COVID AND CRIMES : PREDICTING THE OUTCOMES

Vishal Pathak, M.S

Nima Zahadat, Ph.D.

The George Washington University

**ABSTRACT**

In Last 3 years the meaning of living a normal life has changed. The world which was used to hug and shake hands, moved to social distancing and elbow fists. To say, *“The Pandemic has changed the world”*, would be an understatement. It has reduced the impersonal meeting and increased the virtual existence of humans on internet. Pandemic has imposed the restrictions on everyday lives, thus effecting the people behavior. Pandemic has also affected the crime rates. In this study we intend to find the effects of covid on frequency of crime rates (includes violence, theft, murder, home abuse) in the United Kingdom (UK). Particularly, we will study the shift of crimes from one form to another across different states before and after covid. Secondly, we also want to figure out what are the factors which decide the outcome of the crime offender. We expect to use various data science visualizations techniques from analyzing the patterns of form & frequency change of crimes. For figuring out factors responsible for the outcomes of felony we intend to use various statistical Machine Learning, Statistical and Probabilistic techniques.

**Keyword**: Machine Learning , Statistics , Covid , Crimes

**INTRODUCTION**

The pandemic that started in China in December 2019 and in the United Kingdom (UK) in February 2020 (sources) has caused the biggest impact on the economy in the history of the UK. It has also transformed the mode of crimes, criminal penalties, and outcomes. This paper explores the crime pattern in the UK pre and post covid. Since people were staying at home, this led to reduction in street violence, but home violence has increased drastically. The divorce rates were high as people were getting caught for infidelity. In the starting few months of the pandemic all kinds of crime went low but as time progressed, they increased as well. Unemployment rates were all time high during the starting months of the pandemic. We expect our study to give us an analysis about monthly and overall changes during the pandemic. We are going to use all types of Machine Learning techniques to do so. We will try to find out which type of crimes has increased across each state in the UK, and which has reduced. We expect to analyze the underlying reasons for these changes if there are any.

Secondly, we intend to find out the factors responsible for the person’s trial. If a person has been charged with a crime, we need to find out whether he will be charged as guilty or not. This involves heavy Data Science Analysis, research and lots of practical knowledge as we want to deep dive the crime analysis and use different classification techniques with ensembles of multiple models for getting better results. We also want to involve multiple types of cross validation techniques including some Bayesian optimization techniques.

The problem with grid and random search is that they do not use previous results, instead it picks next results which makes it less intuitive. In Bayesian techniques, which is also known as Sequential Model based optimization, implements an idea by building a probability model of the objective function that maps input values to the probability of loss. The proposed technique will work and give us better results because it is more intuitive, and it considers last output for the current input.

At the technical level we will use Python programming language due to its support for different types of packages pertaining to Machine Learning. The acceptance of Machine Learning and artificial intelligence for the usage in public administration has already been presented [Cerrillo I Martínez, 2019]. Also, a decent amount of research work has been done related to crime prediction.[Batarseh and Yang, 2017] Our intention is to find out the hotspot areas of the crime and connecting the dots pertaining to covid.The structure of this article starts from the introductions followed by the literature reviews then discussions about the analysis, methodology, statistical techniques used for analysis and at last prediction results and conclusion.

**LITERATURE REVIEW**

Crime Analysis and prediction has already been a very import research topic over the years. There is an ample of original data present in the market and many research papers have been submitted. The good part about crime data is that the data is ever increasing and always available. With the abundance of data, there are many types of data available over the internet. Crime Analysis is a huge problem which has been partially solved by many researchers. One of the major one is [Bander and Vijayalakshmi, 2019]. They had used various statistical techniques such as ANOVA, Chi Square etc.

For classification of the problems, they have used heavy loaded Gradient Boosted Models which gives decent result but consumes lots of memory and time. With the advent of cloud-based services and deep learning, these problems have been solved up to some extent but using high core machine and high-level models is not the solution to every problem.

