Assignment 2 - Modeling (Classification)

1) Auto-classifier or auto-numeric output.

The auto-classifier node was selected to run predictive models on the crime dataset with District Region identified as target, and IUCR, Crime Type, Crime SubType, Location Description, Arrest, Domestic, Time, and Date as available inputs. The top three models generated from the auto-classifier node were C5.0, Logistic Regression, and CHAID. IUCR had the predictor importance of 0.25, followed by Location Description of 0.21. Domestic, Crime Type, Crime Subtype, and Arrest all had about the same predictor importance, between 0.13-0.14. District Regions 4, 5, and 6 experienced the highest crime rates.

The C5.0 model selected IUCR, Domestic, Location Description, and Arrest as input variables. For this model, the accuracy for the training set was 37.4% and the accuracy for the testing set was 28.2%, yielding an overfitting of the data by 9.2%, which seems high. Because this model was first split by IUCR, it was difficult to discern patterns due to the wide range of IUCR codes.

In addition to the fields selected by the C5.0, the Logistic regression model also included Crime Type and Crime Subtype as input variables. The accuracy of the training set and testing sets were 33.5% and 26.9%, respectively, indicating an overfitting of the data by 6.6%. While this model yielded a slightly lower overall accuracy, additional variables were included and the overfit was significantly lower compared with C5.0 model. Again, because this model was split by IUCR codes, the equations for each district were lengthy with difficult discerning a pattern or draw comparisons. Reclassifying IUCR codes in to a smaller number of available values would most likely improve the reading of both this model and C5.0.

The CHAID model selected IUCR, Crime Type, Location Description, and Arrest as input variables for the model. The training set yielded an overall accuracy of 30.1% for the training set and 26.9% for the testing set. Although this model included fewer variables, the overall accuracy for the testing set appeared to be the same as Logistic Regression, with a lower overfitting to the data at 3.2%. The decision tree for CHAID first split on Location Description, since it had the highest predictor importance. Crimes identified at Apartments, Residence, and Schools generally occurred in District Region 6. Crimes identified at locations Store, Drug Store, Chicago Housing Authority, Chicago Transit Authority, and Police facility occurred in District Region 3. This finding is not entirely surprising, since District Region 3 is primarily downtown Chicago. For crimes in Airports, Vehicle, and Parking Lots, arrests were generally made in District Region 2. Which is to be expected since both the O'Hare Airport was classified in this district region and Airport parking lots were classified in Parking Lots. Crimes with all other Location Descriptions were more diversely distributed across District Regions.

2) Rationale for model selection.

Although the CHAID model was listed third, the overall accuracy of testing data performed only 1.3% lower than the C5.0 model and the same as Logistic Regression. Considering the overfitting of the data to the C5.0 model was 9.2% and 6.6% for Logistic Regression, the CHAID model is the least overfit fit to the data. Additionally, the inclusion of two additional variables in the Logistic Regression model did not produce a higher accuracy or much additional benefit. Although a small amount of accuracy is sacrificed, the benefit to reducing overfitting to the data outweighs the cost.

3) Model refinement process
Using the best fitting model's node, make any adjustments needed to improve the model accuracy. Consider the inputs and whether or not they should be excluded or transformed, consider the settings in the modeling algorithm tabs, and consider the impact of these changes on model accuracy using the analysis node.

To improve the model, the main objective was changed to focus on model stability through bootstrap aggregating. While I performed transformations on the Time variable, this did not appear to improve the model. Moreover, neither boosting District Regions nor Domestic/Arrests improve the model accuracy. In the auto-classifier node, CHAID selected IUCR, Crime Type, Location Description, and Arrest as optimal input variables. In following with this, I limited the inputs to these variables and Domestic. Including the Domestic variable still provided some improvement. The analysis node did produce some modest improvement in the model, with the training accuracy at 32.9% and the testing accuracy at 28.8%. This testing accuracy did increase enough to exceed the C5.0 testing accuracy, the highest ranked model. Although the model increased the overfit to the data by 4.1%, this result is more desirable compared with the C5.0 and Logistic Regression models. In addition, the tradeoff for the refined CHAID model was approximately a 1% increase in overfit to the data but yielded a 2% increase in model accuracy.

4) Model results explanation

To estimate Districts for the predictive model, they were reclassified into six District Regions. The first region is comprised of northwestern districts 19, 20, and 24. The second region is comprised of northwestern districts 14, 16, 17 and 25. The third region is comprised of central Chicago districts 1, 2, and 18. The fourth region is comprised of western districts 10,11, 12, and 15. The fifth region is comprised of southwestern districts 7, 8, and 9. Lastly, the sixth region is comprised of southeastern districts 3, 4, 5, and 22. By collapsing individual districts into district regions, the application is broader in scope.

The CHAID model identified Arrest, IUCR, Crime Type, and Location Description variables as equally important with a predictor importance of 0.21. Domestic, as

expected, ranked lower at 0.15. Overall, southeastern Chicago districts (District Region 6) experienced the highest amount of crime, followed by western districts (Regions 4) and southwestern districts (Region 5). When looking at where to deploy human resources, Location Description can be used to identify which types of places are more likely to experience high crime rates. Apartments, for example, experience a significant amount of crime, so those areas may be more heavily patrolled. Crime type was also a strong predictor in the model, and can also be incorporated when making decisions about where to deploy resources.

Unfortunately even after selecting the best-fitting model, the overall model quality was suboptimal with 32.9% accuracy for the ensemble. With the Districts divided up into six regions, there is a 1/6 in chance a random estimate will produce the correct result. With an accuracy around 1/3 chance, the model does make better estimates for which District Region will experience crime compared with random selection; however, with such a low percentage accuracy the model is less than ideal. For this reason, the impact of this model would be modest at best given the low accuracy and low precision with respect to individual districts. Before applying this model to police districts, I would first recommend adding more data since the dataset only covers two months or explore other potential variables as targets. For example, Crime Type or Time may yield better predictions about future crime than District and would be worth exploring.