DSC423 – Data Analysis and Regression

Milestone 03 – Data Set Description

By:

Group Dogwood Rose Doppelgangers

Jesse Franks, Pete Gillespie, Reid Case, Robert Cordrey, Salik Hussaini, M. Omar Khan

For the purposes of this analysis, the dataset chosen originally contained approximately 425,831 data objects, each indicating a single reported incident of crime within the Boston Metropolitan area, and 21 attributes. The data spans a range of dates from January 1, 2016 until September 25th, 2019. This dataset currently contains data that will be used to establish a relationship between crime and potential predictive attributes such as geographic location, weather data, distance from police stations, population density, day of the week. Mixed into this data are some additional attributes including, the nature of the criminal activity (i.e. shooting, robbery, drug trafficking, etc.), and reporting area/district. The goal is to determine what, if any, relationship exists between the frequency and nature of criminal activity, and the various explanatory attributes listed within the dataset; With these relationships in mind, we hope to create a regression model, to speculate the nature of a future crime. Prior to conducting this analysis however, several changes needed to be made to provide a clearer and more concise dataset.

Prior to organizing the approach for conducting the analysis, it was determined that the nature of criminal activity presents several challenges unto itself. Primarily, criminal behavior varies in its severity and type, and as such, this data object needed to itself be categorized to best interpret the results of any analysis. Fortuitously, the dataset that was obtained contained a form of categorization based on the nature of the offense committed labelled "OFFENSE_CODE_GROUP" where in classification was more easily determined based on the code assigned. As such, for the purpose of this analysis, the nature of the crime was fit accordingly to the coding presented as follows:

Violent Crime	Aggravated Assault, Ballistics, Biological Threat, Bomb Hoax, Criminal Harassment, Disorderly Conduct, Explosives, Firearm Discovery, Home Invasion, Homicide, Human Trafficking, Human Trafficking - Involuntary Servitude, Manslaughter, Offenses Against Child/Family, Robbery, Residential Burglary, Simple Assault

Property Crime	Arson, Auto Theft, Auto Theft Recovery, Burglary – No Property Taken, Commercial Burglary, Confidence Games, Evading Fare, Larceny, Larceny from Motor Vehicle, Other Burglary, Property Lost, Property Related Damage, Vandalism
White Collar Crime	Aircraft, Counterfeiting, Embezzlement, Fraud, Gambling, Prostitution
Minor Crime	Assembly or Gathering Violations, Harassment, Investigate Person, Investigate Property, Landlord/Tenant Disputes, License Plate Related Incidents, License Violation, Liquor Violation, Motor Vehicle Accident Response, Operating Under the Influence, Phone Call Complaints, Police Service Incidents, Prisoner Related Incidents, Restraining Order Violations, Towed, Verbal Disputes, Warrant Arrests
Drugs	Drug Violation
Other	Fire Related Reports, Harbor Related Incidents, Medical Assistance, Missing Person Reported, Other, Search Warrants, Service, Violations, Warrant Arrests

With these classifications, the analysis is designed to better predict what explanatory variables affect criminal behavior by specific category.

Within the dataset obtained, it was noted that some attributes were unnecessary information and as such, it was determined that in order to produce a relevant analysis only those attributes deemed essential to the model were incorporated (see Appendix A for a list of these attributes). While it is understood that in any regression the potential for hidden attributes exists, it was determined that the essential attributes covered in broad terms the most direct explanation for the response variable. Thereby it was noted that several attributes were themselves redundant in comparison to the essential attributes or provided no meaningful value to the interpretation of the data in question.

Specifically, eight total attributes were eliminated from the available dataset. This included fields such as the INCIDENT_NUMBER, which contained data specific to the Boston Police Department and offered no additional information, OFFENSE_CODE, which contained numerical codes that were again only relevant to the Boston Police Department and contained no additional information, OFFENSE_DESCRIPTION, which contained a descriptor of the reported incident. Additionally, UCR_PART, which contained the universal crime reporting part

number and ID, which contained the station ID for the weather station recording the relevant weather data, were removed to provide for cleaner data. This information, while interesting, did not provide any immediate or relevant data related to the predictability of an incident. Additionally, several redundant attributes were noted including columns for YEAR, MONTH and Location, all of which were reported elsewhere within the essential attributes.