There are still needs of light weight models and tools which can analyze data in lesser time. [David and Suruliandi, 2017] have an interesting approach of predicting crime and clustering crime types in the similar fashion. The research also considers geo location-based crimes, as a feature for the prediction of algorithm. This completely makes sense because the type of crime differs as per the country, in fact it changes as per the city, the various studies discussed by [David and Suriliandi] also take internet data into consideration for predicting crime. They had used various NLP based statistical techniques for predicting crime patterns.

Another research paper [Sathyadevan Gangadhran] deals with the classification and pattern recognition of crimes problem. They used a very different approach as compared to other researchers which is, they have naïve bayes, apriori for the prediction and clustering algorithm.Although, they have also used decision trees for classification but their approach towards solving the problem was different. Instead of focusing on causes of crime occurrence like criminal background of the offender, their research was much more biased towards features of the day. We have taken this intuition in our research as well. Their way of statistical analysis is also different in the sequencing as Data collection – classification – pattern identification – prediction – visualization, mostly visualization comes around pattern identification.

Another research paper [Shyam Varan Nath, 2006], the research intended to take crime analysis problem as a geo spatial problem and tried to solve it using patterns creation with geo spatial plots w.r.t location and time. His observation about crime was that location is feature for explaining crime, makes his research quite efficient. From the perspective of modeling, he had used KNN like many other researchers who had researched in this field.

**RESEARCH METHODLOGY**

Initial Data Preprocessing was done using Python3 (packages included CSV), CSV manipulation was critical because data must be clubbed together as it was present on monthly basis. The Data was completely unstructured in every column. For Ease of use we ran 1st exploratory analysis on pre Covid data and then we did same process with post Covid data. With this we were able to tackle high amount of data. Data Cleaning required a lot of efforts because it was heavily skewed and it required removal of rows with missing values, transforming text categorical data to ordinal and cardinal data types. For modelling and analysis purpose we have used various packages like numpy, pandas, matplotlib, plotly, shap etc. For Visualizations we have used multiple visualization techniques such as Treemap, geolocating crimes, bar charts, histograms etc.For outcome prediction purpose we have used light Gradient boosted trees followed by my Bayesian cross-validation technique.Since it is a tree-based model which involved Gradient Boosted trees, we have used full data set for training and testing purposes. Everything was done using Python and Jupyter Notebook.The associated website and presentation were developed with Hypertext Markup Language (HTML) hosted on Github pages

**DATA INFORMATION**

The data was taken from the UK police crime data. There were 3 types of data viz Streets, Outcomes, Stop & Search. The data is available monthly for each type from 2018 to present. For the sake of Analysis, we have used streets data for 2018 to present (13 Million Records). We have divided it into 2 parts 2018 to February 2020 and February 2020 to present. This partition made the data easy to manipulate and process. The major features are Date, Reported-by, Falls-within, Coordinates, Crime Type and Last outcome category.

