*Analyzes over 300 global protests based on their duration, impact, country, and other variables.*

**Final Project**

FP

ALY6040 Data Mining Applications

Final Project – Global Protests

**PREPERATION:**

By: Matthew Powell & John DiSessa

For: Professor Ellis

On: July 2nd, 2022

Introduction

One of the most basic tenants of freedom is a citizen’s right to protest their government. Without this right, corruption can grow and power can solidify, leaving marginalized citizens trapped in economical and societal prisons. For this project we have decided to analyze a list of 329 major protests from all over the world to understand the 20 mechanics behind this phenomenon. Each protest has information like, motivations, triggers, and outcomes. The dataset also has more quantifiable features such as the size, dates, and durations. The first step to analyze the data was to examine a subset in Excel. Using Excel to examine the data was extremely helpful for intuitive exploration and data quality checks. Fortunately, there were not any duplicated data and the data was not missing many values. After the data was moved to R the more detailed exploratory data analysis began.

Exploratory Data Analysis

The dataset contains tons of high-quality information and an adequate sample size. But, in order to use that information, the data must first be preprocessed. When the data was initially downloaded all the fields were stored as text values. Nearly every column had to be adjusted before the data could be used. First, we converted all dates stored as text into a Datetime data type. The dataset initially only stored the data in month and year, so we assumed every protest started on the 15th of the month. Next, we wanted to look at the binary columns. The dataset stored the binary columns with an X signifying true. To correct this, we wrote a quick function to scan every column and convert those that are binary into a logical form column. Logical form columns in R can be treated the same as a 1:0 storage allowing things like averages and sums to give us counts and allocation. Next, we went back to Excel and converted all the non-numerical values for the “Peak Size” and “Duration” columns. For duration to maintain a consistent metric, we converted all values to days assuming 30 days/month and 360 days/year. While this is not 100% accurate given the data does not have any significant digits, the data quality was not largely changed. For the Peak size columns, we took the value given for higher or lower (e.g. “>4000” became “4000”). These fields were then converted to numerical for further analysis. Finally, we converted the categorical variables like “Freedom Score” to ordered factors. The data stored each countries freedom rating as either “Free”, “Partly Free” and “Not Free”. This seems to be one of the most interesting features because we can tell the political climate of that country. One would assume countries with higher freedom are more likely to have a higher protest success rate.

Chart, histogram

Description automatically generatedWe first examined the countries where these protests took place. Most of the countries only had a single protest, but larger countries like France, India, and the USA had more total protest. Next, we wanted to examine the binary columns we cleaned up before. We made a column chart and correlation plot so we could examine the rows on their own and as a group. Looking at the columns on their own we can see most of the protests in the dataset stemmed from a political motivation. Economic motivations were the second most prevalent. The full histogram can be seen below.

Chart

Description automatically generated with medium confidenceThe interesting thing here is looking at the correlation plot, we can see those two metrics are inversely correlated. This means most of the protests can be simplified down to politics and economics with politics being the strongest motivator. One of the more promising statistics from this portion of the analysis is the number of protests classified with a significant outcome. In the dataset over 30% of the protests achieved a significant outcome. We can also compare these values relative to the protest size and see that the larger the protest was, the more likely the protest was to have succeeded. This was expanded further in the plot below to show a timeline of the protest and their outcome.

The above plot contains a few insights beyond the timeline of protest. We can see on the right side of all three plots that there are not a lot of large or successful protests that achieved a significant outcome. We can see large protests in free countries like South Korea and England, but they drop off after 2020. This in no doubt is correlated to the COVID-19 pandemic but it still represents and interesting cultural trend. One of the largest protests with a significant outcome was the “Police Brutality Protest” in June of 2020. That protest is the large blue circle on the “Free” line and is the most recent large protest in the dataset.

Chart, scatter chart

Description automatically generatedWhile still examining the peak size we can look at how various countries let their people protest. To answer this question, we made a plot that showed the duration and size along the x and y axes with the color signifying how free that country is. In this plot we can clearly see in the top right corner (Large size and Long Lasting) there are mostly Free country datapoints along with a few Partly Free. This makes sense if we consider large totalitarian governments often limit free speech to control their populations. Other interesting insights from this graph are the number of smaller protests in “Not Free” countries that were able to go on for a long time. It is also noticeable that the “Partly Free” countries are very well dispersed over the plot of duration and size. The dataset contains around 110 protests from each freedom rating to provide an even sample. In this plot we can see that they are much more variable in their protest size and durations. This could be because these countries are in turning points where civil intervention is needed.

