the same intuition behind Random Forests. We also wanted to experiment with Spark, so we introduced pyspark in order to implement Random Forest. Firstly, we split the Date column of the dataset into year, month, and day. Next, we combine a few of the Primary Type Crimes mentioned in the dataset to achieve higher accuracy, for example, we combine prostitution and sex offence as Criminal Sexual Assault. After that, we instantiate the model, but only with the Latitude and Longitude features. The random split of the dataset is 70% training data and 30% test data. After fitting the training dataset and evaluating the model with only two features, we achieved low accuracy rates, so we decided to train the model with more features. Next, we train the model with additional features such as Arrest, Beat, Ward, Community Area, etc. We repeat the process of fitting and evaluating the model, but this time achieving higher accuracy than before. We then predict the probability of a crime type at a particular location along with what District and Ward the crimes are likely to occur.

The challenges we faced during this project were during the preprocessing of the dataset. The original dataset had around 7 million rows, so in order to process that we had to rely on Hadoop and MapReduce to only extract columns we need and also delete rows with the "NaN" values since they are not helpful in predicting the results. The other challenges we faced were running the models for different algorithms. We had to come up with features in the dataset that would be helpful in predicting the location of crimes. Since this dataset is so dynamic and has a lot of moving parts, we had to come up with models that would best help us in predicting the crimes at a particular location.

Results

The amount of crime committed is directly related to the number of arrests. Statistically, higher the number of crimes, higher the number of arrests. Crime has evolved positively over the years in Chicago, which means crimes have decreased over the years. Using MapReduce, we came up with the number of arrests for a particular crime and plotting that graph, we can see that the number of crimes have decreased over the years. Narcotics is one of the highest number of crimes committed in Chicago, the number of arrests for Narcotics in 2002 were around 52,000 but the number of arrests in 2020 was around 10,000. Since the number of arrests decreased, we can safely conclude that the number of crimes committed decreased as well.

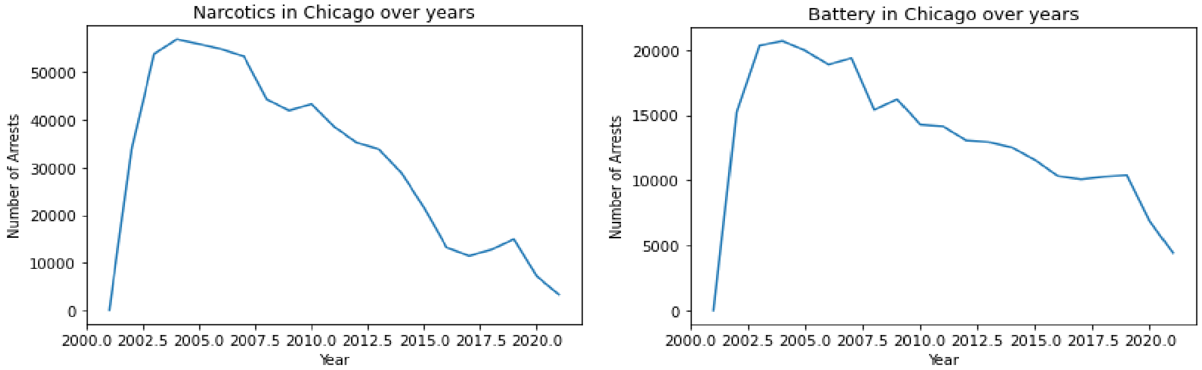

Figure 1: Narcotics and Battery Number of Arrest in Chicago over the years.

Map Reduce was very successful on this dataset. We were able to easily pull and combine data in order to make graphs to visually represent our data. This step was very useful in understanding our data further. First off, we ran map reduce to get the percent of time an arrest occured in each ward. Then we graphed the top 15 of those wards. From this, we were able to conclude that wards 37, 24, and 28 had the highest arrest rate when there was an incident. Ward 28 was the highest at just over 40% shown in figure 2. We then ran map reduce on a dataset to get the count of incidents for each ward. This showed us the count distribution across the top 15 wards. If you look at figure 3 you can see that ward 28 had the greatest number of incidents among other wards. We then broke down ward 24 incidents to see that the highest type of incident was Narcotics which can be seen in figure 4.

