Explanation of the Data Preparation Process

I plan to use the Chicago crime dataset predict which police district experiences the highest levels of crime. By knowing which police districts experience the highest crime levels, resources, such as police personnel or police cars, can be efficiently deployed across districts. Although I am focusing on only one target variable for this predictive model, I would also be curious to research. For this article summary, I will be reviewing the Data Audit and Preparation process.[1]

The Data Audit node revealed variables containing both good quality data and quality issues with the data. First, looking at the Location Description field, some values are too granular and can be consolidated, such as Airport-related values, CTA-related values, and residence-related values. Second, a look at the Latitude histogram appears to resemble a bimodal distribution rather than a normal distribution and may need further investigation. Third, the Longitude histogram has a noticeable long left tail, with a skewness of -0.213. Further, the Longitude variable contains 55 outliers. Fourth, the Data Audit node also revealed missing data in the Latitude, Longitude, District, and Location Description fields. The Location Description field contains 5 empty string/white space missing values. More concerning, the District, Latitude, and Longitude fields are all missing 157 values, effectively reducing the percent of complete fields to 71.43%. The data nodes used to improve the data modeling process were Reclassify, Distribution, Histogram, Filter, Statistics, Partition, and Anomaly. Other fields, such as Date, IUCR, Crime Type and SubType, Arrest, Domestic and Community area were good quality fields as outliers, skewness, and missing data were not present. Although my focus is on the spatial variable, the Date variable shows a noticeable temporal pattern in crime rates.

To address the issue of granularity, I used the Reclassify node to consolidate Location Description values to AIRPORT, CHA, COLLEGE/UNIVERSITY, CTA, RESIDENCE, SCHOOL, OTHER, PARKING LOT, VEHICLE, STORE, FINANCIAL INSTITUTION, HOSPITAL, BUILDING, PUBLIC/PRIVATE SPACE, HEALTHCARE FACILITY, POLICE FACILITY and SCHOOL. The consolidation of these fields to broader categories allows us to more easily identify patterns by looking at the overall picture.

To investigate the distribution of the variable Latitude, the Histogram node did indeed reveal a bimodal distribution. The values were separated into two equal bands highlighting the trough at 41.834 compared with the mean of 81.482. The peaks occur at 41.758 and 41.879. This could be important as it reveals two main crime areas rather than one; however, this could also be indicative of population density rather than high-crime areas. Looking at the Histogram node for Longitude reveals outliers and a skew to the left. One way to address the issue of outliers is to run the model both with and without the outliers to determine their impact on the predictive model. Additionally, the Statistics node revealed a strong correlation between Ward and the Latitude/Longitude fields. Consequently, the Ward field was filtered out to eliminate redundancy.

1. For information on the Data Understanding phase, please refer to the Appendix.

Taking a look at the Feature Selection node, the only fields filtered out were Latitude and Longitude due to their low variation; however, location is an important variable in predicting crime. Therefore, I will retain these variables and adjust the model as necessary. All fields were ranked as important, so all fields remained. While the Feature Selection node was not particularly useful for this dataset, it is still important to explore.

Lastly, anomaly detection is important because it can skew the prediction model, making it less reliable. To address missing values in the Latitude, Longitude, and District fields, the values were imputed. The Anomaly node revealed a total of six anomalies in two peer groups. Two anomalous records identified in the first peer group were IUCR, Crime SubType, and Longitude fields. The four anomalous records identified in the second peer group were Crime Type, Community Area, IUCR, and Crime SubType. The Longitude field is of particular interest since the target variable is District. A histogram revealed anomalies present in Districts 4 and 16.[2] Both of these districts also experienced higher levels of crime compared with other Chicago districts. I will retain these district anomalies for the predictive model. All other fields generated by the Anomaly node were filtered out since they will not make any meaningful contribution to the model.

2. The number of bins for the histogram were set to 23 since there are 23 police districts.

Appendix - Data Understanding

To start the Data Understanding Phase, the data was imported using the Excel node in the Sources palette to read the data in the Excel file. During the import process, the following redundant and/or irrelevant fields were filtered out: Case Number, Block, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, and Location. The redundant fields were filtered out to prevent multicollinearity which can produce poor models. The FBI Codes were filtered out rather than the IUCR Chicago codes because the area of interest is crime in the city of Chicago specifically, as opposed to comparisons between cities, towns, or across state lines. The Latitude and Longitude fields were chosen because they are universally accepted geographic standard and some SPSS data visualizations require two continuous fields. The Case number field was excluded since we only need one unique identifier and ID is a continuous field. The remaining irrelevant fields were filtered because they did not add any value and added unnecessary noise.

Using the Type node, the measurements of Categorical variables were further defined to improve the Data Audit analysis. Both Arrest and Domestic variables were changed from Categorical to Flag since the values were true/false. The remaining categorical fields were changed to nominal. Latitude, longitude, ID and date fields were correctly classified as continuous. Lastly, the role of the ID field was specified as the Record ID and the District field to target, since I would like predict which Districts will experience the highest crime rates.