**Using Machine Learning Techniques to**

**Perform Predictive Analysis on a Durham Crimes Dataset**

Mackenzie B, Kevin C, Austin D, Tony L, Alicia A

STOR 565-001: Machine Learning

Spring 2018

Dr. Yufeng Liu

**Abstract**

Predictive policing is used in law enforcement to identify potential criminal activity by utilizing mathematical and analytical techniques, namely machine learning. In this project, we used a dataset consisting of 120,000 crimes that occurred in Durham County during the past few years with predictors such as latitude/longitude and time of day. Using the original dataset, we cleaned the data by formatting and grouping crimes by similar factors. Given the existing latitude and longitude variables, we created new classifiers of each crime that attributes the crime to certain major areas in Durham, such as Duke University or the American Tobacco Campus. Then, we used both the old and new variables to answer questions such as: what time of the day certain crimes are most likely to occur, what areas of Durham are most prone to certain types of crime, and if there are groups of crimes that are similar in time and location.

To investigate these questions, we used a number of machine learning techniques; namely linear models, logistic regression, LDA/QDA, classification trees, and clustering methods. Due to a number of factors, such as lack of diversity and quantity in the predictors, test errors for many of our methods were high. However, we did find that location was not a significant predictor in crimes and that crimes tended to occur during the day, with the exception of vehicle crimes. For future analysis, more diverse variables such as criminal demographics and income level of locations would likely prove more useful (in addition to our current predictors) in predicting crime time, type, and location.

**References**

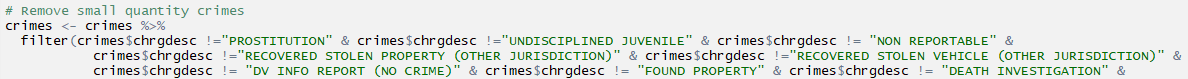
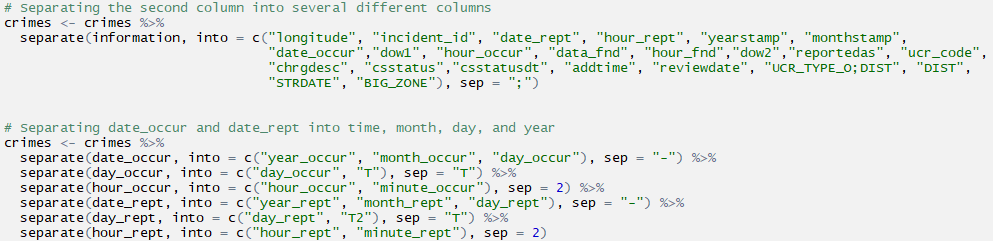
Dataset Source: City and County of Durham. (2017). *City of Durham Police Crime Reports*. Available from Durham Open Data Website: <https://opendurham.nc.gov/explore/dataset/durham-police-crime-reports/api/>

**Using Machine Learning Techniques to Perform Predictive Analysis on a Crimes Dataset**

Durham city is notorious for having a much higher crime rate than its surrounding cities. In fact, three of its neighboring cities, Apex, Cary, and Morrisville, are some of the safest cities in North Carolina. To explore this discrepancy in crime rates between nearby cities, we must consider what types of crimes are being committed in Durham. For our project, we investigated the relationships between types of crime in Durham, hour of the day, time of the year, and location. To examine these relationships, we formulated three major questions: what time of the day are certain crimes most likely to occur, what areas of Durham are most prone to certain types of crime, and are there groups of crime that are similar in time and location. This project is important and interesting because law enforcement can use the data to catch potential criminals before a felony is committed. By researching patterns in crimes, the police force can station more officers in areas where a major crime is likely to occur. The end goal is to create a safe community for citizens to live in.

*Cleaning the Dataset*

Our original dataset of 120,000 crimes had several issues. Although there were 24 different variables on each crime, the excel file condensed all the variables into 2 columns. To separate these columns, we used the tidyr and tidyverse packages in R. In addition, we separated each date of occurrence and date of report into their respective year, month, and days to see if these additional variables might be useful in our analysis. The separate() function was used to split the columns into sub-columns. We also found that counts for certain crimes were small in quantity. In addition, some of the crimes were labeled as “non-crimes” or “non-reportable”. We took these two categories out of our dataset using the filter() function.

