1 Abstract

The Minorities at Risk (MAR) is a long running project to monitor an analyze the situation of minorities around the world. While flawed due to only considering minorities that the are already at some sort of political, social or physical risk, it has been used in a number of cited papers to analyze using primarily linear regression to determine what key risk factors are. However, fitting into a broader trend in social sciences these investigations have been broadly analytical instead 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 research conducted within this project to be placed within a proper context, and the significance 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 computing world, making problems that were once considered near impossible easy, and opening up entirely new areas of insight. Unfortunately as it currently stands much of this same process has not yet fully 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). Fundamentally 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 which attempts NORMAL which simply attempts to determine based off of role call voting the political allegiance of members of congress. It does even attempt what I am trying to do, based off of a set of voting behavior predict how congress, or individual congress members will vote on specific piece of legislation(Grimmer 1).

Where Machine Learning however has had a significant impact on Political science then is in sentiment analysis , attempting to determine the emotion associated with the piece of writing. In an era of decentralized news, and increasing political polarization then these techniques are crucial to achieving a firm grasp on the plethora of on line communities that people express them in. Sentiment analysis 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 research I have been interested in is two fold: there is a lack of good data available, and they have repeatedly 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 , that well suited to the type of algorithms favored by machine learning are hard to find in the political science world. Most interesting features of political sciences: trends, elections, and conflicts are slow moving and infrequent, a situation that leads to a dearth of data.

2.2 Research on the MAR Project

The Minority At Risk data set as one of the most famous, largest, and influential data sets within political science. Science its conception in 1986 by Ted Gurr in 1988 at the University of Maryland it has been cited innumerable times, and several books have been written analyzing its data. fitting however into a broader trend of political science failing to fully embrace Machine Learning the field has yet to apply predictive methodologies to it. In the only case that i could find of any one attempting predictive behavior, through "risk analysis" it was in the context of least squares regression.

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 discussed previously there has been a distinct absence in the use of these technologies within the field. Therefore I was largely in uncharted territory, and was simply 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 creators are credited in the sources section of this paper, to conduct my analysis. As will be 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 Bayesian classifier, a basic Perceptron, and a slightly more complicated multi-layer neural network. I chose these methods as they have been used on similar data sets, in other fields to reasonable levels of success. In particular, I must complement all of the work that went into the PIMA data set, as I used many of the techniques developed there to approach 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 originally stood each of the data entries was individually labeled with over 40 different variables. As we only had Slightly over 850 different cases, it was clear that we would be forced to prune the data to allow it to be more accurately surveyed. The exact specifics, and 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 identifier such as name. This type of variable is ineffective 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 categories to numerical ones that the computer could full process. In the chart below I refer to each of the variables that were changed or eliminated by variable name as defined in the code book for the MAR project. The data is broken up into tow subsections. Those that I eliminated from the project and those that I simply 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 general 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 predictive capabilities about future groups this clearly has no relevance and can be safely eliminated

Continued on next page

Table 1 – *Continued from previous page*

Variable	**Reason For Deletion**
VMARGROUP	VMARGROUP is simply the name of the group. Once again this is far to specific, and will not accurately create generalizable trends. Therefore it is also eliminated
country	This is the actual name of the country, as there is also a numerical representation of the country already in place, it is safe to eliminate this as it is duplicate information, and the numerical values are much easier to work with anyway
year	The year the data was collected in while certainly helpful in understanding how situations changed for ethnic groups across time in nations, is not particularly insightful into their risk factors.
AUTONEND	This variable originally referred to the year that the last major loss of autonomy occurred for the group. This would be very helpful for historical researchers, bu as the date was already discretized in the variable YEARWT, it was deemed unnecessary 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 conferred with AUTOEND.
FACTCC1	while the presence of intergroup conflict is particularly important, the name of the group that is involved in the conflict is not particularly 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 except this is name of the second group a minority group is in conflict with.
FACTCC3	Exact same logic as FACTCC2 and FACTCC1
CCGROUP1	We have now moved on to instead of inter community conflict to intra-community conflict. The logic behind not being particularly concerned with the names of the groups involved, just their presence or lack thereof still stands however.
CCGROUP2	Ditto
CCGROUP3	Ditto

Continued on next page

Table 1 – *Continued from previous page*

Variable	**Reason For Deletion**
REPNVIOL	This is outside the scope of the project. The purpose of the project as explained above is to attempt to use machine learning to attempt to predict whether or not a state will inflict violence or persecution on a minority group within its borders. Being repressive to protesters, while normatively reprehensible is outside the scope of the project.
REPVIOL	Similarly to REPNVIOL the governments crackdown on violent separatist while illuminating is not particularly informative on their relations to civilians. Therefore this too is deleted.

