*Creates a categorical decision tree and analyzes the relationships between variables of interest*

**Assignment**

**3**

A3

ALY6040 Data Mining Applications

Assignment 3 – Decision Trees

**PREPERATION:**

By: John DiSessa

For: Professor Ellis

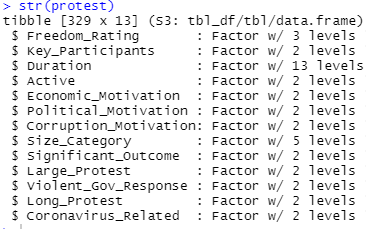
On: June 12th, 2022

Introduction

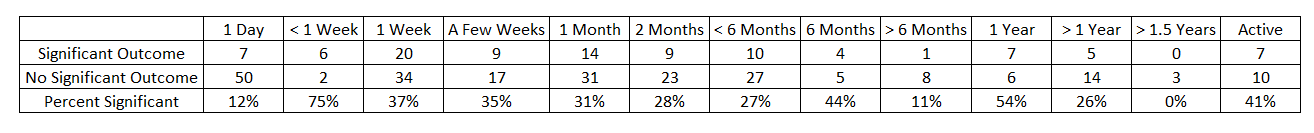
Matthew and I chose to analyze a Global Protest dataset from Kaggle. With about 330 protests and over 13 variables for each protest to analyze, we each conducted various analyses using different methods we have learned from class. For my analysis, I used a Decision Tree to see which variables were able to predict if a protest would result in a significant outcome.

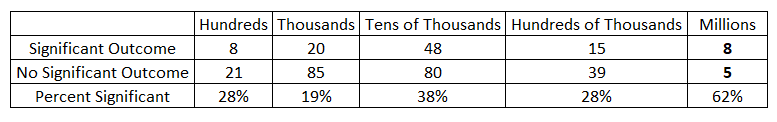
Dataset Exploration

Despite having an adequate sample size, the dataset had many inconsistencies. I changed variables to be Yes or No instead X or n/a. I changed certain numerical variables to categorical variables. I changed categorical variables to numerical. I fixed various typos in the spellings of categories. I grouped different classes within a variable together to reduce the small-sample class. Once I imported the dataset into R, I coded the variables as factors to ensure the decision tree was accurate and that the dataset could be used to conduct future analysis.



