1 Abstract

The Minorities at Risk (MAR) is a long running project to moniter an anyalze the situation of minorities around the world. While flawed due to only considering minorities that the are aleready at some sort of political, social or physical risk, it has been used in a number of cited papers to anyalize using primarily linear regression to determine what key risk factors are. However, fitting into a broader trend in socieal sciences these investigations have been broadly anyalitical istead of predictive.The purpose of this project is to reverse that trend, and instead of simply identifying key factors, try to predict how changes in minority group status will affect a group's security

2 Literature Review

The research for this paper focused on two distinct points: the broader history of the use of advanced statistical anylaisis and Machine Learning within the social sciences, and then an investigation of other research on the MAR project. These allow the reearch conducted within this project to be placed within a proper context, and the significane of the result truly comprehended.

2.1 A History of Machine Learning in the Social Sciences

Machine learning over the course of the last 15 years has revolutionized the computring world, making problems that were once conisdered near impossible easy, and opening up eniterly new areas of insight. Unfortunatly as it currently stands much of this same process has not yet fuly transfered over to political science, and in the few cases that it has been used in the political science realms it has been often misused (Grimmer 3). Fundementally the problem is that political scientist currently use the techniques of machine learning in similar ways to existing statistical models, to measure latent tendencies and causality, instead of what Machine Learning has been used for in computer science: prediction and classification. A prime example of this it the largest application currently in place whic attempts NORMAL which simply attempts to dtermine based off of role call voting the political alligience of members of congress. It does even attemp what I am trying to do, based off of a set of voting behavior predict how congress, or indivudual congress members will vo on specific peaice of legislation(Grimmer 1).

Where Machine Learning however has had a significant impact on Political science then is in sentiminate anylasis, attempting to determine the emotion associated with the peice of writing. In an era of decentralized news, and increasing politcal poizaeration then these techniques are crucial to achieving a firm grasp on the plethora of online communities that people express them in. Sentimint anylasis technology has already been adopted by major companies for this purpose, and increasingly social scientist are as well(Pang and Lee 3).

The reason that political scientist have been so unwilling to fully engage in the type of predictive reaserarh I have been interested in is two fold: there is a lack of good data avaiable, and they have rpeatedly been wrong. In fact one major survey of results found that, "Chimps randomly throwing darts at the possible outcomes would have done almost as well as the experts."(Stevens 2). But in many ways this hinges on the first observation , dat well suited to the type of algorithms favored by machine learning are hard to find in the political science world. Most intresting features of political sciences: trends, elections, and conflicts are slow moving and infreuent leading to a dearth of data.

2.2 Reseach on the MAR Project

Coming soon...

3 Methodology

Approaching this problem there were two main decisions that had to be made: how to prune the data, and how to find sets that were particularly illuminating as the complete size of the data was far too large for accurate classification to occur. Beyond that my primary challenges were as expected simply attempting to find the correct level of sampling, learning rate and epochs that would maximize the accuracy of the results. The exact methodologies are explained separately.

3.1 Modeling

In many ways this was an authentic experiment of how machine learning could be used to introduce predictive machine learning into the filed of Political Science. As dicussed priviously there has been a distinct absence in the use of these technologies within the field. Therefore I was largely in ncharted territory, and was simplay attempting to see if I could get reasonable results, and using which methods, not necessarily attempting to optimize it using any one set of data. Therefore I pretty much used "off the shelf" technology, whose creaters are credited in the sources section of this paper, to conduct my anylasis. As will b discussed in the results section some of these technologies ended up being very successful and others did not. The three that I investigated however were a simply Basyian classifier, a basic Percepton, and a slightly more complicated multi-layer neural network. I chose these methods as they have been used onsimilar data sets, in other fields to reasonable levels of success. In particuar, I must complemet all of the work that went into the PIMA data set, as I used many of the techniques devekloped there to apporach the MAR data set

3.2 Data Grooming

One of the main challenges with doing a proper statistical anylaisis on the project was that it was not entirely set up for statistical anylaisis of the type that I wanted to conduct. As it origionally stood each of the data entries was individually labeled with over 40 diffrent variables. As we only had Slightly over 850 different cases, it was clear that we would be forced to prune the data to alow it to be more accuratly surveyed. The exact specifics, amd justification of why each group was excluded is explained below, but the general approach was to eliminate anything where th data was woefully incomplete, or severed as a unique identifir such as name. This type of variable is innefective as it is far too specific as identifies just one element, making the predictive powers employed here far too weak. The other main change that I made was to discretize some of the data, or convert it from text based catagories to numerical ones that the computer could full process. In the chart below I refer to each of the variables that were changed or elimiated by variable name as defined in the code book for the MAR project. The data is broken up into tow sunbsections. Those that I eliminated from the project and those that I simpley changed. For each there is an explanation of why this was done, and for any changes an exact description of the changes.

