SLIDE ONE

Hello Everyone!

I hope you are all doing well today, my name is Matthew McFall and for my first practicum presentation, I will be presenting on my analysis of the geospatial data given on Kaggle.com for criminal offenses in the City and County of Denver for the previous five calendar years plus the current year to date. The main idea behind this project was to see if you could effectively use supervised and unsupervised technics to classify criminal incidents effectively thus predicting the likelihood of which crime will occur when and where.

SLIDE

So, for the overview of this presentation

First, I will go over the dataset and dimensions

Next, I will discuss how the data was cleaned

From there, I use Hotspot Analysis for an exploratory evaluation

Then, I will go over the KNN Models that were created

Followed by Issues Using K means for geospatial data

Then, I will go into the DBSCAN Model

finally, I will go into the conclusion with what was leaned and what can be done for future analysis

SLIDE

First up we have the Dataset Overview,

The total length of the dataset is about 493 thousand rows with 19 columns, after the first initial clean the rows dropped to 483 thousand and I only used about 3-4 columns of the data given. All the missing values that composed the piece were predominantly sexual assault longitude and latitude coordinates, all except for 1 sexual assault case which I just dropped. It was to my understanding this was done to preserve the victim’s identity due to the nature of the crime. The data again is over the past 5 years plus the current one of the city and county of Denver Colorado. This is shown in The image in the top right shows all the data points displayed on a map which essentially highlights the Denver city and country area. The bar chart below shows how the crime categories broke down and compared to each other in terms of amounts committed, as is obviously inferred traffic incidents ranked as the most likely occurrence while murder ranks the least.

SLIDE

Next is Cleaning the Data

After the sexual assault category was dropped, I had noticed some strange points initially when I graphed the data in the image given in the right of the slide, encased with the blue box. After reading the documentation given on the dataset, I still had no idea why there where points out in the middle of nowhere on the map so I decided to get rid of them as is shown in the image in the bottom right of the screen. Lastly, I made a lot of subsets with the data when I go into my modeling and I will be indicated that when I discuss the models.

SLIDE

Hotspot Analysis

For an exploratory analysis of the data points given I decided to approach it with a hotspot analysis first. I did this through excels 3d mapping tool that is available in the 2016 version. Initially, I noted as is evident in the image in the top right when all the crime incidents are toggled on the heatmap shows crime permeating from the center of the downtown area and the major highways that go through Denver. Thus, I decided to see what happens when I toggle off the traffic incidences off which resulted in a similar display as is shown in the bottom right of the screen. This is unfortunately due to how the points are logged which is based on the police car position. Next, I we go into modeling.

SLIDE (Supervised leaning technique)

KNN Model

My first models that I attempted for my supervised learning approach with the data were KNN based. The data was split into an 80/20 train test sets before the models were created. the first model that was created with the data was with all the crime categories, or 15 of them, which preformed very poorly at a 35% success rate on classifying crime. Therefore, I decided to eliminate some categories that likely had a high degree of sporadic occurrence likely traffic incidences, larceny, the catch-all category for miscellaneous crimes, etc.. unfortunately, this resulted the next few models performing worse with about a 30% average. Therefore, I decided to start pairing the crimes respectively off similar occurrences to see if I could effectively tell the difference. The best model I was able to find was telling the difference between robbery and white-collar crime which performed at a 60% rate initially then performed at a 65% after optimization. Unfortunately, when I did a color plot of the classification groups, the algorithm simply split the data straight down the middle, asserting that you are more likely to be victimized by white-collar crime on the right of the city and robbery on the left which doesn’t yield much information. Next I’m going to show you some of the charts of KNN model with just white-collar and robbery.

SLIDE

First is the Andrews curve of the dataset, which illustrates that there are certain areas that have a mixture of both white-collar crime and robbery, singular occurrences of both, but overall it doesn’t look promising for cluster analysis.

SLIDE

The next image shows a coordinate plot of the x y coordinates with the precinct id and district id. Overall the image tells a similar tale to that of the Andrews curve, which is that there is a mixture of grouping within the geographical spread resulting evident in the charts portrayal of district id to precinct id which shows some of the lines is having a mix of blue and black while others of singular colors.

SLIDE

This next image shows a chart of how I optimized the algorithm an additional 5% through calculating the best number of neighbors to use. In this case there where two crime categories but 3 clusters optimized the algorithm the best.

SLIDE

Lastly I created a pairwise distribution of the data which shows that there are two geographic locations or precincts, 3 and 6, which have more of a singular occurrence of a specific crime 3 being white collar and 6 being robbery, while the others seem to have to have relatively the same amount of both crimes.

SLIDE

Issues Using K means unsupervised learning approach

k-means is not an ideal algorithm for latitude-longitude spatial data because it minimizes variance, not geodetic distance (Boeing, 2016). This results in the algorithm working but performing poorly, and not having much room for improvement. Therefore, this left me with one other option which is the DBSCAN model for an unsupervised learning technique.

SLIDE

DBSCAN Model (Unsupervised leaning technique)

For my unsupervised learning technique I used the DBSCAN model which I initially wanted to try and see if the algorithm had similar results with the entire dataset as the first model did however, due to performance issues on my laptop I was only able to do small datasets with this algorithm. So, I just stuck with the robbery and white-collar crime dataset. The algorithm took in the data and made three clusters with it. The image in the top right shows the center points of the clusters surrounded by some of the original data. Interestingly that puts a cluster in the top left of the Denver area or the downtown area, the cluster in the top right is around the airport area, and the third cluster in the bottom right is around where Hampden South is.

SLIDE

Conclusion

Lesions leaned

Cluster analysis on crime seems to be a bust for predicting what type of crime will occur where in the Denver area due to the sporadic nature of the criminal occurrences and it being an overall safe community comparatively.

Moving forward

For Potential models you could try to create an Artificial neural network due to the timeframe given on the criminal incidences which might create a more accurate model. For crime analysis in general the geographic constraints of only seeing a small part of an area without being about to connect to the immediate areas surrounding give a pigeon holed view into incidences that occur, so creating a database or data set that has an entire geographical region would definitely aid in understanding the criminal incidences in a area.

SLIDE

THANK YOU all for viewing my video I hope you all enjoyed it.

SLIDE

References

Boeing, G. (Oct 25, 2016) DBSCAN clustering to reduce spatial data set size. Retrieved from https://github.com/gboeing/2014-summer-travels/blob/master/clustering-scikitlearn.ipynb

Mooney, P. (2019, June 29) Denver Crime Data. Retrieved from https://www.kaggle.com/paultimothymooney/denver-crime-data

Python Tutorials (2015) k nearest neighbors. Retrieved from https://pythonspot.com/k-nearest-neighbors[/](https://pythonspot.com/k-nearest-neighbors/)

Skalski, P. (2016) K Nearest Neighbor. Retrieved fromhttps://www.kaggle.com/skalskip/iris-data-visualization-and-knn-classification

Video available on: https://youtu.be/CfLQm5GkP-Y