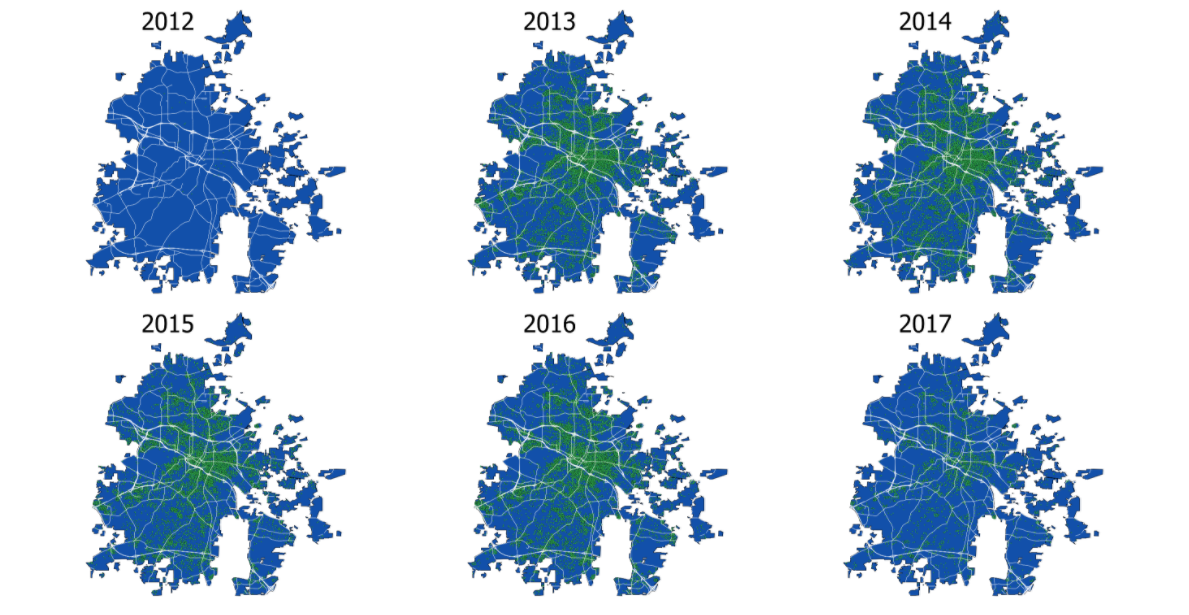
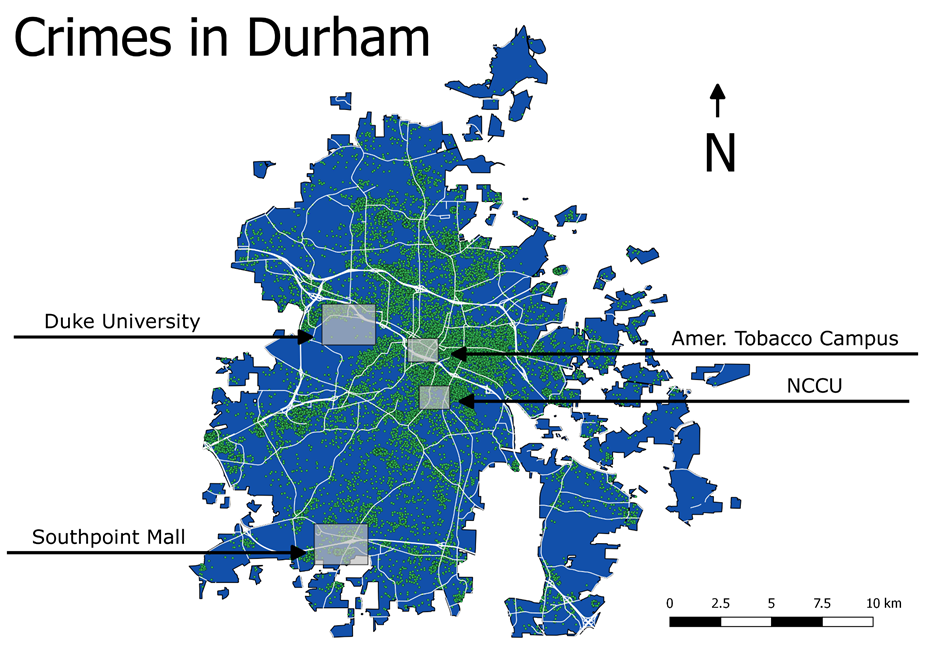


