Crime Prediction and Analysis using Temporal Crime Data

George Mason University, Volgenau School of Engineering, Fairfax, Virginia
Rishab Banskota, Nabeel Choudry

Abstract

Crimes are increasing over time. There are several types of crimes that take place and they are exponentially growing. For this project, we have picked Chicago and DC as our two major cities for analysis. Chicago itself has a total reported over 51 thousand crimes just for the year of 2018. These crimes have been a great motivation for us to work on a project like ours where we build models which predict the type of offense and location where it might be taking place given some other attributes of the crime. The objective is to understand the cause of occurrence of a crime and understanding the factors that cause them to help law enforcement officials narrow down while during the investigation so that they can better act on it and can ensure they have the adequate resources to make it happen beforehand.

In this project, we implemented several models that predicted geographical pieces of information about crimes while given some other attributes. We also had models that predicted the type of offense that might take place given some attributes of a crime. We are using clustering and classification algorithms to make our predictions and are using several data mining techniques to make it happen.

Literature Review

Rasoul Kiani and Siamak Mahdavi [1] had performed a similar work in the past. They were successfully able to cluster crime data based on the crime to group different types of crimes. This project took place in India. They used KMeans to cluster crime and perform analysis on it. There were able to mark spots on a hot spot to see which places had the highest possibility of a crime occurrence so that it can be monitored. Additionally, they used PHP script to load add and retrieve data. They used Google API to color code different crimes on the map and depending on the users' location, they show an initial view of the crimes around them. Their dataset was basically the crimes that took place in New Delhi, India.

Priyanka Gera and Dr.Rajan Vohra have also conducted a similar experiment where they look at 59 years worth of past data to predict crime like theft, robbery etc. They used

linear regression to create a system that produced a formula and squared correlation(r^2). The coefficient was useful because it outputs the proportion of variance of one variable from another variable. It was a measure that helped determine how one variable can turn out to be from a certain model/graph.[6]
We performed a hybrid of the two above where we cluster first and then classify next within the clusters. We first narrow down our pool of crimes where we look at and then based on the other crimes similar to it, we try and predict what the type of offense would be and the neighborhood cluster where it might be taking place.

Methods and Techniques

Preprocessing

As part of pre-processing, we started off with our DC crime dataset as it had fewer records relative to the Chicago one. Furthermore, they had similar features so it did not make that big of a difference. The total number of recorded crimes in our dataset was categorized by years. Our crime data ranged from 2012-2018. As the data was not clean, we did a lot of pre-processing to it. Some of these include removing the misleading/null values and feature selection. Some examples of nominal columns which were disregarded are CCN number, case number etc. These features did not help us identify the crime itself. We also took into account continuous features and how to handle them. We had to do a lot of research to understand the relationships between out features(eg.precincts and districts)to see how it could help us predict a class label.
We split the data into two parts. The first contained crime records from years 2012 to 2017 and was used as train data. The other part contained records from the year 2018 and was used as test data for our models. Additionally, for pre-processing since our data had strings we had to encode and scale a lot of our algorithms. For this, we used the LabelEncoder for encoding and MinMaxScaler for scaling our data from(-1,1) so that it was highly optimized and pre-processed to be fed to other algorithms.

Random Forest Classification with Pandas DataFrame

We used features such as shift, offense, ward and district for this model. We used the features above to predict the neighborhood cluster. DC is divided into 29 clusters. There are 39 neighborhood clusters throughout the city, each made up of three to five neighborhoods. These clusters are being used by the D.C. government for budgeting, planning, service delivery, and analysis purposes.[2] We predicted the cluster value for a given crime given the offense, shift, ward, and district. We used Pandas DataFrame for manipulation of our dataset, to filter and pre-process it and to extract three additional columns after parsing one of our features.

Classification and Clustering

For this model, we performed classification and clustering on our data. We first clustered our data into 10 clusters. We initially started with 4 clusters and went up to 10 and monitored the changes to come up with the best optimal number of clusters. We used the K-Means clustering for our clustering algorithm. We did this for both our train and test data since our final plan was to classify a crime by going into its cluster and then predicting the value of it based on the cluster's interpretation. We noticed how this narrowed down our train data by a lot which made the overall execution faster and more accurate.

