Predictive Analytics: Analyzing Complex Threats for Operations and Readiness (ACTOR)

Courtney Beckham

Honors College, College of Charleston, Charleston, SC

SMART
SCIENCE, MATHEMATICS & RESEARCH FOR TRANSFORMATION
PART OF THE NATIONAL DEFENSE EDUCATION PROGRAM

Abstract

This project explores the accuracy of the Analyzing Complex Threats and Operations Readiness (ACTOR) model employed by the United States government and theoretical approaches to surmount the challenges quantitative predictive models pose when used to predict qualitative information. The ACTOR model, also known as a "second image" model, views conflicts as arising from "factors and failings internal to states." In accordance with this belief, the ACTOR model aggregates data on 12 macro-structural factors from 159 countries over the period 1975-1999 to predict the level of intensity a country will experience instability. This paper will explore the accuracy of such findings as well as the future directions and developing technologies that can be applied in accordance with later projects employed by the United States such as Integrated Crisis Early Warning Systems (ICEWS).

ACTOR Model

The ACTOR model aggregated data on 12 macro-structural factors for 159 over 1975-1999. These macro-structural factors are outlined in Table 1.

Table 1. The table below details the 12 macro-structural factors used by the ACTOR model to predict country instability.

Macro-Structural Factors	
Percent of history spent in conflict	Infant Mortality Rate
Trade Openness	Youth Bulge
Civil Liberties Index	Life Expectancy
Political Rights Index	Democracy
Religious Diversity	Caloric Intake
GDP per Capita	Ethnic Diversity

The ACTOR model serves as a "second image" model which views country instability as arising from the internal factors of a country. This method is illustrated in Figure 1.

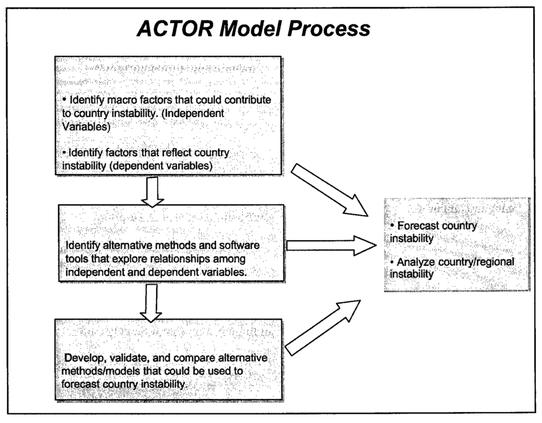

Figure 1. The figure above illustrates the methodology behind the ACTOR model.

Theoretical Extensions and Updates to ACTOR

Decision Tree Modeling

- Decision tree modeling is a different way of modeling the ACTOR issue, and will lead to easier interpretability by analysts

Nearest Neighbor Outlier Detection

- This method centers on the assumption that "instances of normal data occur in dense neighborhoods, while outliers occur far away from their closest neighbors."

Multinomial Logistic Regression

- Finds the best model that accurately describes the relationship between the dependent variable (i.e. level of stability) and the independent variables (i.e. the 12 macro-structural factors).

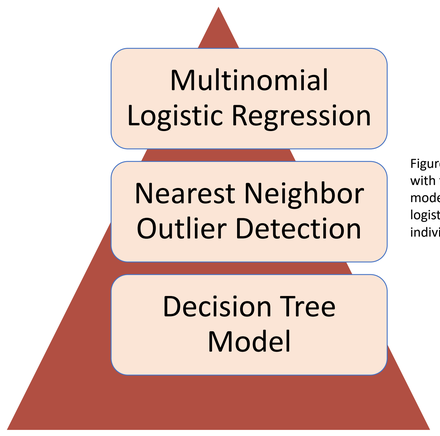

Figure 2. This figure illustrates the scope of each model, with the decision tree model taking a holistic approach to modeling the ACTOR data, all the way up to multinomial logistic regression providing a very specific look at the individual data points.

Decision Tree Model

A decision tree is a machine learning algorithm that splits the data into subsets. In approaching the large dataset that the ACTOR model is working with, finding patterns in the data is imperative to proceeding with the model. To begin, the subsets of the data must be determined. Specific variables will be split at a location and two determinations will be formed: the predictor variable used for the split, and the set of values for the predictor variable. To aid in possible partitioning values, an equation for information gain is used. The higher the value of the information gain variable, the better a split. The information gain is determined by the following equation

$$Info = -\sum \left(\frac{N_j(t)}{N(t)}\right) log_2 \left(\frac{N_j(t)}{N(t)}\right)$$

Due to this model being a classification model, using the Gini Index value (also known as the Gini impurity value) will help determine where to split variables. The formulas for these variables are displayed below.

$$impurity = 1 - \sum \|p(j)N_j(t)/N_j\|^2$$

$$Gini = impurity(Parent) - \sum (p_k) impurity(Child_k$$

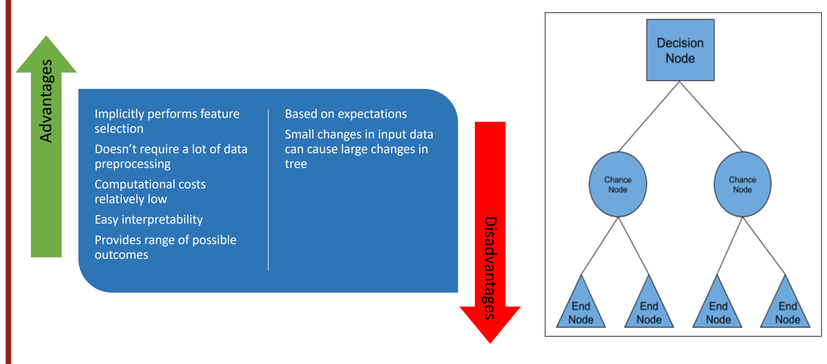

Figure 3. The left figure illustrates the advantages and disadvantages in using the decision tree model. The diagram on the right represents the standard structure of a decision tree model.