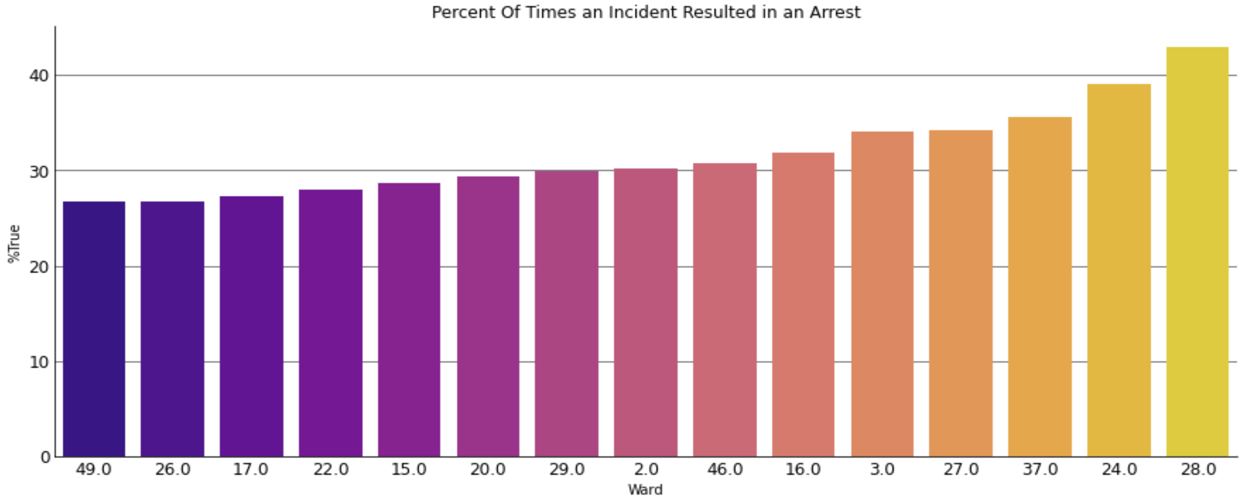

Figure 2: Top 15 wards with percent of times an incident resulted in an arrest.

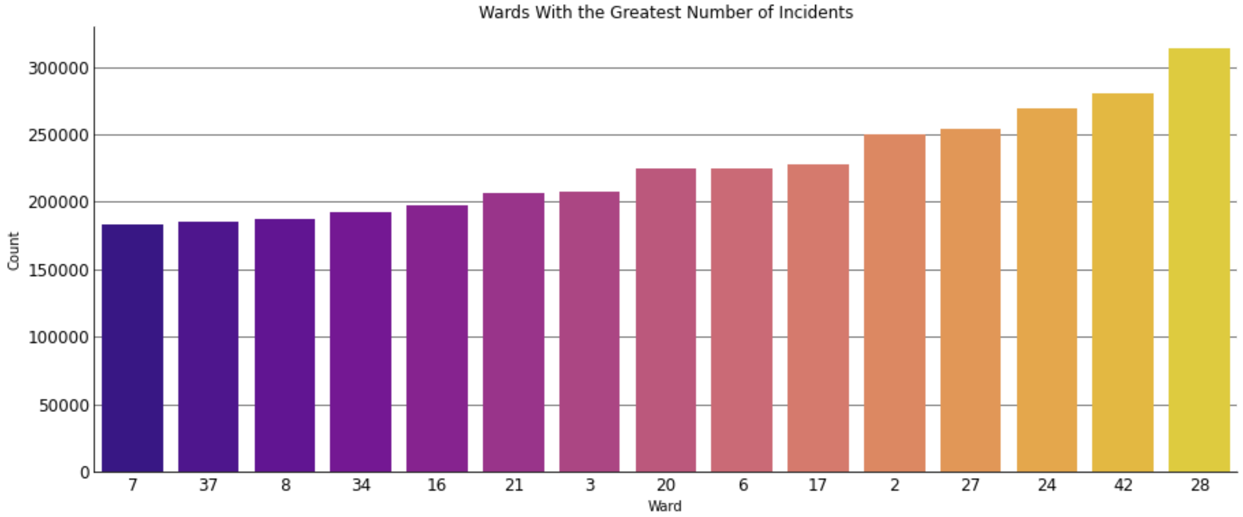

Figure 3: Top 15 wards with greatest number of incidents..

