Joseph Silagi
11/22/22

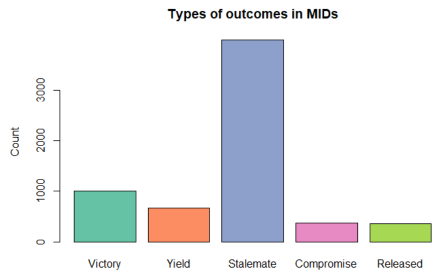
Armed Conflict Memo

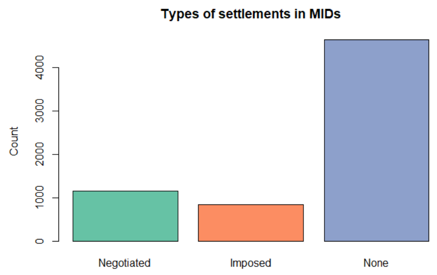
The question I seek to explore is: How accurately can we predict the outcome of a militarized interstate dispute, given its number of fatalities, its highest military confrontation, its highest level of hostility, whether it was a reciprocated dispute, whether it was a part of a war, its cumulative duration and the highest action taken by either side? Can we predict the settlement type of the dispute given these same attributes?

The data used comes from the Dyadic Militarized Interstate Disputes (MIDs) Dataset, Version 4.02 (Maoz and Miner, 2021), which is part of the larger Correlates of War (COW) project. The data can be found at: https://correlatesofwar.org/data-sets/mids/. The COW website defines MIDs as, "cases of conflict in which the threat, display or use of military force short of war by one member state is explicitly directed towards the government, official representatives, official forces, property, or territory of another state. Disputes are composed of incidents that range in intensity from threats to use force to actual combat short of war" (Jones et al. 1996: 163).

I treated the `outcome` variable and the `settlmnt` variable as the response variables, each of which being a function of the other variables present in the dataset. My first step was to trim down the number of predictors to be used in my analysis. I determined that the most likely variables to affect these response variables were the number of fatalities (`fatlev`), the highest military confrontation action (`highact`), the highest level of hostility (`hihost`), whether it was a reciprocated dispute (`recip`), whether it was part of a war (`war`), the cumulative duration (`cumdurat`) and the highest action taken by the dyad (`mid5hiact`). I also removed missing values and other noninformative response indicators (ex: an ongoing dispute).

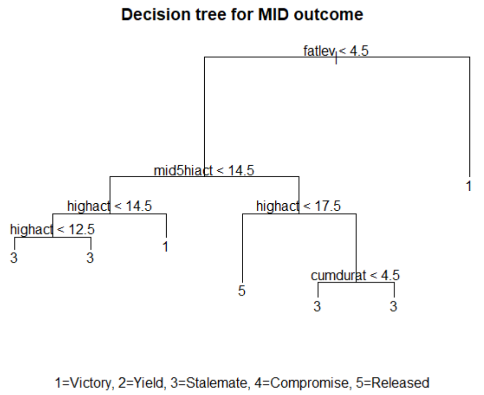
I plotted the different types of outcomes and settlements to get an idea of the distribution of data:



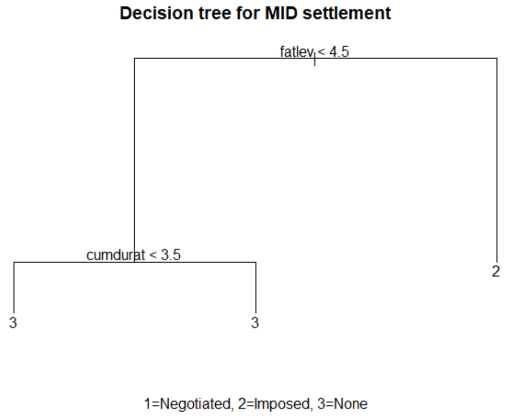


Stalemate is the most common outcome type (3993 occurrences), while having no settlement is the most common settlement type (4639 occurrences).

