Baltimore Crime Data

Ion Barbus

With 58 murders per 100,000, Baltimore crime levels rival and exceed some of most troubled regions of the world regarding crime (Sean, 2020). The skyrocketing levels of crime have put the city on edge and has depleted the resources of the weary Baltimore police department that has been rocked by scandals and corruption. As such, it is important to be able to identify the key predictors of crime, characteristics of crime by neighborhood, and to be able to forecast the situation into the future to better allocate resources. More specifically the project will focus on using the data provided by the Baltimore police department to forecast crime into the future, using time series forecasting to predict trends.

The main goals of the forecasts will be to predict what might at first seem unusual spikes in crime. These spikes might be due to things such as scandals within the Baltimore police department, events in the city, or other events and to identify more traditional patterns as a result of weather patterns, holidays or annual pattern occurrences. I also want to identify neighborhoods that are at risk and understand the key characteristics of what makes them unique. This might lead to different policing and management practices based on the neighborhood. Finally, I want to be able to identify unique predictors of crime and readjust the model based on those.

The tools used for this project will primarily include MongoDb to store the 293k rows of police department data, and to update the data monthly. I will also use the PyMongo to connect the mongo database to the anaconda distribution and more specifically jupyter notebooks where the analysis and machine learning models will be created. The primary data will be the crime data from the police department, however additional datasets will be identified and used for the models. Datasets such as weather patterns, demographic datasets, city activity events and news articles will be used to supplement the main data. I will be primarily using time series forecasting to forecast the crime incidents into the future but will also utilize the coordinates in the dataset to cluster neighborhoods into different entities and better understand each. I will also be classifying which neighborhoods are most at risk by the values from the original dataset along with the supplementary data.

Similar academic work has already been established. A few relevant papers are deal with the topic and merit attention for when undertaking such a project. The first paper I want to highlight that focuses on the same topic is ‘Crime forecasting Using Data Mining Techniques’. The paper focuses on classification for the forecast model and the t-month approach. The t-month approach states that crime which happened this month can be described by the events that came before it. Thus, the training may be January and February to predict what happens in March. The paper also focuses on the Broken Windows Theory for the attribute set, this theory describes how related categories and events may be used to describe crime. For example, certain crimes act as signals that a neighborhood is primed for crime, so foreclosures, drug dealers and bus shelters for example may indicate that a neighborhood may be at increased risk. The second source is research performed by the rand corporation to develop a reference guide for departments interested in predictive analytics. I focused gravitated towards this paper because it summarizes the entirety of the field and the most promising approaches. The methods of predictive analytics identified focuses on Methods for predicting crimes, predicting offenders, perpetrators identities and victims of crimes. It raises some concerned about privacy and civil rights, major pitfalls, and myths. My project has similarities with the work in these papers as I will also be focusing on forecasting and classification methods to predict crime. I will not however be focusing on just residential burglary and while I will be using similar methods, I will also be focusing more on how certain events will impact overall crime trends. These events will be identified by a dataset of all the special permits available in the city.

The dataset that has special permits information was in much need of some data wrangling as was the main dataset which features crime data. On the Baltimore crime dataset, I focused on preparing the data for use in forecasting model. First, I handled null values for certain columns which would have logical data gaps. For instance, larceny crimes do not necessarily have weapons associated with them, so the values were left empty, this was rectified by filling in values with placeholders. All other null values that could not logically be accounted for were dropped, which resulted in less than 5% data loss and 270k incidents to work with. The data was standardized across columns, and data types. Datetime objects were added for dates available to make forecasting easier, and Geopoints were created by utilizing the longitude and latitude listed in the data. Columns which had redundant, poorly defined, or default values were dropped to make the data frame smaller and easier to work with.

Initial analysis was also performed on the dataset. By importing shapely files of Baltimore streets and Baltimore crime cameras, I was able to plot the location of homicides, rapes, and street burglaries. From these plotted data points, you can easily identify where crime cameras are clustered and how this correlates with the amount of crimes shown in each area. Initial analysis of the dataset also painted a clear picture of which districts of the city were most at risk; Northeastern and Southeastern, which types of crimes are most common; Larceny and Common Assault, and what sort of weaponry is used most for each crime and in each district.

The third phase of the project began with some preparation of a supplemental dataset that includes dates and descriptions of special events permits granted in Baltimore. The location format in the dataset was mixed, containing both coordinates with notes, and street addresses. Using regex, I collected all the street addresses and by utilizing Geopy I was able to retrieve coordinates for the data. The coordinates already used in the dataset were also expunged of additional notes, such as ‘(closest intersection)’ and prepared for use with Geopoints and Geopandas. I also created datetime objects out of the date strings to prepare the dataset to be used for time forecasting. Events were listed with a start and end date, as well as number of days the event would go on for. In these instances, I replicated the rows and changed the start date so that an event that has gone on for ten days, would have ten instances and be accounted on each day the event moves forward. After this data was prepared, I conducted some initial analysis, plotting the events on a map, along with crime cameras to see how the two correlates.