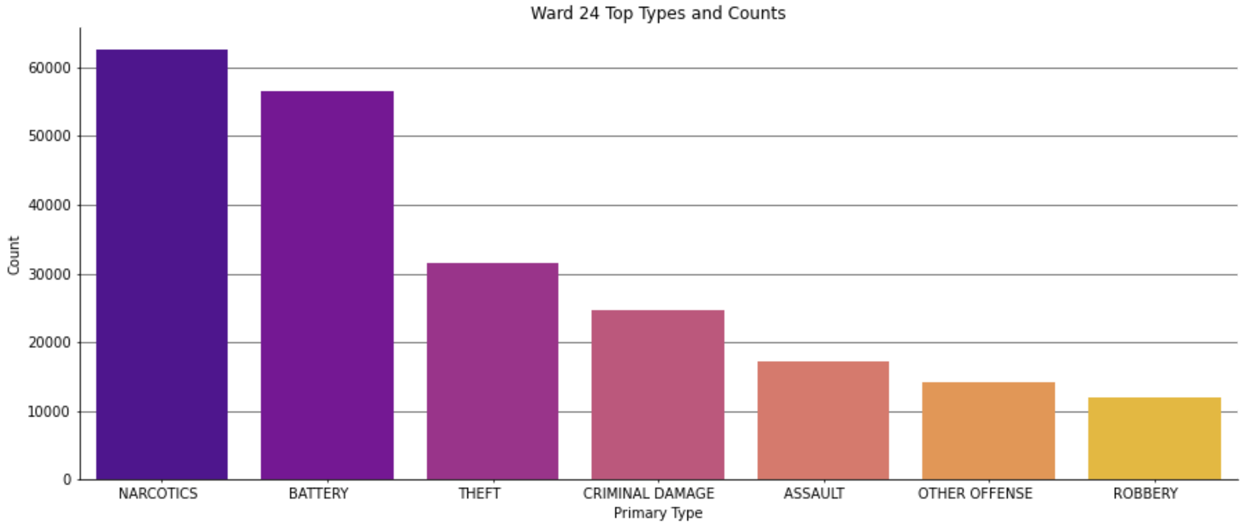

Figure 4: Top 7 ward 24 primary types and counts.

The results found from our K-Nearest Neighbor (kNN) model were not as accurate as we had hoped, but with such a big dataset, we did get plenty of correct predictions. The accuracy of our model with a k value of 15 was 28.5%. This was while using “Longitude” and “Latitude” as features to determine a location. The most common occurrences of correctly predicted crimes were “Theft” and “Battery” with 215,259 and 179,872 which was expected because these two crimes are the most common in the entire dataset. The next most common occurrence was “Narcotics” with 73,805.

Although the accuracies reached were low for the predictions, the clustering graphs and bubble diagram can give us a much better representation of the data. The graph distinctively shows a significant amount of crime in the western regions, as well as scattered all throughout Chicago. Southern regions have slightly higher rates of crime than the national average, however they are still significantly lower than the rates shown in the western districts. In the bubble graph, you can observe some of the tightly confined areas in the middle of the sphere, the most common crimes plaguing the city of Chicago are Homicide, Battery/Assault, and Theft. Western Chicago once again has a much higher concentration of these crimes in a smaller region. It is important to note as stated before, the overwhelming majority of crimes go unsolved or otherwise undocumented, and are not included in the dataset. Having a faster police response time and more units is what needs to be done to not only prevent or catch these criminals, but also for more reliable data that can be used further to help save citizens as well as law enforcement from dangerous criminals.