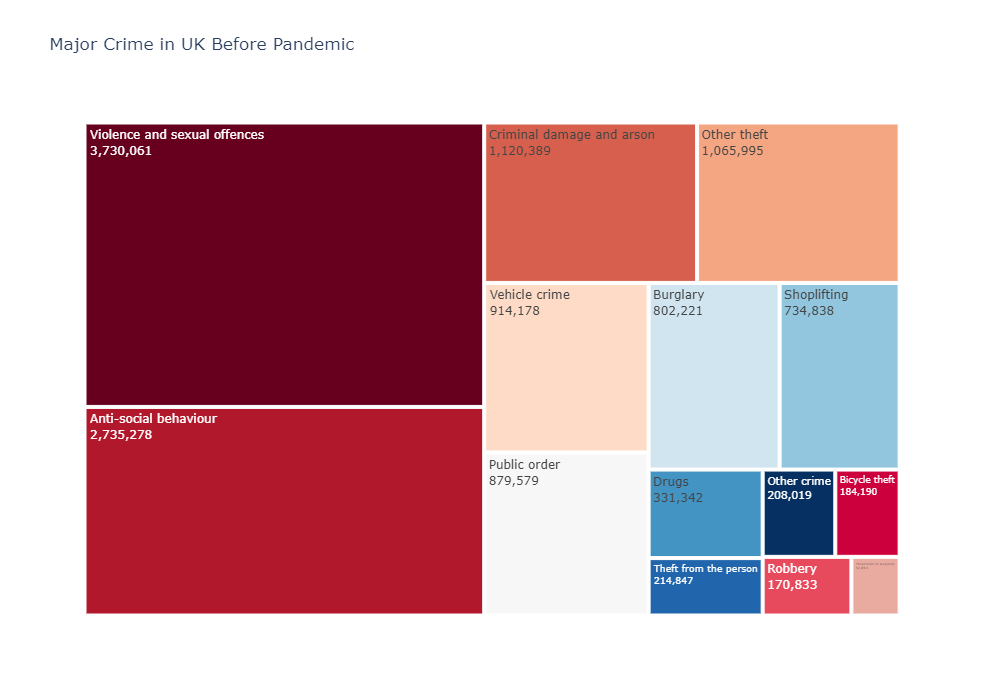
For the modeling purpose we have used outcomes data set which is around 13 million as well. This data is also available month wise, which on combining becomes 13 million. The Major features of this dataset were Date, Falls-within, Reported-by, LSOA code, Latitude, Longitude, location, Outcome-type (Label). There index column was a random JSON, which was no use for us, so we tweaked it without normal index. Data dictionary has been included in this, can be found in the Appendix.

**DATA CLEANING AND ANALYSIS**

When the data was loaded after merging, first problem we faced was no data was showing its type, after research we realized every dataset is being merged with their respective column names thus disturbing every column datatype. Then we removed the ‘crime\_id’ column because it was a random ‘JSON\_id’, as data was decent in count, we removed rows with missing values for both the dataset (outcomes & streets). Then we indexed the data with resetting the index. The Dimensions for pre-Covid streets data is (13185655, 12), which is around 1.5 years of data pre Covid. Since the columns were less, we had to create more data from existing columns, to extract columns like month we extracted the same from date column, we had also visualized the data using geo-plots to see which area has most crime involved and observe overall significance. We have also checked the outliers in the data, since we have decent size of the data, we had the power to let go few rows since there were not so many outliers. For analysis purpose we mostly relied on the plots like Treemap, frequency plot, bar plots. Since we did the analysis on street data, we knew that there were not too many things wrong with the outcome data, because many columns were common and the data belonged to the same set of crimes mostly, so for cleaning purpose we did the same things as we had data for 2 years and the count was too much, so the outliers and missing values were treated by removing it. We created many columns to have better predictions having size [13185655, 12] Weekday, weekend, month, caught by same police, year block month, day of week, latitude, longitude, there were few outcome types like investigation complete; no suspect identified, unable to prosecute suspect, suspect charged, local resolution, offender given a caution. So, we converted it into charged or not charged for ease. The total count of the outcome data before cleaning was (13011476, 10). We also removed few variables like latitude, longitude, LSOA code and LSOA name. While converting variables text to categorical we made sure to maintain the ordinality and cardinality of data to get proper results and to also avoid overfitting.

**KEY FINDINGS**

The first part of the research led us to deep dive in analyzing the pattern of behavior of crimes in UK. This is one a major part of the research because these key finding can help in reducing the crime rates post covid era. So, the first thing we wanted to compare was the crime by type.



While & After Pandemic

Chart, treemap chart

Description automatically generated

As we can see the counts of each crime type has dropped drastically and the reasons were obvious, people were at home they were not going out due to fear. Before covid public order crimes were high but since people were not coming out and scared to be around people . Also, the major chunks of vehicle crime also dropped drastically , and this impact is only due to the locked down imposed across the UK. Shoplifting we reduced by almost 70 percent because shops malls ,stores were closed.

Now We are going to see top 5 crimes before and After Pandemic.

Chart, bar chart

Description automatically generated

While and After Pandemic

Chart, bar chart

Description automatically generated

As we can see violence and sexual offences tops the chart , because this data might include home violence as well, there are possibility of increased home violence. Anti-social behavior is also very high as like violence and sexual offences. The thing which needs to be observed is the drop of violence as compared to anti-social behavior. Before Pandemic there was much difference between violence and anti-social behavior, but after pandemic these things almost get similar. One thing important we need to see is public order during pandemic is no more a major problem , it doesn’t even exist in major chunks. vehicle crime is all time low and its not even a top 5 crime.