**Top**: Separating the original dataset of two columns into several columns that corresponds to each variable.

**Bottom**: Removing crime counts less than 150.

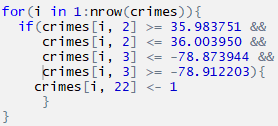
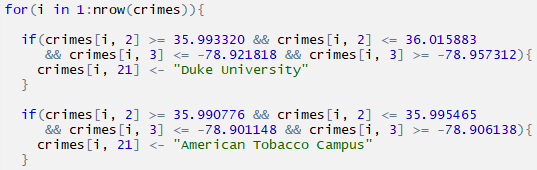
Next, we created a binary and a categorical variable. For the binary variable, we created a column that classifies each row as 1 if located in downtown and 0 otherwise. For the categorical variable, we used the latitude and longitude coordinates to classify whether a crime occurred in a major area in Durham. These major areas included Duke University, American Tobacco Campus, and Durham Tech. Our goal when creating these classifiers was to see whether certain areas or college campuses are more prone to certain types of crime. To assign these binary and categorical variables, a for-loop can be used to iterate through each row of the dataset and assign values based on the latitude and longitude coordinates of each observation. A major drawback of this process, however, is that the coordinates used to classify each observation create a rectangle shape on the map. Obviously, each landmark isn’t a perfect rectangle, and so the classifications are only a rough approximation of each landmark. The process can be improved by using non-rectangles that best fit the shape of each area such as circles or triangles. Due to the scope of this project, however, we chose rectangles to approximate each classification.

Next, we found that some of the crimes in the dataset were reported as crimes that occurred several years ago. One particular crime was reported in 2014, but occurred in 1945. We removed these observations in order to look at recent crimes, because time might be an extraneous factor when running our analysis. In the end, we saw that the scope of our dataset was between the years 2012 and 2017. 2012 had very few crimes, however, because only those that occurred on the last day of 2012 were reported. Similarly, 2017 had few crimes because only those that were reported during the first half of the year were included.



**Left**: Rectangular plots of land were used to generate categories that classified each observation in the dataset. These plots are only an approximation, since each area isn’t perfectly rectangular.

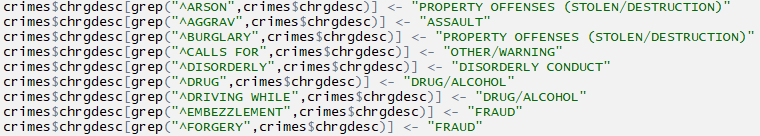
**Right**: There were very few crimes in 2012 because the dataset began on Dec. 31, 2012. Similarly, there were only a few crimes in 2017 because the dataset ended in mid-2017.



**Left**: For loop used to classify each observation to a major area: Duke University, American Tobacco Campus, City Center, Duke University, Durham Tech, North Carolina Central University, Southpoint Mall, or “other”.

**Right**: For loop used to classify each observation as 1 if in downtown and 0 otherwise.

Lastly, we found that there were over 150 different types of “reported as” and “charged as” crimes. Note the distinction between “reported as” and “charged as”, because not all crimes were charged as the same thing that they were reported as. To simplify future classifications, we kept the “reported as” column the same, but greatly reduced the number of distinct “charged as” crimes. We simplified the 150+ types of crimes into 9 major categories by grouping together crimes that are similar by nature. After cleaning the dataset and filtering out information that would be useless to our analysis, the number of observations was reduced greatly from 120,000 to 70,000 crimes.



A snippet of reducing more than 150 different types of crime to 9. Each crime was manually classified into a broad category. One of the major challenges of this project was figuring out what crimes could be grouped together, so that future regressions and classification methods would run more smoothly.

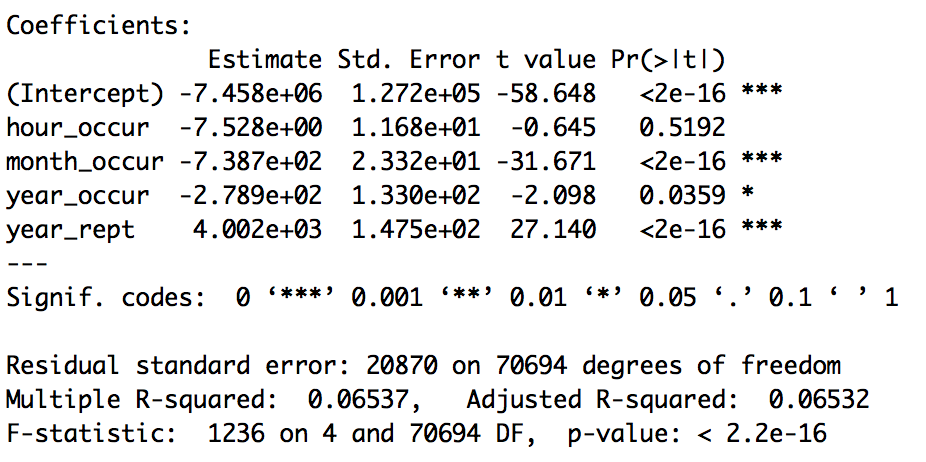
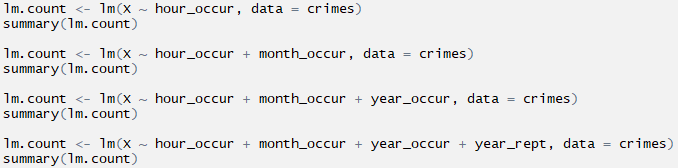
*Methodology*