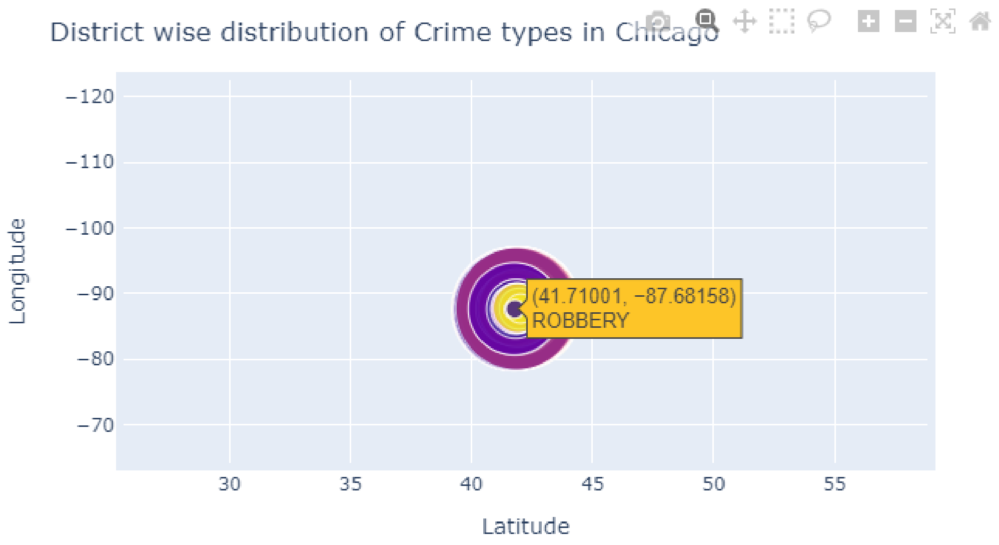

Figure 5: District Wise Distribution of Crime Types

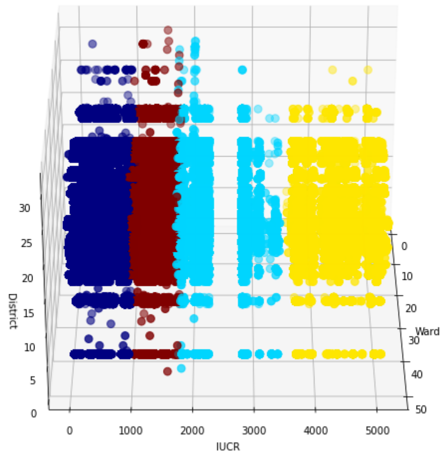

Figure 6: K Means Clustering using Ward, District, IUCR Codes

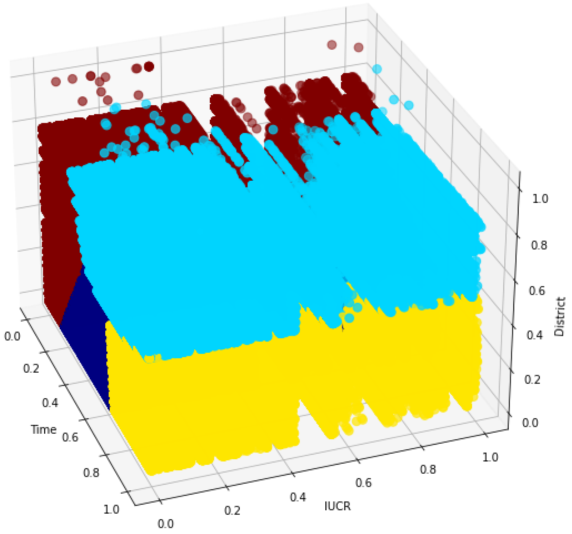

Figure 7: K Means Clustering using Time, IUCR, and District

Random Forests algorithms reduce overfitting and work well with both categorical and continuous data values. After fitting the model and predicting the results, this algorithm successfully predicted the probability of every crime type at a particular location along with the district and ward the crime is going to take place. For example, at location (41.852346, -87.665346), the probability of Theft is around 19% followed by Battery which has a probability of around 18%. For this particular location, the District and Ward predicted are 12 and 25 respectively. These results predict the probability of crime being committed at a particular location and which District and Ward are located at the particular location.

Conclusion

In conclusion, we ran the Chicago Crime Dataset through various models to predict the types of crime at a particular location along with different attributes such as whether an arrest was made or which districts and wards are more prone to what types of crimes. Machine Learning models are as good or as bad as the data you have. Correlation between features is important for predictions. In our case, we experienced low correlation features with our predicting variable. We experimented with different features in order to get better predictions such as using Latitude, Longitude, IUCR codes, Districts, and Wards to predict crime type based on location description and arrest. The results became better, however, not significant enough. The original dataset was highly imbalanced. Dropping/merging related crime types only helped the balance by smaller margins and not all crimes had a good correlation with latitude and longitude. Some future work can include additional data such as demographic and weather data and focusing on specific crime types can provide better predictions.

Project Contributions

Tasks Implemented	**Names**
Data Processing: The dataset had a lot of “NaN” values, so in order to predict the data, we had to clean the data using MapReduce and Hadoop. We also extracted columns that we thought were helpful for our project goals. We also created new datasets based on specific columns and data, and used those data to create graphs and answer the questions of the term project.	Jimit Bhalavat, Kyle Cummings
Data Visualization: Create different kinds of plots from the new datasets and try to answer questions defined in the term project. We used Matplotlib and Seaborn as data visualization libraries in Python.	Kyle Cummings
K-Nearest Neighbors (KNN): Implement the KNN model and try to predict what crimes have the highest chances with regards to Latitude and Longitude. Also, plot correlation's matrix to explore dependency between features.	Logan Cuccia
K-Means Clustering: Implement the K-Means Clustering model by clustering the data according to different features available in the dataset such as District, Ward, IUCR, etc.	Mike Peyton
Random Forest: Implement the Random Forests model using the different features available in the dataset. First, run the model with fewer features and then run the model with multiple features to ensure higher accuracy and accurate predictions of what proportion of crime types can occur at a particular location.	Jimit Bhalavat

Works Cited

Clustering Chicago robberies location with K-means algorithm. Arkadiusz Kondas | Software Architect and Data Scientist. (n.d.). Retrieved December 2, 2021, from https://arkadiuszkondas.com/clustering-chicago-robberies-locations-with-k-means-algorithm/

"Crimes - 2001 to Present - Dashboard: City of Chicago: Data Portal." Chicago Data Portal, https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

K-means. TowardsMachineLearning. (2021, December 1). Retrieved December 2, 2021, from https://towardsmachinelearning.org/k-means/

Peterson, Leif E. "K-Nearest Neighbor." *Scholarpedia*, 21 Feb. 2009, http://scholarpedia.org/article/K-nearest_neighbor.

Wang, Hongjian et al. "Crime Rate Inference with Big Data." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016. 635–644. Web.

Yuki, Jesia Quader, et al. "Predicting crime using time and location data." Proceedings of the 2019 7th International Conference on Computer and Communications Management. 2019.