Now let’s see the count from monthly perspective.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

As we can see , February saw a major spike, this is a peculiar observation for data , apart from that most of months it went low. The strange thing to observed in this is crime rises during the first half of the year and go low by the second half. The reason could be a weather situation as crime is low in those months when there are festivals of low temperature.

Now we are going to see top county wise crime count.

Chart

Description automatically generated with medium confidence

Chart, bar chart

Description automatically generated

There is always a possibility of Police department efficiency , how well they manage their area.

The geolocation also affects the crime rates , it always depends on the how much population that county holds . we took the population counts for top counts and which are as follows.

[Source : Wikipedia]

Metropolitan : 7.2 million

West Midlands : 5.9 million

West Yorkshire: 2.235 million

Great Manchester: 2.822 million

Thames Valley : 2.34 million

Kent : 1.855 million

Lancasaster constabulary: 1.5 million

Essex : 1.8 million

Northumberland : 1.5 million

Hampshire constabulary : 1.9 million

Merseyside Police : 1.5

As we can see that the population count is the answer for the maximum crimes in UK county wise. Lanchaster Constabulary has joined the top 5 count group which normally was not there . the point needs to observe is that the counts before and after pandemic hits is same. This could be due to area was running normally and nothing changed while pandemic.

While Coming to Prediction Part:

For prediction problem , we started with basic model and then move to more advanced models. For cleaning the data, we performed various statistical techniques such as skewness, kurtosis . we also removed those rows which has any variable missing. For outliers we simply replaced it with the mode of the feature. We also performed many statistical techniques such as dummy variable induction, transform variables to get new variables. After cleaning the data properly , we started with basic models and moved it too heavy models. Our first model was logistic Regression it gave us 80 % test score , then we ran SVM , model gave better results but than logistic, but it requires lot of computational resources, after running decision trees , gradient boosting and light GBM with cross validation we got best score with LightGBM which uses gradient boosted trees in the backend of the sklearn . The model accuracy was 84 % for the best model . this shows that features we took and ran our model gives us decent results and are good predictors such crime reported by same county , no of days in which it has been charged , reported by which county plays a major role , which month of the year it is. Day of the week . SVM took lot of time to work on such kind of data so for this research purpose we realized it gives similar results but by taking lot of computational resources. Our research motive was to get decent result without using heavy models. The AUC score is 0.61 which seems to be a decent result.

**RECOMMENDATIONS**

There are several ways to improve the findings of the project. One of the ways is to go deep in locations and analyze the behavior of how it affects the predicting power of outcomes.

This will be a great area of research to go ahead as it can untangle the peculiar behavior of areas w.r.t crime rate and we can also go deep in figuring out which areas are biased towards a particular kind of crime. Like theft or home violence is correlated to areas in which people are living.

Another recommendation would be to analyze the anti-social behavior of people during Covid times. As we know people started to live inside their houses and the home violence incidents have increased due to people are bound to home and their psyche is being affected by the pandemic and how it is affecting their personal lives. This can be connected to happiness factor of that area, how much it has affected the lives of people with pandemic.

Also, we haven’t analyzed the 3rd type of dataset stop and search dataset, there is a lot that can be done with the 3rd type of the UK police Crime Data Set, as stop and search is a problem in itself, from that we can start predict the drug usage along the people or they are carrying weapon or there were any criminal evidence associated with the person who was stopped.

**CONCLUSION**

The project Gathered low level insights on the UK crime Data pre and post the pandemic of Covid. We were also able to figure out factors which can predict the outcome of the crime happened. The decent accuracy of models leads to good results and since they are boosted trees, so chances of over fitting is low. Results seem to be good but took a lot of time. Models seem to show expected result but since everything can be improved over time and it has been only one year of the pandemic data, with time data will evolve and patterns in the data will change as we need to take care of that as well. As the time will increase, we will also require high computational power with more stronger models. Results are easily interpretable and effective but if the data will increase, we must improve our baseline model with time for better results.