Continuing our exploration, we questioned how long is the average protest? With the converted duration data, we were able to create a density plot showing how long the average protest lasted in days. We can see the most common protest duration is 1 day. For larger issues the dataset even had protests going on for over 2 years. The final plot we created showed the duration and size of the protests given the freedom rating. While the duration plot did not yield any notable insights, the peak protest size did. Looking at Figure 6, we can see that the interquartile range is largely similar between the 3 different categories, but the mean shifts greatly. In the free countries, the mean peak size is roughly 1,000. For the partially free countries that mean was also their 75% mark. For the Not Free countries their mean lined up with their 25% level. So, while these ranges are not extremely varied there is a notable difference.

Chart, box and whisker chart

Description automatically generated

In this plot we can see the clear outlier for “Not Free” countries with the main one being a protest in Algeria about President Bouteflika seeking his fifth term, a highly unprecedented move. This action mobilized up to 1 million people and lasted a year. Through all this vocalization, the people were able to force President Bouteflika to resign and not seek an additional term. These are the insights that examining data like this can provide. We plan to learn what factors about protests make them more likely to be successful. Would it be better to have all your protestors show up on the first day and have a bunch of them burn out? Should they play the long game and take shifts? Once we mine the text data, we hope to understand what factors led people to believe their only recourse was to take directly to the streets. Martin Luther King Jr. said that a “riot is the language of the unheard people” and we believe a protest is as well. If we can understand what problems people are facing, the government can help those people before it escalates.

Decision Trees

We started creating the decision tree by splitting the data into training and testing datasets. Decision trees work by finding the split at each point that will provide the end prediction with the minimum loss. We can control how complicated the algorithm will build itself using the complexity parameter and simplify it with pruning. 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 us to create the decision tree (Figure 8) and the testing dataset allowed us to evaluate the accuracy of our dataset. We also assigned an equal penalty matrix between type 1 and type 2 errors. We could not think of an obvious justification for assigning a stronger penalty for false positives or false negatives. Predicting a protest having 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 could just result in no protest at all. We 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.

Diagram

Description automatically generatedThe protest duration was used for 4 of the 6 nodes in our decision tree. Future analysis could 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 we 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. We are not surprised to see this in the decision tree since 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. We would have expected Freedom Rating to be a core determinant of protests.

We also performed some additional normalization to the dataset to try and create a second decision tree. The first split for the algorithm comes from the start date. The dataset only goes back as far as 2017 so all the current protests are in the same relative time frame. We then looked at a plot (Figure 2) of the likelihood of protest success over time and can clearly see fewer protests in the more recent history. While this does not give us any additional information on how to set up our protests, it does inform us of the current global conditions. The next split examined the motivations and if they were driven by corruption. The table checks if the corruption motivation column is a 0 (False). In this case if corruption is a motivation, then we would inverse the logic since it is a double negative (The corruption motivation was no 0 (False)). Corruption seems to be a major point of success for the protestors. Protests that stemmed from corruption had a 44% success rate compared to the total average of 27% (Figure 3). After we split on corruption, we can classify most of the remaining data with the start date. We can also dig slightly deeper for the protests that do not relate to corruption. We can see the next split says that a protest longer than 405 days is less likely to have a successful outcome. This could be caused by 2 things, either the protest has lost steam over time, or that if problems are not fixed after 405 days it is unlikely they will ever have the intended outcome. Then, we split again along the freedom rating of the country with the “Free” countries succeeding with a longer duration. The less free countries success still depend on when they started. Given that the date made up 3 of the 7 splits we decided to rebuild the tree without that feature for additional insights. In the new decision tree (Figure 4), all the previous start date columns were replaced with duration and longer duration protest tendeds to be more successful. Overall, this decision tree taught us that it is getting harder and harder to have a successful protest and that unless it is something that most people can easily agree on, like government corruption the protest has a Diagram

Description automatically generatedmuch lower chance of success.

Clustering

The second method we used was t-SNE clustering. t-SNE stands for t-distributed stochastic neighbor embedding. This model works by decreasing the number of variables with embedding. Embedding is the process of examining the surrounding data points and minimizes the data into 2 coordinates. Looking at the output of t-SNE you can see the data dividing into a few key groups of roughly 4 clusters. There are some clusters with a higher average success. This was also extended to show those 4 specific clusters in the following appendix figure. The t-SNE by itself is less understandable than the decision tree, but in theory, has the most information extracted. This algorithm also works better with large sources of data so only having 329 rows limited its performance. Looking at the clusters after being merged back into the original data we can identify a few key characteristics. Of the 4 clusters, one has an average success of 40% which is 33% better than our 30% baseline. If we wanted to run a successful A picture containing chart

Description automatically generatedprotest, we should aim to behave similarly to that group (group 3). Group 3 had the longest average duration being almost double all the other clusters. While this is double all the other clusters it is still only an average duration of 9 days. This again goes to show that there may be some protests that will Chart, scatter chart

Description automatically generatedforever go unheard.