3.2.2 Deletions

The data has been modified extensively as the original data was not particularly well suited to the type of analysis 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 significant challenges that I experienced with my data, and as a result I was forced to extensively modify the data, to improve the results that I was experiencing. The purpose of the data was not to force a certain result, but was designed to reduce th overall complexity to allow for more efficient analysis. A prime example of this was the recoding 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 that signified the existence of the variable. As simply deleting this data was snot an option due to the size of my data set, I simply recoded every -99 in the data as -1. In addition smaller changes were needed to several variables to achieve similar effects a process I have documented in the table below:

Table 2: A simple longtable example

Variable	**Reason For Deletion**
VMARREGION	Regionalism, and broader social and cultural actor may very easily play a role in determining 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 Communist States, 4). Sub Saharan Africa, 5). Western Democracies and Japan
REPGENCIV	The purpose of the project as defined above was simply to estimate whether violence would occur against a minority group given a set of circumstances. Therefore, trying to decide what scale of violence was being committed seemed to me to be rather more difficult. Therefore, I recoded the data binerally, with 0). representing no violence and 1). representing violence bing conducted by 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 converted the data down to a reasonable scale of around 5 different values. This actually massively skyrocketed the consistency 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 noticeable improvement in the quality of my results.

4 Results

The results of this experiment were promising. I am by no means an expert in Machine learning and yet I was able to achieve the high accuracy rating of around $> \%88$ on predicting what amounts to ethnic cleansing. It is possible to see from this results then how these tools could be used both in creating policy as well as in researching various trends in political science.

The results here were presented in three parts and represent the three approaches

that I attempted in my efforts to apply machine learning to the MAR Data:

4.1 Bayesian Start

In reality this was my second piece of technology that I attempted, after a fairly unsuccessful run with the perception, which I will describe later. However, this basic approach was quite successful at least at first glance with an accuracy of.

One area that I spent a significant amount of computational power on was attempting to determining the ideal distribution of data between my training and test data sets. The initial attempt to do this simply divide the data in the rations $(0.5, 0.67, 0.75, .80)$ and then ran the Bayesian classifier 15000 times on the data to achieve the following results

Ratio of Training to Test	Results after 15k iterations
0.5	83.33%
0.67	83.69%
0.75	86.85%
.8	83.63%

While the data behind this seemed to be quite accurate I was concerned that this might be influenced by the sorting of the data in the original CSV file which was sorted by region, potentially leaving entire regions out if the ratio was too small. Therefore I ran the exact same experiment again, but this time I shuffled the data on each run of the classifier to ensure a more equal spread (on average) of the data. The results of that were as follows:

Ratio of Training to Test	Results after 15k iterations
0.5	84.27%
0.67	85.10%
0.75	87,32%
.8	84.79%

As can be seen this significantly affected the results of the data, and as it was a more accurate slice of the data, I continued this process on the rest of the data that I collected for the experiment. Regardless of the result however these results were rather significant. Being able to predict the results of violence this accurately based of a collection of essentially arbitrary values seem like quite a satisfactory result. However, there are issue with these values that should be apparent in the following confusion matrix, compiled from a fairly average result of the classifier:

	Labeled False	Labeled True
Predicted False	213	19
Predicted True	22	28

A can be seen from the confusion matrix the high level of true negatives is essentially confusing the true accuracy of the result which is not nearly as high. The sensitivity for example is only 60%, and the sensitivity is precision is 55%. In other words I am only correctly identifying around 60 percent of actual violence committed against minorities, and my prediction of such is only around 55% accurate. In such as sensitive matter, where speed and discretion are necessary these values are clearly not what would be expected, or even desired. Therefore necessitating the creation of more robust standards for analyzing the data set.

4.2 The Perceptron

A perception may allow for a more nuanced approach to the solution, by learning to weight the variables that it is offered. However as will become apparent from the data presented here this was not entirely successful in this case. As we can see from the initial report there may have been an over saturation of the data, and much of the data was badly coded for a machine learning perspective. The initial results at a 1000 runs were particularly problematic as they had only the following accuracy:

Run	Results after 1k iterations
1	18.90 %
2	34.54 %
3	29.34

While the complete and total lack of accuracy was concerning, an equally large consideration was the lack of consistancy. I attempted this eperiment at 10k iterations of training and only recieved no signifiacnt reduction in the spead of the data. Therefore, fearing that the dimmesnions of the problem were simply too large for the data that i had avaliable I attempted to prune the data down to a more reasonable subset, Unfortunatly I did not have the the rescource to do a power set of these values, as 62! is $3.14E^85$, to see which the correct values to prune where and simply had to make a judgement call on what I thought would be particualarly useful. These variables are listed in the following chart:

VMARRegion	GPRO	LANG	CUSTOM	BELIEF
RACE	GROUPCON	GC119	GC2	GC10
GC11	AUTLOST	SEPX	SEPKIN	EMIG
DISPLACE	POLDIS	ECDIS	CULPO1	CULPO2
GOJPA	AUTON2	LEGISREP	EXECREP	GUARREP
POLGR	ECGR	CULGR	KINSUP	STASUP
NSASUP	INTRACON	INTERCON	PROT	REB

The results form this were much better from the very outset with values stabilizing. As a result I attempted to then do a standard windowing of the data, training a variety of duffrent learning rates as well as training iterations to see if I could improve the

frequency in some significant way. This was partially successful as can be seen from the table below:

Run	Results after 1k iterations
1	18.90 %
2	34.54 %
3	29.34

Not only was the accuracy greatly improved by doing this, but the consitancy was as well. Unfortnatly the results were still not as good as that offerd by the Bayesian anylsis in straight up accuracy, asnd when the confmat was examined, there were similar issues w