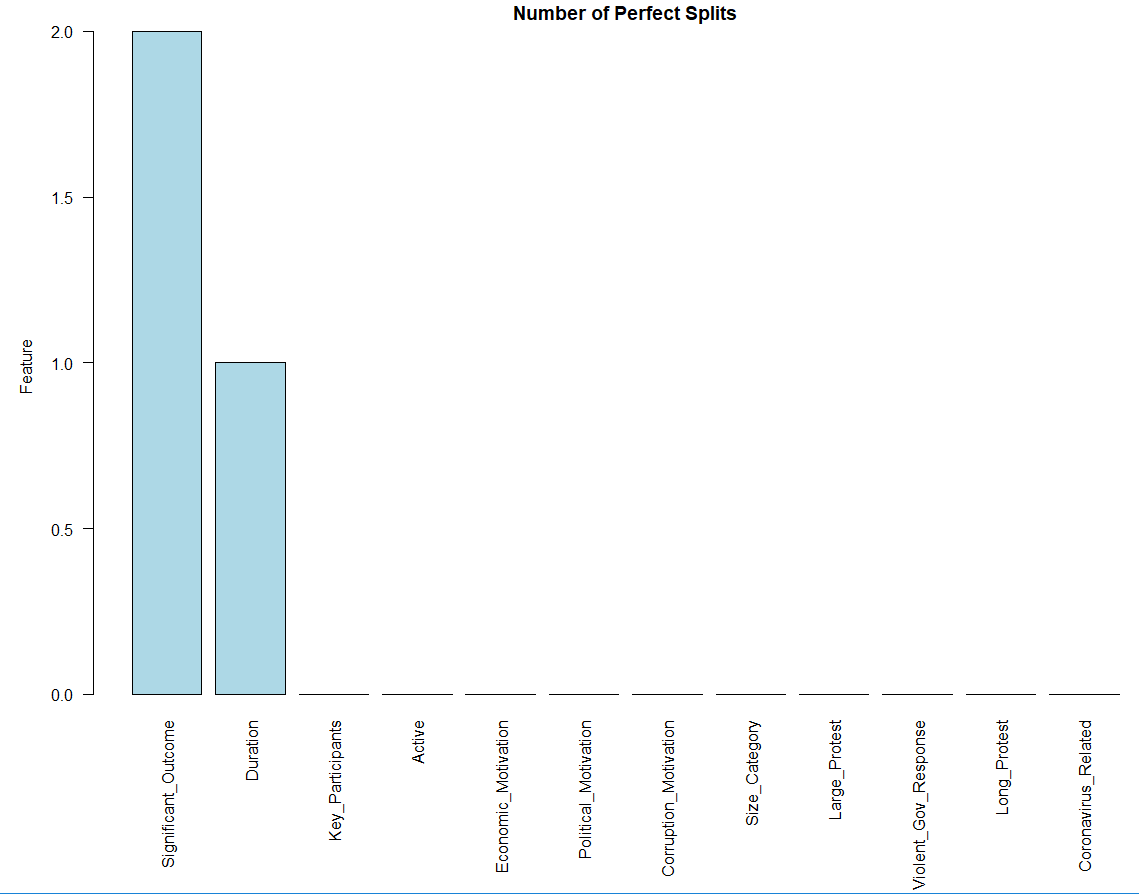
Before creating the decision tree, I thought it would be helpful to analyze the relationship between the duration of a protest and the peak size of the protest. I thought that as the protests lasted longer, they would build up more momentum and inspire more people to join the protest. These two variables had a correlation coefficient of .33. This seems like a moderate correlation so there is potentially a positive relationship although, without further analysis, no determination can be made yet. I then thought that, as the protests lasted longer, they would be more likely to induce government changes. I looked to see if there was a relationship between the protest duration and if there was a significant outcome. In the first table, the protest success rate was all over the place. This could be due to small sample sizes, but likely there are other factors affecting a protest outcome. I also looked to see if there was a relationship between the size of the protest and if there was a significant outcome. In the second table, we only saw a noticeable increase in success rate once a protest reached a million people. This could be caused by small sample sizes again, but it is at least worth looking into.