To investigate crime distributions by time, we used a simple linear model. Next, we used logistic regression, LDA, and QDA to classify and predict whether a crime is more likely to occur during the daytime or at night. Classification trees were used to classify and predict types of crime. In this method, we compared the results of bagging, random forest, and a new method that we did not learn in class called C5.0. The reason that C5.0 was used is because classification trees only allow for binary splits. However, we had 9 major types of crimes that we wanted to classify, so C5.0 allowed us to have non-binary splits.

An unsupervised learning method that we used in our project is clustering. We used k-means and hierarchical clustering to investigate whether there are patterns in the dataset that are not obvious to the eye. After all, there are over 150 types of “reported as” crimes, and 9 types of “charged as” crimes. Therefore, clustering allows us to see what types of crimes are similar to each other in terms of location and time of the day.

*Linear Models*

In this dataset, the number of data points is much greater than the number of predictor variables (~20 useful predictor variables vs. ~70,000 observations). Since the dataset is low-dimensional, a linear regression would be preferred over a ridge or lasso regression. Forward selection is used to find the optimal linear model. The lm() function is used to fit a linear model to the dataset.



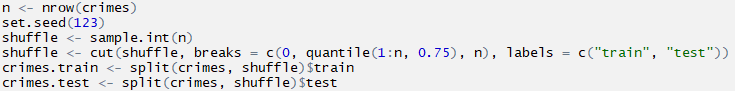
**Left**: Forward selection is used iteratively. Variables are added to the model one at a time. In each iteration, each variable that is not already included in the model is tested, and the one that is most statistically significant is added.

**Right**: The summary results of the last iteration.

At the beginning of forward selection, the hour occurred is the most significant predictor variable of crime counts. Each subsequent iteration added month occurred, year occurred, and year reported to the model. From the summary statistics of the last iteration, we found that the adjusted R-squared is only 0.07. The adjusted R-squared does not explain nearly as much variation as we would have liked. In the last iteration, the hour occurred, which was initially the most significant predictor variable, resulted in a t-test value of 0.5192, which is not significant at all. Therefore, we can conclude that fitting a linear model to this dataset is not useful. One reason that this may be the case is that most of the variables in the dataset was associated with either time or location (hour occurred, latitude, longitude, etc.). These variables alone may not be enough information to predict crime count. Useful variables that might be important but were not included in the dataset are income levels of the location that the crime was committed and description of the criminal such as gender or age.

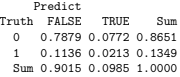
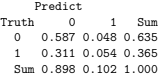
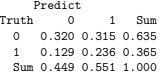
*Logistic Regression, LDA, QDA*

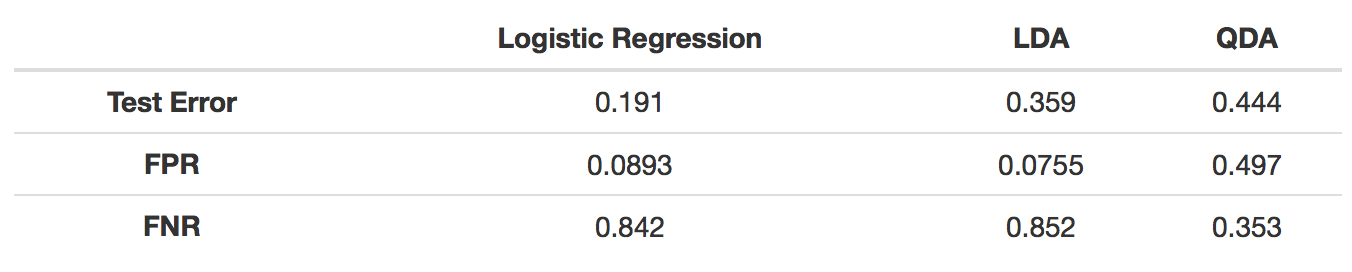
Logistic regression, linear discriminant analysis, and quadratic discriminant analysis were used to predict when a crime occurs based on the type of crime and the location. Before performing the analysis, we created a binary variable in the dataset that determined whether it happened during the day or at night. Night is defined as between the hours of 8pm and 6am. So, logistic regression and LDA/QDA is appropriate because we are working with a binary classifier of night or day. Next, we randomly divided the data into a training and test set. 75% of the observations were assigned to the training set, and the remaining 25% were assigned to the test set. The sample.int() and cut(), and split() functions are used to create the two sets.