Data

Our source of data was Chicago Data Portal[4]. Our dataset had 6752229 records. This is because of its crime data from Chicago from 2001-present. Some columns of our csv datasetareID,CaseNumber,Date,Block,IUCR,PrimaryType,Description,Location,Arrest,Domestic,District,CommunityArea,FBICode,XCoordinate,YCoordinate,Year,UpdatedOn,Latitude,Longitude,Location, etc. There were columns like FBI code, ID, Case Number, etc that were irrelevant to us, therefore, we didn't use them during our prediction. We were mainly interested in knowing more about the location/area of the crime, therefore, some of the columns that are relevant to us are the spatial ones with the location information and the ones that describe the nature of the crime.

The relatively smaller dataset we used was the DC crime dataset which had a total of 246879 records[7]. We were able to pre-process and test our model with this data before proceeding to the larger one. We used our train data as data from the past years from 2001-2016 and we used 2018 data to be the train data for our model. We had some interesting observations while during this approach. We noticed how some of the offenses did not exist in our train that existed in the test data. Also since we are dealing with crimes here, their location varied as time progressed (not a whole lot but noticeable). We did not have enough features to classify with a good prediction. Our main goal was to not use any geographical data to predict an area of the crime, but since the location is a continuous feature it was very difficult to get a good accuracy on the prediction.

As a part of pre-processing we append two columns as the month of the crime and year of the crime. These two columns were parsed from the reported date of the crime column where we were given a timestamp. We extracted the month and date and appended them as a column when the crime took place.

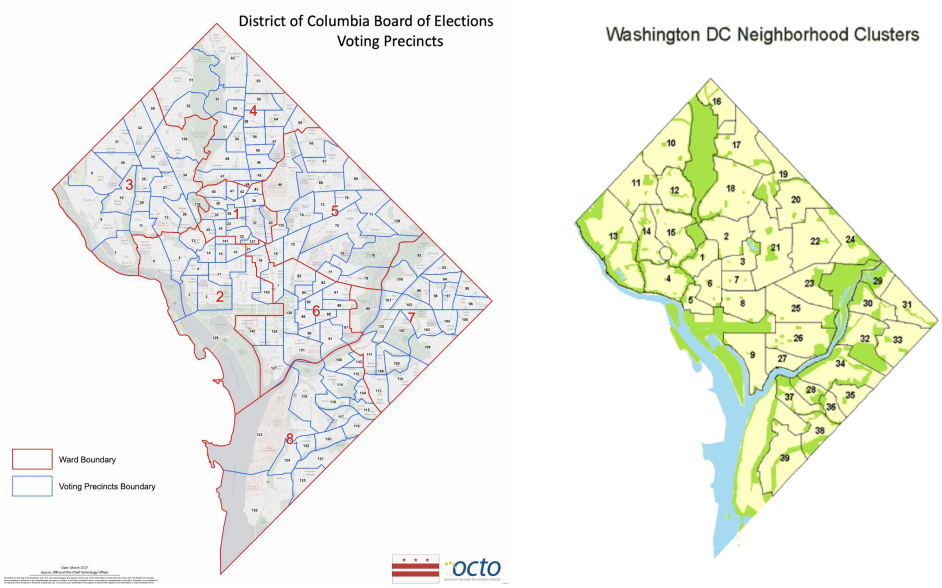


Fig: Division of DC into wards, voting precincts, and neighborhood clusters.