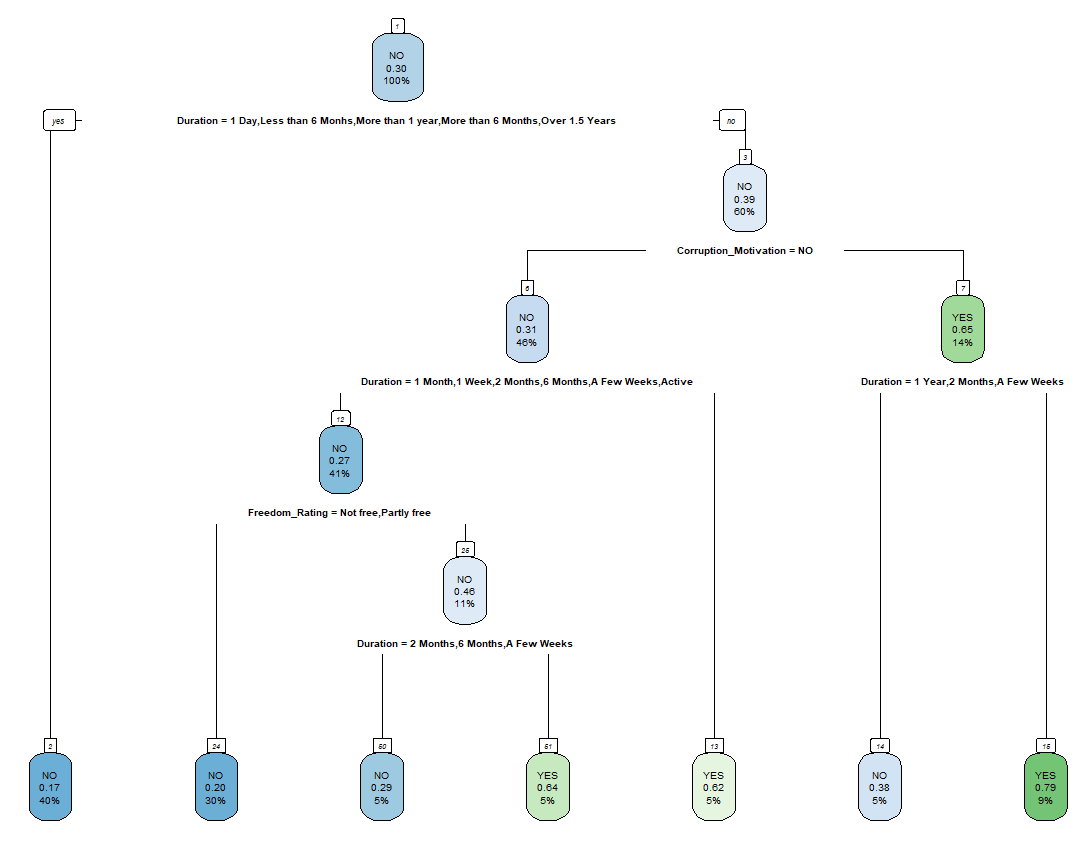
Decision Tree Setup

Even though I decided to make the Significant Outcome variable the target variable for this preliminary analysis, I still checked to see how many perfect splits we had with all our variables. Unfortunately, only 2 variables had perfect splits and both were due to very small sample sizes. Despite having 330 rows of data, when I categorized each one, the sample sizes dwindled for each category. We will need to rely on the decision tree calculations in order to predict if a protest will have a significant outcome or not.



Decision Tree Creation

I started creating the decision tree by splitting the data into training and testing datasets. 80% of our data was randomly selected for the training dataset and the other 20% was assigned to the testing dataset. The training dataset allowed me to create the following decision tree and the testing dataset will allow us to evaluate the accuracy of our dataset if we decide to proceed with this method for our final project. I also assigned an equal penalty matrix between type 1 and type 2 errors. I could not think of a obvious justification for assigning a stronger penalty for false positives or false negatives. Predicting a protest will have a significant outcome but not actually having the outcome would lead to false hope, but predicting a protest wouldn’t have a significant outcome when it actually would could just result in no protest at all. I played with the penalty matrix by changing the various penalties and each scenario resulted in a new decision tree. These would be worth exploring for future analysis, but for now, an equal penalty matrix was used to create this decision tree.



Despite having insignificant preliminary results, the protest duration was used for 4 of the 6 nodes in our decision tree. Future analysis will determine if this is due to small sample sizes, chance, or if there is actually a significant relationship here. It was also interesting to see Corruption Motivation as a node since I did not expect that to be any more or less influential than our economic or political motivation variables. The 6th node looked at the Freedom Rating of the people in the country where the protests took place. I am not surprised to see this in the decision tree since I believe freedom is a human right and is one cause that most people would agree is worth protesting for. On the other hand, there could be more protests in countries that are free due to the lack of fear of retaliation compared to some dictatorships. I would have expected Freedom Rating to be a core determinant of protests.

Summary

My report shows a clear desire for further analysis with so many questions stemming from the decision tree. We could also consider changing the target variable as well. There could be many reasons why protests have significant outcomes, why they attract many people, why they last a long or short time, how freedom affects protests, or other significant relationships that we haven’t even considered. Since many variables are categorical, our first decision tree was able to predict the protest outcome based on duration, freedom, and corruption motivation. Our dataset can be used in many different ways that can answer many different questions over the development of our final project.

Citations

jhuno137. “Classification Tree Using RPART (100% Accuracy).” *Kaggle*, Kaggle, 11 Jan. 2017, https://www.kaggle.com/code/jhuno137/classification-tree-using-rpart-100-accuracy/report.

Kuhn, Max. “The Caret Package.” *17 Measuring Performance*, 27 Mar. 2019, https://topepo.github.io/caret/measuring-performance.html.

Martin, Wendy. “Cohen's Kappa in R: Best Reference.” *Datanovia*, 7 Nov. 2019, https://www.datanovia.com/en/lessons/cohens-kappa-in-r-for-two-categorical-variables/#:~:text=kappa%20can%20range%20form%20%2D1,rater%20agreement%20exceeds%20chance%20agreement.