Randomly splitting the data into 75% training observations and 25% test observations. The training set is crimes.train and the test set is crimes.test.

After running the regressions, we created proportional confusion matrices for logistic, LDA, and QDA. Then, we calculated the test error, false positive rate, and false negative rate for each regression in order to compare and contrast each of the methods. It might be worth noting that in the context of our project, a false negative is worse than a false positive because a false negative means that we are missing a potential crime, and a false positive would just mean that we are spending extra resources on a crime that didn’t actually happen. The confusion matrices can be created using the table() function.



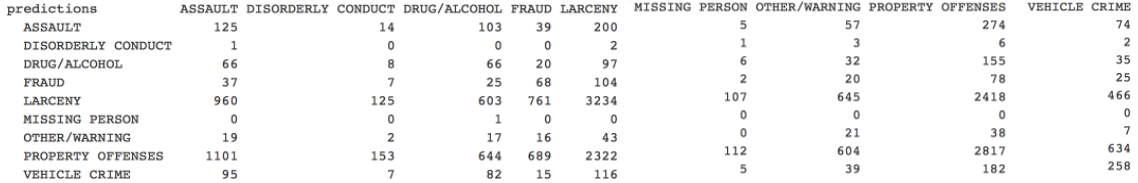
**Top, left to right**: Proportional confusion matrices for logistic regression, LDA, and QDA.

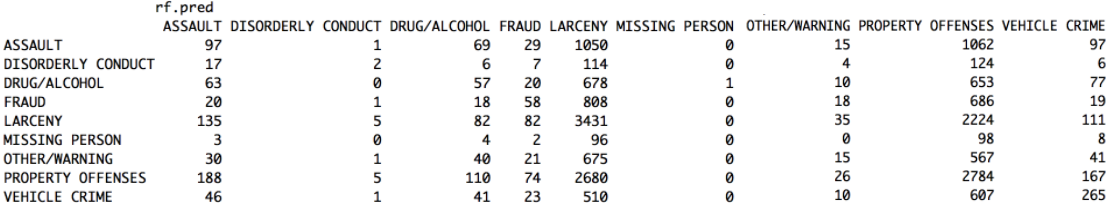
**Bottom**: A summary of the test errors, false positive rates, and false negative rates for logistic regression, LDA/QDA.

Based on test error, the logistic regression vastly outperformed LDA and QDA, with an error rate of 19%. However, we are more interested in the false negative rate for this dataset, and logistic regression gave the worst FNR out of the three. QDA performed substantially better than logistic regression and LDA in terms of FNR, with a rate of 35%. The reason why the test error was so low for logistic regression but the FNR rate was so high is because the FPR rate was relatively small. Although we are less interested in the false positive rate, logistic regression and LDA had a relatively low FPR, at 9% and 8%, respectively.

*Classification Trees: Bagging, Random Forest, and C5.0*

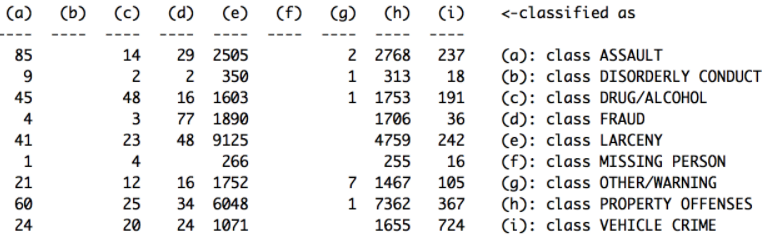
Next, we used bagging, random forests, and a new method called C5.0 to classify and predict types of crime based on the variables in the dataset. These methods were implemented in hopes of reducing the variance and thus improving classification. C5.0 is a method that uses a different splitting algorithm than the classification trees discussed in class. While classification trees measure Gini impurity, C5.0 chooses features that provide the greatest information gain. The advantage of using this method over traditional classification is that it allows for non-binary splits. In other words, since there are 9 total types of “charged as” crimes, binary splits would not accurately depict 9 types of crimes. Therefore, we need to use non-binary splits.





