Jiayi Zhou

Professor Nicole Dalzell

STA 381

9 December 2022

Write Up on Goal Two Decision Tree

**Section1: Goal**

This research project studies the variations in criminal trial rates on jury trials in felony cases for counties in North Carolina. There are two stages of the project. The second stage is fitting models with variables to explain what causes the variations in criminal trial rates among counties in North Carolina. This write-up is building decision tree models with all the explanatory variables and response variables to explain the main components that lead to the variations.

**Section 2: Data**

In this task, explanatory variables included *year, county, acquittal rates, dismissal rates, prosecutorial district, judicial district, the volume of cases in the county, types of cases in the local docket, delivery system for criminal defense, violent crime rates, property crime rate, the total county population, the percentage of the total population that is Black,* *the percentage of the total population that is Hispanic ethnicity,* and *political views of the county*. Two response variables are Y1–*the percentage of trials as a percentage of all convictions obtained* and Y2–*the percentage of trials as a percentage of total case disposal*.

For building the decision tree model for Y1, we used *Combine2.* It is the final combined dataset with all explanatory variables and Y1, with 909 observations and 15 variables. For building the decision tree model for Y2, we used *Combine3.* It is the final combined dataset with all explanatory variables and Y2, with 909 observations and 15 variables. Both of them are located under the folder Data Sets–Goal2VariableOrganizaiton–FinalDatasetsForModeling.

**Section 3: Data Cleaning, Processing, and Creation**

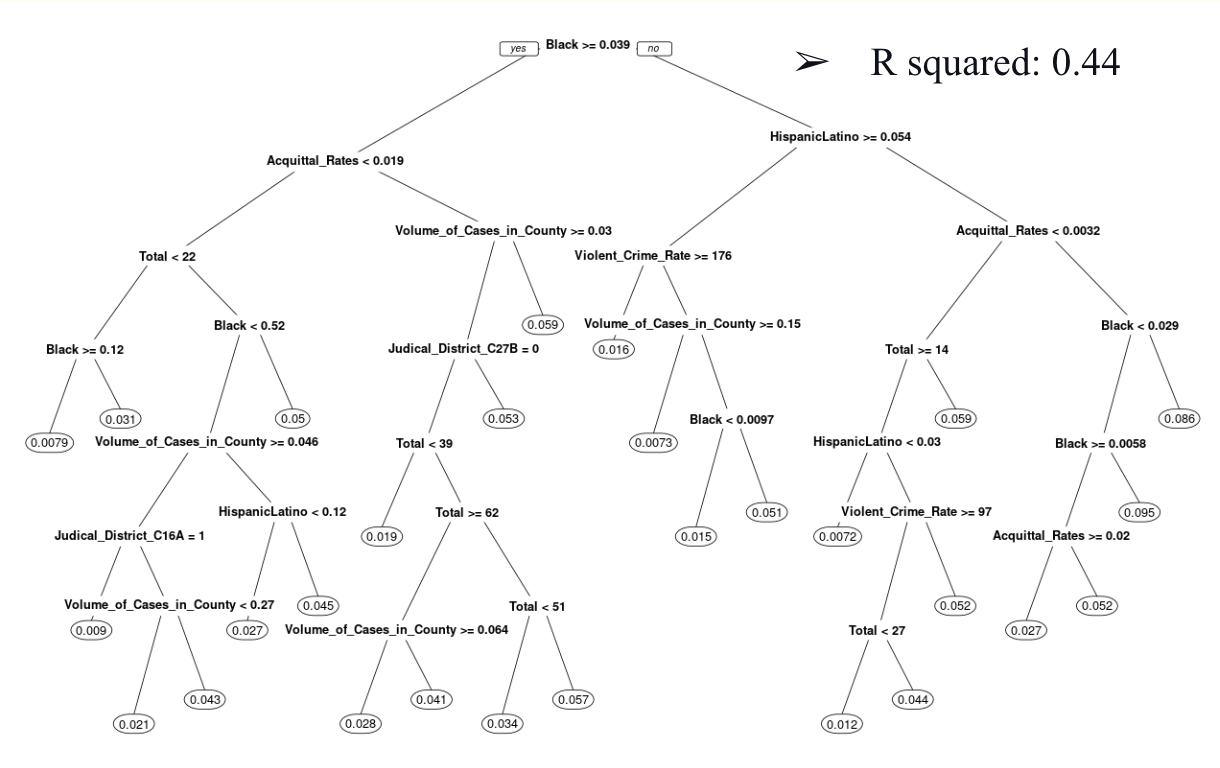
*Combine2* and *Combine3* already changed the column names and filled in missing values. They also changed variables *violent crime rates, property crime rate,* and *delivery system for criminal defense* to numeric types*.* Variables *prosecutorial district* and *judicial district* are changed to categorical types. The correlation between *Total, HispanicLatino,* and *Black* is breaked by dividing the *Total* by 1000 to make the numeric value smaller, *HispanicLatino* by *Total* to get *the percentage of the total population that is Hispanic ethnicity*, and *Black* by *Total* to get *the percentage of the total population that is Black.* As a result, *Combine2* and *Combine3* only need minor changes when they are loaded into Goal2DecisonTree.Rmd for building decision tree models.