As part of my project proposal, I was very interested in seeing how each neighborhood is different from one another so that we can gain insight based on individual communities and propose solutions for them as well. As such, the first modeling tool I used was DBSCAN with the intent to find crime hotspots in the city. DBSCAN, which stands for Density-Based spatial clustering of applications with noise, is a relatively common data clustering algorithm in data mining and machine learning. DBSCAN groups together points close to each other based on a metric and can be used to separate high density areas from lower density areas. This should give us certain hotspots for whatever we’re looking at. The size of the dataset was taxing to my system, as such a sample of the data was used when looking at all crimes. The algorithm used was Ball Tree as it provides the function to calculate Euclidian distance between neighbors, as well as haversine distance. Haversine distance is ideal to use when dealing with coordinates, such as in this dataset, as it factors in the circle distance, such as seen on the spherical earth. The initial search yielded 11 hotspots but encompassed the entirety of Baltimore. When the data was broken down into subsections for shootings, rape I was able to get interesting hot spot clusters that resembled the Baltimore neighborhoods.

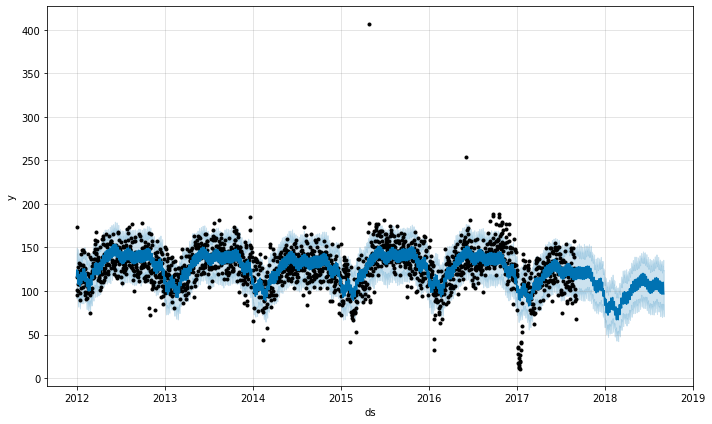
Running the DBSCAN model for Baltimore shootings has yielded 43 clusters that I was then able to plot onto a map of Baltimore. The same model yielded 51 clusters for rape, with a major cluster seen around Patterson park and the Mt Vernon as well as other regions. This data could potentially be used to adjust policing practices in certain regions of Baltimore, or to change how and which social resources are expended in those areas. The results from the model can also be used to focus special attention to certain neighborhoods and adjust my crime forecasting to fit the special characteristics of each area.

In preparation for a time series forecasting model I have also performed some time series analysis on the data to better understand how to alter the models. First, I indexed the dataset by the Date Time objects and aggregated all the instances of an event by day, week and month. Right away I notice some trends emerging that could be investigated. I also noticed a sharp outlier with a spike in crime April 27, 2015. A quick google search shows us that this was the first day when the infamous Baltimore riots turned violent. There is also a steep decline in January of 2017 that is unexplained, google searches only yielded an article confirming the drop off according to FBI data, which is based on the Baltimore city data as well. Next I wanted to see if there was any initial correlation between the events dataset I have and the crime dataset. Visually there seems to be some correlation with a spike in May on both datasets, but this needs to be confirmed. Finally, I began my initial time series modeling forecasting using Facebook’s Prophet.

According to the Facebook GitHub account, Facebook’s Prophet is an open source procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It handles outliers well and can be easily tuned. The initial model used takes only the date object and the y variable. Using these variables, I was able to make a future data frame for the next year and predict what crime occurrences will look like. Prophet returns the estimate, and the lower and upper bounds estimates as seen below:

ds yhat yhat\_lower yhat\_upper

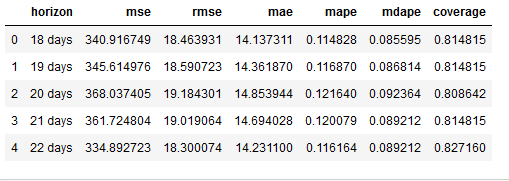
2018-08-29 104.280489 78.613682 128.743013

2018-08-30 102.961917 78.495364 129.230602 …

I will tune the model to include several regressors such as holidays, events as seen in the special permit dataset, and seasonality. The quick forecasting model also has yielded some surprising results in showing reduced forecasting for crime in the future. While looking at the analysis and the model there seems to be a drastic drop off in crimes towards the end of 2017. I will investigate this to see if its due to unreliable data or missing data.

The first work performed after the initial forecasting was to see how accurate my model had been. Initially includes functionality for time series cross validation to measure the forecasting error. The cross validation involved training the model with an initial 3 years of data and choosing a horizon of about 180 days, or roughly half a year. Calling on this feature, I received a MAPE of 11.54 at 18 days of forecasting from the cutoff, 11.68 at 19 and 11.6 at 22 days. MAPE in this case stands for the mean absolute percentage error. It is a statistical measure of how accurate a forecast system is and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values. The formula is as follows:



Later I wanted to see how big of an impact holidays would have on my model. When including holidays, the MAPE was as follows. 

This shows that holidays alone did not have a dramatic impact. The MAPE was only moved slightly for certain days. Next the event dataset was cleaned and added into the model as a regressor for each day. The dataset of events was incomplete, as such the model was only constructed for the year of 2015 which is when the dataset featured events. The assumption in using this dataset was that there would be a degree of regularity between when these events occurred and that they will occur again in the future. Adding the events to the overall model for all crime across the city did not have a significant impact.

Citation: Kennedy, Sean. “'The Wire' Is Finished, but Baltimore Still Bleeds.” *The Wall Street Journal*, Dow Jones & Company, 7 Feb. 2020, www.wsj.com/articles/the-wire-is-finished-but-baltimore-still-bleeds-11581119104.