**Top**: Results of bagging **Bottom**: Results of random forest

For bagging, the out-of-bag estimate of misclassification error is 45%. For random forests, the misclassification error on the test set is 68%. In the results of the bagging method, most misclassified cases were classified into categories that were similar to them. For example, larceny was largely misclassified as property offenses, and property offenses were largely misclassified as larceny. Random forest tends to misclassify larceny and property offenses regardless of true classification.

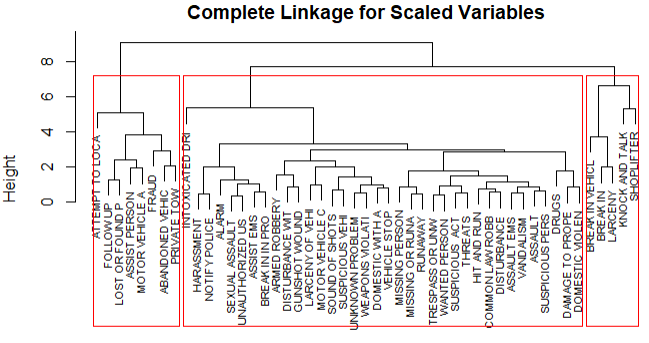
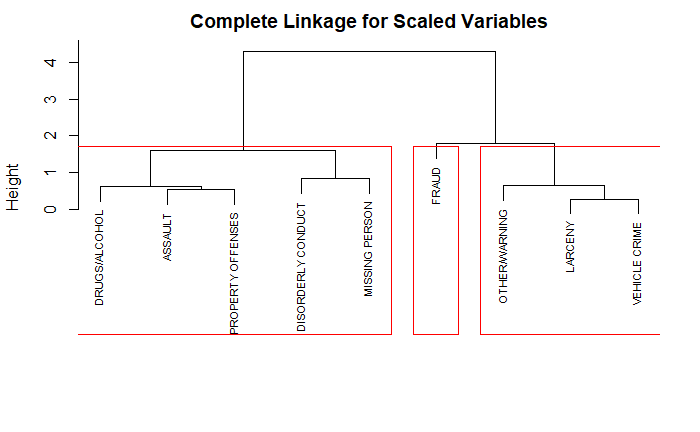


Results of C5.0

In C5.0, as in previous tree models, there is a high number of larceny charges that were misclassified as property offenses, and vice versa. There were also zero classifications of disorderly conducts or missing people. The misclassification error for C5.0 is 65%. In conclusion, classification trees were not ideal in classifying and predicting crime types.

*Clustering: K-Means and Hierarchical*

Lastly, we used unsupervised learning in order to learn potential new patterns that are not obviously visible to the eye. We used two clustering methods: K-means and hierarchical.

**Left**: Dendrogram for hierarchical clustering using complete linkage of “reported as” data.

**Right**: Dendrogram for hierarchical clustering using complete linkage of “charged as” data.

The two methods were used for both “reported as” and “charged as” categories. For hierarchical clustering, we used complete linkage because the observations are spherical. One interesting finding is that the classifications for both clustering methods were very similar for both “charged as” and “reported as”. For “reported as”, K-means and hierarchical had 98% similarity (48/49) in classifications. The only difference was that “knock and talk” moved from the smallest category to the largest in K-means. For “charged as”, the classifications were exactly the same for both K-means and hierarchical.

*Conclusion*

Many of the test errors in our regression and classification methods were quite high. One of the major limiting factors could be that there were not enough predictor variables in the dataset. Most of the predictors had to do with only location and time. For example, we had latitude, longitude, district, zone, and major areas as predictor variables. However, these can all be written as a linear combination of each other. In other words, latitude and longitude alone can give you the district, zone, and major areas. Some variables that may have been useful but were not included are the criminals age, gender, and the income levels of the area.

We can, however, draw a few conclusions from our analysis. Location is not a very good predictor of type of crime. More specifically, crime categories are evenly spread out through Durham. A majority of the crimes can be classified into either larceny, fraud, or property offenses. Almost every type of crime occurs more frequently during the day than at night, except for vehicle crimes, which happen most at night. The method that resulted in the most useful information was QDA to predict how often we predict a crime to happen at night actually happens at night.