**BIOGRAPHY**

**Vishal Pathak** is a graduate student in the Data Science Program at The George Washington University. His interests include Natural language processing, Machine Learning, and reinforcement Learning. He has worked in Telecom and e learning systems in various positions for five years. He enjoys competitive coding, running , and cooking in his spare time.

Dr. **Nima Zahadat** is a professor of data science, information systems security, and digital forensics. His research focus is on studying the Internet of Things, data mining, information visualization, mobile security, security policy management, and memory forensics. He has been teaching since 2001 and has developed and taught over 100 topics. Dr. Zahadat has also been consultant with the federal government agencies, the US Air Force, Navy, Marines, and the Coast Guard. He enjoys teaching, biking, reading, and writing.

**REFERRENCES**

[1]Alonso, Ricardo Dec 2020 Applied Machine Learning in Social Sciences: Neural Networks

and Crime Prediction Ricardo Francisco Reier Forradellas \*, Sergio Luis Náñez Alonso

[2] Shyam Varan Nath 2006 , Crime Pattern Detection Using Data Mining

[3] Shraddha Ramdas Bandekar1 , C. Vijayalakshmi2 Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India 2019

[4] Shiju Sathyadevan , Surya Gangadharan. S Crime Analysis and Prediction Using Data Mining 2017

[5] David. H, Benjamin. (2020). Kernel from Kaggle comprising the starter code for demonstrating how to read the crime propensity prediction data and begin exploring it for your research projects. 10.13140/RG.2.2.18264.93445.

[6] Sathyadevan, Shiju & S., Devan & Gangadharan, Surya. (2014). Crime Analysis and Prediction Using Data Mining. 10.1109/CNSC.2014.6906719.

[7] Joshi, A., Sabitha, A.S., Choudhury, T.: Crime analysis using k-means clustering. In: 2017 International Conference on Computational Intelligence and Networks. IEEE (2017)

[8] Chen, P., Kurland, J.: Time, place, and modus operandi: a simple apriori algorithm experiment for crime pattern detection. IEEE (2018)

[9] Babakura, A., Sulaiman, M.N., Yusuf, M.A.: Improved method of classification algorithms for crime prediction. In: 2014 International Symposium on Biometrics and Security Technologies (ISBAST). IEEE (2014)

[10] Gupta, M., Chandra, B., Gupta, M.P.: Crime Data Mining for Indian Police information System. IIT Delhi, India (2006)

**APPENDIX**

**Streets Data Dictionary**

|  |  |  |
| --- | --- | --- |
| column\_id | Definition | Data Type |
| Crime ID | JSON ID | Object |
| Month | Month ID | Date |
| Reported by | Reported by which police | Object -category |
| Falls within | crime falls under which county | Object -category |
| Longitude | Longitude of the location | float |
| Latitude | Latitude of the location | float |
| Location | location name | object |
| LSOA code | Area code | object |
| LSOA name | Lsoa area name given by UK | object |
| Crime type | Crime type of different types | category |
| Last outcome category | outcome of the category | category |
| Context | context of the crime | object |

**Outcomes Data Dictionary**

|  |  |  |
| --- | --- | --- |
| column\_id | Definition | Data Type |
| Crime ID | JSON ID | Object |
| Month | Month ID | Date |
| Reported by | Reported by which police | Object -category |
| Falls within | crime falls under which couny | Object -category |
| Longitude | Longitude of the location | float |
| Latitude | Latiitude of the location | float |
| Location | location name | object |
| LSOA code | Area code | object |
| LSOA name | Lsoa area name givrn by UK | object |
| Crime type | Crime type of different types | category |
| Last outcome category | outcome of the category | category |
| Context | context of the crime | object |
| dayofweek | which day of week | category |
| weekday | is it weekday | category |
| n\_days | n th day from last crime | int |
| year | year | numerical |
| Block | which block does it belong | category |
| Same Police | reported and falls with same police | category |