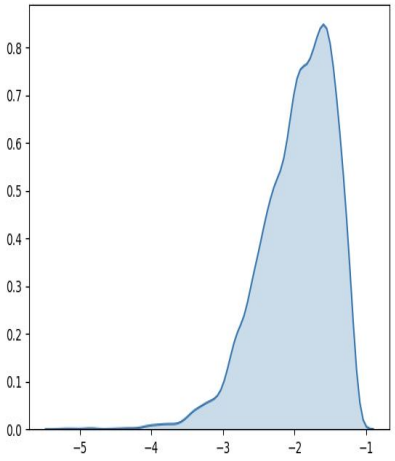
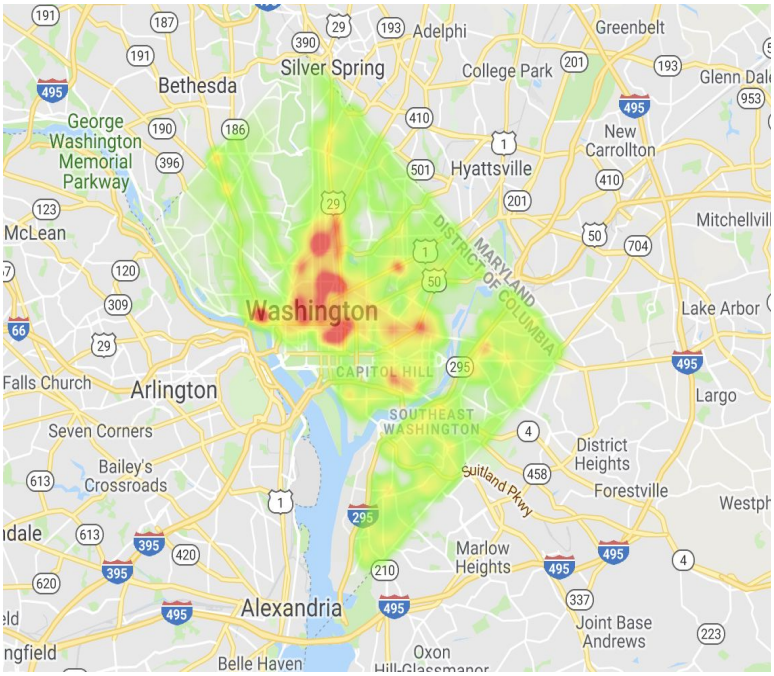


Fig: KDE Score Samples of the prepared train data.



Heatmap of DC dataset generated using Google API

Evaluation Metrics

We used a variety of evaluation metrics for our different models. For our models with RandomForest classifier, we used the built-in mean accuracy score. With some of the classification algorithms, we used their built-in accuracy score. However, for some models, we used the f1 score from the sklearn metrics library which computes the weighted average of precision and recall.

Experimental Results

Interesting observations

Even though ‘method of crime' sounded like a deterministic feature, it did not serve as a decisive attribute as we did not have a variety of methods. We also noticed that for our train and test data, we had a lot more records for one type of offense than the other one. This was a noise while classifying and clustering data.

Even though our accuracy was pretty good, after research we noticed that K was bigger and neighborhood cluster was inside a precinct. In other words, one precinct had many neighborhood clusters in DC which resulted in our accuracy to be very high. That was not very interesting therefore we did some further analysis and steered towards a different kind of model.

Random Forest Classification with Pandas DataFrame

We were able to predict the cluster where the crime would belong to using the columns above using Random Forest Classifier with an accuracy of 91% using the mean accuracy score metric from Random Forest library. This was our model as we progressed to find the location of a crime or a common area given crime data from past crime data. In other words, in this experiment, we were successfully able to predict the cluster that the crime belongs to given attributes of the crime.

We felt the random forest classification would be best for our model because it builds a training model using many decision trees that are build by dividing the data into many pieces.

Classification and Clustering

This model was our heaviest model and we spent a lot of time here. For this model, we performed classification and clustering.

- **Offense prediction models**

  After analyzing the data using the heatmap and graphs on the crime statistics (see data section), we observed two things:

  1. There were certain spots on the maps that showed a higher concentration of crime compared to others.

2. There is a high concentration of certain crimes during certain months of the year and time of the day.

With those observations, we concluded that there was some correlation between location, time and crime. So, we decided to build offense (type of crime) prediction, model.

For this model, we used location and time attributes to predict the type of crime that might occur for those given attributes. The time attributes were; month, date, shift (AM, PM) and the location attribute were; block number, precinct, and census tract.

