Literature Review and EDA

Ion Barbus

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.

Papers Researched: Yu, Jacky & Ward, Max & Morabito, Melissa & Ding, Wei. (2011). Crime Forecasting Using Data Mining Techniques. Proceedings - IEEE International Conference on Data Mining, ICDM. 779-786. 10.1109/ICDMW.2011.56.

Perry, Walter L., Brian McInnis, Carter C. Price, Susan Smith, and John S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Santa Monica, CA: RAND Corporation, 2013. https://www.rand.org/pubs/research\_reports/RR233.html.