DSC423 – Data Analysis and Regression

Milestone 02 – Data Set Selection

By:

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

Preventing crime can be expensive. According to boston.gov, for the year 2019, the budget for the Boston Police Department was $400,000,000. $361,589,110 went to personnel services alone. Doing some simple math, that is approximately 90.4% of the budget. Based on this data alone, there is a need for crime research and development so as to better deploy vital resources and reduce the incident of criminal behavior. Solving this problem can be tricky, however we will create a regression model that uses factors like location, time of year, time of day, nature of the crime, affected city district, and local weather to predict what crime will occur.

The response variable we will use in this project will be offense code groups (types of crimes). We hope to predict the type of crime that will be committed. We will also use the population and area of the different police districts to normalize the data, as we expect larger police districts to have more crime in general. The explanatory variables we will explore include the district that the crime was reported in, the area where crime was reported from, whether a shooting occurred, the day and time the incident took place as well as the location (including street, latitude and longitude). We will also include average daily temp (measured at Logan International Airport) to look at the relationship between temperature and certain crimes.

The Boston crime data was retrieved from https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system and covers the period of 2015 – 2019. The population and area data was also retrieved from the Analyze Boston website. The weather data was collected from a CRAN package named stationaRy. The specific data extracted is for the data recorded, hourly, at the Logan International Airport for the same period as noted above. Those rows with a NA value were removed and were deemed not necessary. A new column was added to convert the Celsius temperature to that of Fahrenheit. The daily hourly readings were grouped together for each day and aggregated the mean temperature. A new table was created containing only the date and the mean temperature for each day. It is this new table that is being used for the project.

We have two datasets. A primary Boston crime dataset and a supplemental Boston weather dataset. The date range is 2015 to 2019.

Data legend: Red = columns to remove due to duplication or unuseable, Green = dependant variable, Blue = explanatory variable.

The dataset currently includes the following data:

- **19 columns:**
  - INCIDENT_NUMBER: string, categorical, Internal BPD report number, will be removed.
  - OFFENSE_CODE : string, categorical, Numerical code of offense description, will be removed.
  - OFFENSE_CODE_GROUP : string, categorical, Internal categorization of offense_description, Y
  - OFFENSE_DESCRIPTION: string, categorical, Primary descriptor of incident, will be removed.
  - DISTRICT: string, categorical, What district the crime was reported in, X
  - REPORTING_AREA: string, categorical, RA number associated with the where the crime was reported from, X
  - SHOOTING: Boolean, Indicated a shooting took place, X
  - OCCURRED_ON_DATE: time/data. R has its own time_data data type, Earliest date and time the incident could have taken place, X
  - YEAR: string, categorical, will be removed.
  - MONTH: string, categorical, will be removed.
  - DAY_OF_WEEK: string, categorical, X
  - HOUR : time/data. R has its own time_data data type, X
  - UCR_PART: string, categorical, Universal Crime Reporting Part number (1,2, 3) will be removed.
  - STREET, string, categorical, Street name the incident took place X
  - Lat: spatial, lat is Y coord, X
  - Long: spatial, lat is X coord, X
  - Location: ordered pair, of lat long, will be removed.
  - POPULATION: integer, discrete, X
  - AREA: double, continuous, square miles, X
- **Num of rows: 393,748**

Boston Weather:

- **10 columns:**
  - ID: sting, primary key, will be removed.
  - TIME: time/data. R has its own time_data data type, X. Will join on this column.
  - TEMP: doubles, continuous, X
  - WD: integer, discrete, X
  - WS: doubles, continuous, X
  - ATMOS_PRES: doubles, continuous, X
  - DEW_POINT: doubles, continuous, X
  - RH: doubles, continuous, X
  - CEIL_HGT: integer, discrete, X
  - VISIBILITY: integer, discrete, X
- **Num of rows: 43,825**

As mentioned, crime prevention can prove to be expensive, the primary difficulty with our data is that we do not have granular costs associated with specific crime prevention measures in response to specific crimes, nor do we have detailed budget allocation data of the Boston PD. As a result, we will not be able to project potential cost, or cost savings, through this analysis. We have chosen to collect weather, population, and police district data along with the primary data set of reported crimes. Regarding the weather data, attempts were made to pull specific historical conditions based on coordinates and time stamps. This effort was met with pay-walls established by weather API providers that exceed reasonable expenditure limits for this project. As such, we have settled on a broader data set that is collected from a common location for the date range of the primary data for the study. All four data sets do not share the same indexing. The final challenge will be that we will need to perform extensive joins and transformations to bring all data into a single usable set for analysis. This may result in some data not translating to the primary dataset in an ideal fashion.

In summary, using publicly available data on crimes reported in the Boston Metropolitan Area covering the period of 2015 - 2019, we are attempting to predict how accurately the aforementioned factors indicate the likelihood of crimes being committed. Through our regression analysis, our efforts will ultimately determine what predictive factors exist for criminal behavior, thereby allowing the City of Boston to better analyze criminal activity, and to deploy its resources more efficiently. Additionally, the results of our analysis may hopefully provide cheaper solutions to offsetting criminal activity. By analyzing data such as weather, time of day, day of the week, latitude and longitude of criminal activity, etc. we intend to establish what, if any, relationship exists between these explanatory variables and our dependent variable. We understand modeling human behavior is a daunting undertaking and is certainly rife with the opportunity for pitfalls and drawbacks. Specifically, the broad nature of weather data, and the lack of particularly granular data regarding budget allocations or criminal prevention activities in specific geographical locations presents a challenge as these may be statistically significant variables that remain unaccounted for. Despite this, the very nature of any effort in the prediction of so inconsistent a creature as human beings is to accept a degree of unpredictable variability. We endeavor to meet this challenge.