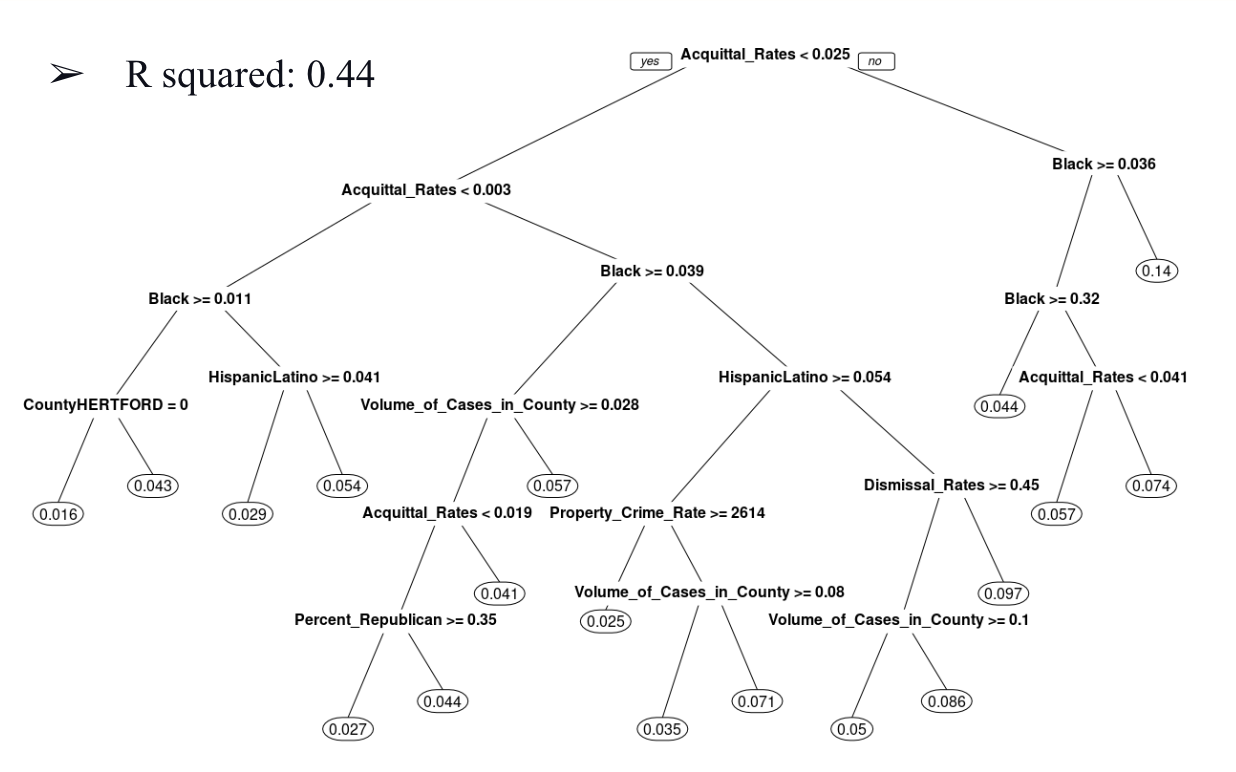
Because the decision tree model works well with both data types–numerical and categorical, thus I didn’t change most variables. With variable *year*, it originally is a numeric type, which assumes a more significant number in year weighted more. So I changed it to factor type to categorize the year. When our teammate Xinyuan Cao fitted her models, she discovered that *county* and *prosecutorial districts* are highly correlated, with *county* covering more information. Thus in our model, we removed variable *prosecutorial district* and focused on *county.* As a result, in both *Combine2* and *Combine3*, we have a total of 909 observations, 14 explanatory variables and one response variable.