We can also look at cluster 1 which had the highest peak size but the second lowest success rate. This also supports the previous decision tree. As that algorithm identified corruption as the most influential factor, t-SNE group 3 had above average success and had the highest percentage of corruption-based protest.

Natural Language Processing

This dataset contains a few text columns indicating more qualitative things like motivations behind the protest and the participants involved. These text-based columns all contain comma separated terms that we can convert into a term document matrix. A term document matrix shows how often each term appears in each document. This can help extract the Text

Description automatically generatedkey values across large text fields. When we apply this analysis, a few interesting things appear. We summarized these findings with a word cloud visualization. The “motivators” column created the most interesting word cloud. The main reason behind protests in this dataset is corruption. While this is not a surprising result it is fascinating to see these trends we hear about actually appear inside of our dataset. Other interesting motivators are systemic racism, police brutality, economic downturn, and poverty. These causes of corruption can be seen globally and need to be addressed by governments all over the world.

Text

Description automatically generatedWe can make another word cloud using only the successful protests to see if there is anything in those protests that separates them from the rest. The successful protests only had three common terms: misuse of government funds, high cost of living, and rising authoritarianism. Misusing government funds makes sense as it is the most actionable problem where the government can simply change its spending. Other problems like systemic racism have complex solutions that are impossible to solve with one bill or one protest. It is interesting to see a term like rising authoritarianism in the successful motivators group as that is not a problem with an easy solution. Again, we feel that this dataset would greatly benefit from a “movement” column so that we can associate protest with certain continuities to examine their overall impacts.

Interpretations and Recommendations

This analysis has been extremely informative but is far from complete. For a topic as complex as government protests, we need to search under every rock for answers. The first thing we would like to add is an overall movement variable. We have seen large societal issues are not solved by a single protest. Often there are larger underlying movements that are slowly gaining ground until they gain critical mass. The most prominent example are the Black Lives Matter protests that have been going on across the world for over a decade. All of those individual protests may not have totally achieved their goal but maybe later protests are able to build momentum from previous protests, Another feature we would like to see is a more detailed description about the outcome of the protests. While there was a column describing the outcomes, the information was scarce. Over half the protest had reported to no large-scale change but the small changes can still add up. We would have liked to see some proxy metrics that can be used to judge the success of the protest on a more objective way. Metrics that show this could be the following change in government approval, or if we know their specific group name, we could search online platforms like twitter for the public sentiment. It would also be helpful to know the type of government that the people are protesting. We have rejoined the new cluster back into the original dataset to examine the results. The complete table can be found in the appendix.

The next steps for this analysis are to mine the text-based columns for additional insights and try to pull in additional sources. For the text mining we would recommend creating a dictionary for each variable listing all the words used and then use embedding on those words. A possible solution could be to use word2vec to turn each protest into their own datapoint that can be compared graphically based solely on the text. Since there are only 329 rows, we elected to avoid deep learning in favor or less data greedy algorithms.

Conclusion

This dataset is critical to understanding the current global climate and how to effectively manage protests for change. Unfortunately, this dataset shows that there is not a lot that you can do for certain issues. Protests work best when you are able to get the general population to side with you and are fighting against something that effects the masses like corruption. The main issue with this dataset is that it treats all the protests as independent actions. In the world of protests, there is rarely a massive scale change right away but by thinking of them as smaller battles in a much larger war, the data begins to make more sense. People will continue to advocate for their causes and hopefully as data becomes more available, people will be able to more effectively make actionable change.

References:

Death. (2010). Counter-conducts: A Foucauldian Analytics of Protest. Social Movement Studies, 9(3), 235–251. <https://doi.org/10.1080/14742837.2010.493655>

Genuer, & Poggi, J.-M. (2020). Random forests with R. Springer. <https://doi.org/10.1007/978-3-030-56485-8>

Pezzotti, Lelieveldt, B. P. F., Van Der Maaten, L., Hollt, T., Eisemann, E., & Vilanova, A. (2017). Approximated and User Steerable tSNE for Progressive Visual Analytics. IEEE Transactions on Visualization and Computer Graphics, 23(7), 1739–1752. https://doi.org/10.1109/TVCG.2016.2570755