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 groups 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

Coming soon....

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

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

Continued on next page

Table 1 – *Continued from previous page*

Variable	Reason For Deletion
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
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 particularly 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

Continued on next page

Table 1 – *Continued from previous page*

Variable	Reason For Deletion
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.
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

Similarly sevral of the variables had to have their data discretized to make the prhgram function better. I have detailed the necesary changes below for convenience. I have listed the changes that were necessary below for clarities sake:

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