Once the substance of the relevant data was determined, further refinement was necessary in order to normalize the data for the purposes of analysis. Specifically, population data for the areas surrounding police stations were copied into the dataset. This data is normalized to provide a reportable figure for the number of crimes per population unit (i.e. crimes per 10,000 people). With this normalization it will be possible to compare the number of crimes reported from one police district to another.

Additionally, the average daily temperature information was data that was arithmetically derived from the hourly temperature reading data available as noted at Logan International Airport. In order to calculate the average daily temperature, hourly readings were summed and divided by the 24 hours of the day to arrive at an average temperature for the date in question. This data was presented in Fahrenheit and rounded to the nearest whole number. Furthermore, in order to synchronize this data with the dataset obtained for this analysis, a join was created using the “Occurred_On” column as an anchor to match average daily temperature to the date a crime was reported. As the weather data covered the time period of January 1, 2016 until September 25, 2019, all crime data subsequent to the latter date was eliminated from the analysis, resulting in a removal of approximately 500 entries from the original dataset.

Furthermore, latitude and longitude data were transcribed into the data set using available GPS coordinates of police stations. The purpose of obtaining this data was to determine the distance between reported crimes and the nearest police station. The analysis uses the Haversine formula to calculate the distance between two coordinates in an “as the crow flies” metric. The resulting “Distance_To_Station” attribute was created and added to the dataset, using the “District” attribute as an anchor for joining the data.

Finally, the time of day of the criminal activity was standardized to ensure uniformity in the data. Specifically, the criminal activity was rounded to the nearest hour and the hour of the crime was converted into the 24 hour “military time” convention (i.e. 11:00 pm is reported as 23). By doing so, the analysis will be able to provide a higher degree of precision regarding the time of day of criminal activity.

After processing our dataset, we are left with 393,188 data objects in the form of reported crimes, and 17 attributes, 16 of which represent potential explanatory variables. From the original dataset, attributes were determined to be essential based on the value of their information, as well as to eliminate any possible redundancies. Additional attributes were added from other datasets, or calculated using data currently available in the dataset, these included data regarding the distance to the closest police station and the average daily temperature for

reported crimes. Using this information, it is the intention of this analysis to determine the various possible factors involved in criminal activity and to better predict such activity.

APPENDIX A

Field Name, Data Type	Description
[OFFENSE_CODE_GROUP] [varchar]	Internal categorization of offense.
[DISTRICT] [varchar]	What district the crime was reported in
[REPORTING_AREA] [varchar]	RA number associated with the where the crime was reported from.
[SHOOTING] [binary]	Indicated a shooting took place. 0 - No, 1 - Yes
[OCCURRED_ON_DATE] [datetime2]	Earliest date and time the incident could have taken place
[STREET] [varchar]	Street name the incident took place
[DAY_OF_WEEK] [varchar]	Day of the week incident took place
[HOUR] [integer]	The hour wherein the incident took place based on a 24:00 hour reporting system, rounded to the nearest hour
[Lat] [decimal]	Latitude GPS coordinate where crime took place.
[Long] [decimal]	Longitude GPS coordinate where crime took place.

[Population] [integer]	Population of neighborhood police district is in.
[Area] [decimal]	Area of neighborhood police district is in (in sq. miles).
[Average_Daily_Temp] [integer]	Daily average temperature from BOS (in deg. F).
[Station_Lat] [decimal]	Latitude GPS coordinate of police station.
[Station_Long] [decimal]	Longitude GPS coordinate of police station.
[Distance_To_Station] [decimal]	Calculated distance of crime to police station (in miles).
[Recode] [varchar]	Generalized grouping of “Offense_Code_Group”