**Section 4: Modeling**

In this report, we use the decision tree model for building models for response variables Y1 and Y2. A decision tree is a tree-like model of decisions and potential outcomes by splitting the dataset based on criteria. We decided to use the decision tree model because it works well with mixed types of variables and is relatively easy to build. The decision tree model is created by recursively evaluating different explanatory variables and using the one at each node that best splits the data and maximizes the R squared. It also provides visual results that are easy to explain and predict the response variables. To read a tree, we start from the root node and go to the next node based on the options provided by the edges.

For response variable Y1, I used the *Combine2* dataset to build the tree. The library we used is rpart. I created a model matrix with the response and explanatory variables. Then I combined the matrix and the response variable and created a data frame. The tree is drawn based on the new data frame, and I changed the value of cp in rpart command to maximize the R squared and keep the tree readable. We also calculated the R squared for the tree. The final output of the tree for response variable Y1 is shown below. The value of cp can be smaller to achieve a higher R squared, but the tree will also get more complicated. As we can see from the tree, *the percentage of the total population that is Black,* *the percentage of the total population that is Hispanic ethnicity,* and *acquittal rates* are three key factors that impact the response variable Y1–*tthe percentage of trials as a percentage of all convictions obtained.*



For response variable Y2, we used the *Combine3* dataset to build the tree. The same steps are performed as building the decision tree model for response variable Y1, except for the data frame change. The final output of the tree for response variable Y2 is shown below. The value of cp can be smaller to achieve a higher R squared, but the tree will also get more complicated. As we can see from the tree, *the percentage of the total population that is Black* and *acquittal rates* are two key factors that impact the response variable Y2–*the percentage of trials as a percentage of total case disposal.*

**Section 5: Conclusions**

Based on the result of both trees, we discover that *the percentage of the Black population and acquittal rates* significantly affect the variations in criminal trial rates on jury trials in felony cases for counties in North Carolina. However, the decision tree models don't provide how much two variables affect the variables. Hence, our project builds linear and mixed-effect models to explain the weight of two explanatory variables. Also, if we want to improve the R squared, the tree can get complicated, and people might need help understanding. So we must find a balance between the value of the R squared and the complicacy of trees.

**Section 6: Next Steps**

For the explanatory variable– *percentage of the total population that is Black,* we only had two years of census data and calculated the rest of the data based on the assumption the population change rate is static across ten years. This assumption is unlikely to be true in real life. So if more analysis is performed, it would be great if we could go back to find more data on the variable *percentage of the total population that is Black* and perform the analysis again to test and verify our results. Also, in building the decision trees, except for rpart package, there are more packages to choose from, such as party. Different packages can provide different visualizations of trees that build prettier trees and help others to understand the tree better.