Since the accuracy score for random forest classifier was low, we decided to use KMeans to cluster the data into smaller similar groups and use the random forest classifier inside those smaller clusters to improve the accuracy. We used the elbow method technique to get the optimal number of clusters for the KMeans. Then we used the sickit's KMeans library to train our clustering model and used RandomForestClassifier inside those clusters to predict the type of crime. This model also gave us an accuracy score of 55%. Although this model used random forest inside the cluster, there was no difference in the accuracy score.

We tried a variety of classification technique for our prediction. We also tried changing our training and testing data. Some of the things we tried for training and testing data construction was test-train split, using past years data as train and future ones as the test, etc. Regardless of all these techniques, we did not have even information about all the types of crimes, therefore, our accuracy wasn't that high. We did a research on how we could make it better and adding weather data to it made a little bit of a difference but not a lot.

Results

Model Description	**Algorithms used**	**Model Accuracy**
Neighborhood cluster prediction	Random Forest	91%
Type of offense prediction	KNN and KMeans	61%
Type of offense prediction	KMeans and Random Forest	58%
Type of offense prediction	KMeans and MLP	57%
Type of offense prediction	KMeans and Linear Regression	56%

Conclusion

The ultimate goal of this project was to help prevent criminal activities from taking place and making the world a better place. We wanted to predict helpful characteristics about crimes so that law enforcement officials can better act on a crime when something is determined to happen. We also wanted to identify the key catalysts of crime and identify them so that steps can be taken towards its prevention. Although we were not able to predict the concrete area of crime other then the neighborhood cluster, we learned a lot about crime data and relationships between them. We also learned how temporal characteristics can be taken into account and how they serve as a condition in the classifier. Also, we learned that a bigger dataset does not necessarily means anything. A good dataset is the one that tells you more about the different types of label you are predicting. You can have a big data set which tells you minimal about the prediction label. Also, we learned that given a dataset you can perform analysis to find out several things that it can imply and correlations amongst the features can be very useful in prediction.

We learned how data mining is such a broad field and applications of it is just immensely. We learned how to apply different clustering and classification techniques to a dataset to provide meaning to it. Even though raw data doesn't mean anything after deep analysis it can describe things. With the increasing amount of technological gadgets and growing data all around us, machine learning and data mining serve a great purpose and have a really high scope in putting meaning to it and predicting foreseeable things and making human lives better and easier.

Directions for Future Work

Crime Indicators: In the future, we want to create a model that predicts the indicators of the crime in a certain area. Looking at the features of different crimes, we want to be able to predict some of the reasons why these crimes might be taking place. An example can be the weather. We want to find the strongest correlators of the offense that is taking place in a certain location so that preparations can be done towards improving them. This would help reduce crimes in the society so that it can be a peaceful place to be at.

References

[1] Sharma, Y. (2017). Crime Prediction using K-means Algorithm. [online] Pdfs.semanticscholar.org. Available at: [Accessed 25 Nov. 2018]. https://pdfs.semanticscholar.org/3643/74119cd633ac6396f81959700912acdf30ee.pdf

[2] Neighborhoodinfodc.org. (2010). Neighborhood Profiles. [online] Available at: https://www.neighborhoodinfodc.org/nclusters/nclusters.html [Accessed 25 Nov. 2018].

[3] " Report Covering The Week Of 26-Nov-18 Through 02-Dec-18". 2018. Home.Chicagopolice.Org. https://home.chicagopolice.org/wp-content/uploads/2018/12/1_PDFsam_CompStat-Public-2018-Week-48.pdf.

[4] "Crimes-2001 To Present". 2018. Data.Cityofchicago.Org. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

[5] https://pdfs.semanticscholar.org/3643/74119cd633ac6396f81959700912acdf30ee.pdf

[6] P. Gera, and R. Vohra, ―Predicting Future Trends in City Crime Using Linear Regression, ‖ IJCSMS (International Journal of Computer Science & Management Studies) Vol. 14, Issue 07 Published Month: July 2014.

[7]2018. Opendata.Dc.Gov. http://opendata.dc.gov/datasets/bda20763840448b58f8383bae800a843_26?selectedAttribute=XBLOCK.