As all the variables in this analysis are categorical, I decided to use a decision tree to classify the observations by their outcome/settlement type. A decision tree sequentially splits the data into smaller and smaller subsets, with each split looking at a certain value of one of the predictor variables. I split the data into training and testing sets, and trained one decision tree with outcome as the response, and one decision tree with settlement type as the response. Each tree used the aforementioned seven predictor variables. The trees are shown below:


Decision tree for MID outcome

1=Victory, 2=Yield, 3=Stalemate, 4=Compromise, 5=Released

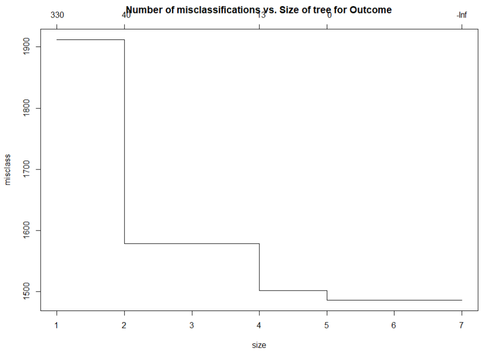

Decision tree for MID settlement

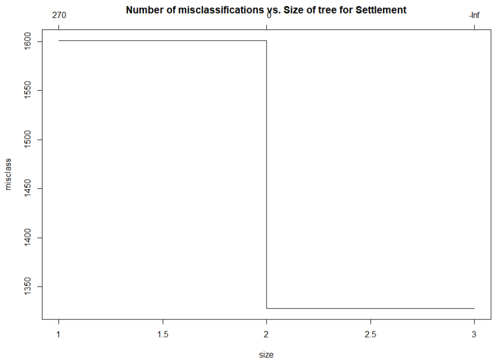
1=Negotiated, 2=Imposed, 3=None

In the outcome decision tree, there are only three possible values on the leaf nodes, meaning that the tree is guaranteed to never predict a yield or a compromise outcome. A similar situation is seen in the settlement type tree, as the indicator for a negotiated settlement isn't even an option in the leaves. This is likely due to the low occurrence of these outcomes and settlement type in the dataset. The trees still achieve moderately successful accuracy rates when used to make predictions on the test data: 0.696 for the outcome, and 0.758 for settlement type.

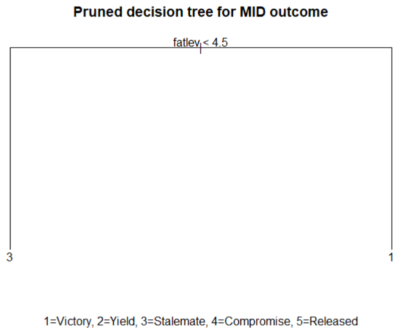
It's also interesting to see which of the predictor variables are most important in predicting the outcome or settlement type. In both models, the number of fatalities is the uppermost node in the tree. If the `fatlev` variable has a value of 4.5 or above (meaning there are more than 500 fatalities), then the tree predicts the outcome will be a victory for one side, and the settlement type will be imposed. Additionally, the cumulative duration doesn't seem to have a large effect on the outcome or the settlement type – in both trees the `cumdurat` variable is used on a node, but regardless of the value it takes it results in the same leaf value for outcome and settlement. We note that the highest action variables (`highact` and `mid5hiact`) do have an effect when it comes to predicting outcome type.

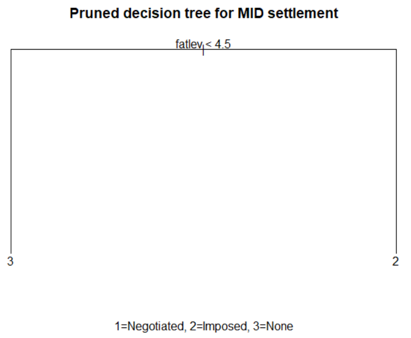
Finally, one concern when it comes to decision trees is overfitting to the training data. Although these trees are already quite simple as they are, we can reduce them even more via pruning. I used 5-fold cross-validation to iteratively add leaves to the trees and then record the misclassification rate. Plots of this process for both trees are shown below:





For both trees, it appears that using just two leaves could be an appropriate way to further cut down on overfitting, as this is where we see a steep drop-off in the misclassification rate. I re-fit both of the trees with two leaves to see what the results would look like:


Pruned decision tree for MID outcome
fatlev < 4.5
3
1
1=Victory, 2=Yield, 3=Stalemate, 4=Compromise, 5=Released


Pruned decision tree for MID settlement
fatlev < 4.5
3
2
1=Negotiated, 2=Imposed, 3=None

Clearly, these trees are extreme oversimplifications of the relationship between the predictor and response variables. The outcome tree will take an observation, look at its fatality level, and label the observation as stalemate if fatalities are ≤500 or victory if fatalities are >500. The settlement tree will label the observation as none or as imposed based on the same criterion Yet, the accuracy rates are surprisingly not horrible: 0.671 on the outcome tree, and 0.758 for the settlement tree. Note that this is the same accuracy rate that the first settlement tree had, as that tree didn't alter the outcome level of the settlement based on cumulative duration (even though there was a leaf for `cumdurat`).