Nearest Neighbor Outlier Detection

It may suffice for the Army to focus on the major outliers in the dataset: those who are highly unstable. Instead of using this to replace the ACTOR model, this may present a more easily interpretable and communicable output that includes the countries experiencing the extremes of instability. This technique would be used best after the initial predictions for the ACTOR model are made.

The nearest neighbor based outlier detection technique requires finding which instances are most dissimilar. For the continuous variables that are used in the ACTOR model, Euclidean distance would be used in measuring dissimilarity. The Euclidean distance is calculated using the following formula $d = \sqrt{\sum_{i=1}^{v} (p_{1i} - p_{2i})^2}$ where p_{1i} and p_{2i} are two data instances that are summed for v variables. For the categorical attributes used in the ACTOR model, a simple matching coefficient would be used to measure dissimilarity. Lastly, an outlier score is computed for each country.

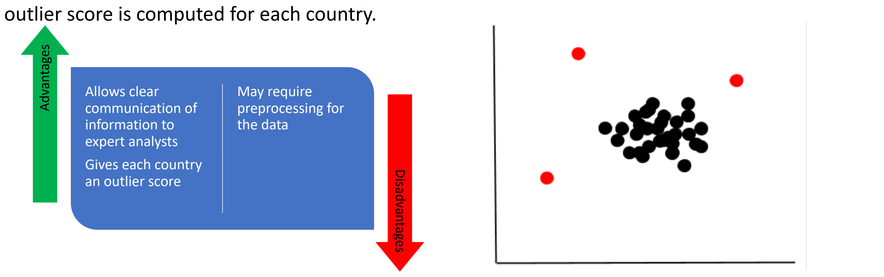

Figure 4. The image on the left outlines the advantages and disadvantages of a Nearest Neighbor Outlier Detection model. The image on the right serves as a simple example of what this model would look like, with the red dots serving as the major outliers in the dataset.

Multinomial Logistic Regression

Multinomial logistic regression uses the following regression equation $logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_kX_k$ where b_0, b_1, b_2,..., b_k are the regression coefficients of the independent variables. If the regression coefficient is not significantly different from 0 (P>0.05), then the variable can be removed from the regression model. However, if P<0.05, then the variable is significant in the prediction of the outcome variable. The logistic regression coefficients illustrate the change in the predicted log odds of having the outcome of interest for a one-unit change in the independent variables. If b_k is greater than 0, then the odds are higher, and similarly if b_k is less than 0, then the odds are lower.

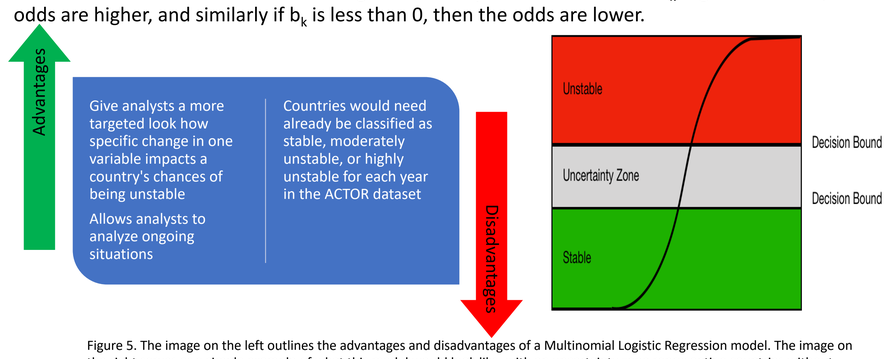

Figure 5. The image on the left outlines the advantages and disadvantages of a Multinomial Logistic Regression model. The image on the right serves as a simple example of what this model would look like, with an uncertainty zone representing countries without enough information to be classified as highly unstable or highly stable.

Future Work

This paper serves largely theoretical approaches to later be hard-coded and used in conjunction with the updated ACTOR model known as the Integrated Crisis Early Warning System (ICEWS). The dual approaches include:

- Creating a decision tree model that uses the 159-country dataset to compile reports that would be useful to analysts.
- Designing a web applet that allows the user to select a variable and a country and outputs how a shift in the variable affects the country's instability.

Conclusions

The ACTOR model is not meant to be an all-encompassing, stand alone approach to predicting country instability. This methodology is to be used in conjunction with expert analysts as well as a myriad of other predictive modeling techniques and information retrieval. The changes that could be made to the ACTOR model, and to its descendant, the ICEWS (Integrated Crisis Early Warning System) rely first and foremost on the question that the government would like for ACTOR to answer. If the government needs accuracy on just a few important factors, the accuracy of the model could increase drastically. However, if the government desires a more holistic view of the individual country, the ACTOR model provides an excellent picture of 159 countries that allow policymakers and expert analysts to help inform them in important decisions for the protection and furtherance of prosperity for the United States government.

Acknowledgements: I would like to express my gratitude to Professor Lancie Affonso for his continued guidance on this project. I would also like to thank the SMART Scholarship program for providing the funds necessary for this project.