3.2.1 Deletions

As explained above some of the data needed to be deleted. The following table offers an explanation behind each piece of Data that I decided to delete, but in gneneral what was deleted was either extraneous, a repeat, or unusable for the reason specified.

Table 1: A simple longtable example

Variable	**Reason For Deletion**
numcode	This is a unique identifier of each minority group that was included in the data set, with it being the same across years for each group. As we are trying to identify predictuve capabilities about future groups this clearly has no relavance and can be safely eliminated
VMARGROUP	VMARGROUP is simply the name of the group. Once again this is far to specific, and will not accurately create generizable trends. Therefore it is also eliminated
country	This is the actual name of the country, as there is also a numerical representaion of the country already in place, it is safe to elnate this as it is duplicate information, and the numerical values are much easier to work with anyway

Continued on next page

Table 1 – *Continued from previous page*

Variable	**Reason For Deletion**
year	The year the data was collected in while certianly helpful in undersatnding how situations changed for ethnic grups across time in nations, is not particuarly insightful into their risk factors.
AUTONEND	This variable origionally refered to the year that the last mjor loss of automity occured for the group. This would be vry helpful for historical researchers, bu as the date was already discretized in the variable YEARWT, it was deemed unceessary and deleted
TRANSYR	Once again this was a case of the variable being covered already, in this case with YEARWT as well. This variable was simply a measure of when authority was transfered an it the vast majority of times confered with AUTOEND.
FACTCC1	while the presence of intergroup conflict is particuarly important, the name of the group that is involved in the coflict is not particuarly important. In addition this measure is very sparse, and thus pruning made sense
FACTCC2	This is the exact same logic as the for the FACTCC1 execpt this is name of the second group a minrity group is in conflict with.
FACTCC3	Exact same logic as FACTCC2 and FACTCC1
FACTCC2	This is the exact same logic as the for the FACTCC1 execpt this is name of the second group a minrity group is in conflict with.
CCGROUP1	We have now moved on to instead of intercommunity conflict to intra-communiy conflict. The logic behind not being particuarl concerned with the names of the groups invloved, just their prenesce or lack therof still stands however.
CCGROUP2	Ditto
CCGROUP3	Ditto
REPNVIOL	This is outside the scope of the project. The purose of the project as expalind above is to attempt to use machine learning to attempt to predict whether or not a state will inflict violince or persecution on a minority group within its borders. Being repressive to protesters, while normativly reprehsnisble is outside the scope of the project.

Continued on next page

Table 1 – *Continued from previous page*

Variable	Reason For Deletion
REPVIOL	Similarly to REPNVIOL the goverments crackdown on violint seperatist while illuminating is not particuarly informative on their reations to civilians, Thereofore this too is deleted.

3.2.2 Deletions

The data has been modified extensivly as the origional data was not particuarly well suited to the type of anylaisis that I wished to accomplish. In some cases this was because the data was coded in text, and I simply needed to recode the data to be strings. However, as will be discussed below there were sigificant challenges that I experienced with my data, and as a result I was forced to extensivly modify the data, to improve the results that I was experiencing. The purpose of the data was not to force a certian result, but was designed to reuce th overall complexity to allow for more efficint anylasis. A prime example of this was the recodeing of missing data. In the original MAR data any missing data is codded as -*99*. However, especially in binary variables, this 99 would be weighted as 99 times the time complexity of the 0 and 1 tht signified the existance of the variable. As simply deleting this data was snot an option due to the size of my data set, I siply recoded every -99 in the data as -1. In addition smaller changes were needed to several variables to achive similar effects a process I have documented in the table below:

Table 2: A simple longtable example

Variable	Reason For Deletion
VMARREGION	Regionality, and broader social and cultural actor may very easily play a role in determing the response to ethnic groups and therefore it is important to include. However, as it was currently it was strings, and thus I decided to convert it to categories as follows: 0).Asia, 1). Latin America and the Caribbean, 2). Middle East and North Africa, 3). Post Communits States, 4). Sub Saharan Africa, 5). Western Democracies and Japan

Continued on next page

Table 2 – *Continued from previous page*

Variable	**Reason For Deletion**
REPGENCIV	The purpose of the project as defined above was simply to estimate whether violance would occur against a minority group given a set of circumstances. Therfore, trying to decide what scale of violnce was being committed seemed to me to be rather more difficult. Therfore I recoded the data binarilly, with 0). represnting no violance and 1). representing violence bing conducted byt the Government.
GPOP	I decided that to reduce the overall complexity of the data, it would be smart to reduce the population of the minority groups down into several discrete sets. I accomplished this through taking the log base 10 of each group, a process that convereted the data down to a reasonable scale of around 5 different values. This actually massivly skyrocketd the consistancy of my data
CPOP	I did the same thing on this data that I did on the GPOP data in discretizing it down to just a few values. This change also marked a noticible improvement in the qality of my results.