1 Abstract

The Minorities at Risk (MAR) is a long running project to monitor and analyze the situation of politically relevant minorities around the world. While suffering from significant issues with selection bias, it has been used in a number of cited papers to analyze what key risk factors for inter-communal violence, and more broadly how minority groups are treated in society. However, fitting into a broader trend in social sciences these investigations have been broadly analytical instead of predictive, focused on understanding questions of "Why?", instead of "What will happen if?". Thur purpose of this project is to address the viability of asking the second question, trying to predict based off of range of variables whether a state will commit violence against a minority group, using modern principles of machine learning.

2 Literature Review

The background for this project is rooted in two distinct points: the use of advanced statistical analysis and machine learning within the social sciences, and an investigation of other research on the MAR project. Through a better understanding of these it is possible to then place both the project itself as well as the results presented below in the correct context:

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 these same process have 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. It therefore has been used 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 of the field in political science, NORMAL, which attempts to determine based off of role call voting the political allegiance of members of congress. It simply attempts to quantify the existing spread of political thought instead of attempting to predict how congressmen will vote on new legislation(Grimmer 1). This fits into the observations of Timothy Gurr, the founder of the MAR Project itself who stated, "Empirical political science has been preoccupied with large-n comparisons either at higher levels of analysis, in the form of quantitative studies of properties of states and the international system". In fact, economists have in general more active, and successful in using their field to predict political models than political scientists themselves.

Where machine learning however has had a significant impact on political sciences is in sentiment analysis, or 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 online 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). Interestingly political parties have already realized this trend and have been using it to the fullest effect in effort to target voters that may be likely to switch their vote in a coming election (Woodie), leading "political parties to engage in an arms to leverage ever increasing volumes of data" (Nikerson and Rogers 51). In fact one major company has used associated analytical technologies to, "entire 255 million person national voter file to digital platforms like Facebook, Google, Yahoo, and MSN by using personally identifiable information " (Woodie). The failure of political scientists to engage, or even understand these technologies then represents a lack of knowledge about one of the most major changes in how political campaigns are run within the last few decades.

It is important then to understand why political scientist have been unwilling to engage in this type of research. The reason 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). In addition when models fail to predict significant trends it is also tends to be a major story, such as the election of Donal Trump. This is not entirely the fault of political scientists however, most of the topics that the general public, and grant committees are interested in, social movements, elections, and conflicts, are slow moving and infrequent. This creates a dearth of data for these types of analysis to be run on. Even the MAR Project, considered one of the largest and most complete data sets in international relations, only has 850 cases which is on the low side of an effective machine learning based analysis.

Regardless of the case however this may be changing. The increasing availability of data sets, particularly in domestic politics makes this type of examination far more palatable. Over the last few years there are increasingly more cases of this research and several major universities have opened up either departments or institutes in computational social science. Political scientists may be behind the times on this issue, but there is no reason to believe this need be permanent.

2.2 Research on the MAR Project

The Minority At Risk data set is one of the most influential datasets within political science. Since its conception in 1986 by Ted Gurr at the University of Maryland it has become one of the canonical source on understanding how and why intra-state violence occurs. The two main books written on it by Ted Gurr have been cited over a 1000 times each, and have shaped much of the current understanding on these topics. Fitting however into the broader trend of political science failing to fully embrace machine learning, predictive methodologies have yet to be applied to it. Most papers using the data focus on topics such as "The Unique Role of Religion in Middle Eastern Ethnic Conflict" (Fox 1) and "Discrimination in International Relations: Examining Why Some Ethnic Groups Receive More External Support Than Others" (Saideman 2). However, the data set also was a massive step forward for the time and scholars such as Simon Hug have credited this data set as one of the first attempts to address the topic in a truly quantitative way and as a result it,"opened the way for much more systematic analyses of ethnic violence and mobilization" (Hug 21).

While the data set is doubtlessly important it by no mean is perfect. In particular selection bias has been identified as a significant error of the Minority at Risk Data set. In the words of Hug, "mobilizing,rebelling, and violent ethnic groups are simply much easier to identify than peaceful, reclusivly living tribes" (Hug 289). This has created a noticeable selection bias in the data, where the groups included are often included on the basis of either them being involved in violence, or being the victims of it (Binir et al. 3). In addition the data is biased towards the Middle East and the America's, largely ignoring minorities in post communist states and communal societies in Africa. As the communities often have some of the most complex interactions with their governments their inclusion in the expanded AMAR project has been identified as a mater of critical importance by those maintaining the data set. (Gurr 2).

As a result there are currently attempts to create a new and improved dataset, the All Minorities at Risk Data set that will include every minority they can possible find that matches certain criteria of population, cohesiveness and organization and then include them in the project. The first phases of this project was completed in 2016, and identified 1200 such groups - a quadrupling of the original data set (MAR). This project therefor contains the possibility that this data set will continue to be relevant far into the future.

3 Methodology

The basis of this project was to see the viability of using basic predictive techniques of machine learning on the MAR project, and in political science by extension. As there was no strong basis on how to approach this project a variety of techniques were employed, noticeably: a naive Bayesian classifier, a basic perceptron, and fully developed neural net. To allow for these approaches to be used a wide variety of data grooming took place, with an emphasis on loosing as little relevant information as possible. The exact specification on each of these processes is contained in the next three sections.

3.1 Modeling

As this project is arguably one of the first to attempt to apply machine learning to these data set a general approach was taken. All of the three technologies used in this project are employed broadly, instead of being specialized for one particular field such as image recognition. Perceptrons for example have been used on everything from basic sentiment analysis to image recognition, being one of the most basic tools in the machne learning toolkit. In fact much of the inspiration for how this project approached the modeling of the problem comes as a result of the PIMA data set, which attempts to predict diabetes risk based off of a variety of health and genetic factors. As a result the results of this project, while certainly "best effort" and thorough, should not be considered the upper limit for what machine learning has to offer the field of political science or even this data set.

It is a also briefly important to consider the specifics of what this project is a tempting to predict, mainly based of a variety of societal and economic factor a minority group will have violence committed against it. This simple statement contains a great deal of complexity and contention and accordingly needs a little explanation. As this project is based of the MAR data set, this project according,y adopts their definitions namely:

Table 1: Criteria for group to be included in the MAR Project

Criteria
Collectively suffers, or benefits from, systematic discriminatory treatment vis-à- vis other groups in a
Collectively mobilizes in defense or promotion of its self-defined interests.
Membership in the group is determined primarily by descent by both members and non-membe

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 discretized 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 2: 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
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

Continued on next page

Table 2 – *Continued from previous page*

Variable	**Reason For Deletion**
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
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 3: 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.

Continued on next page

Table 3 – *Continued from previous page*

Variable	Reason For Deletion
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 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. At 10K iterations there was still significant variation in the results, making any sort of windowing or tother more sophisticated methods to improve consistancy rather difficult. Therfore, to reduce the overall complexity of the problem the data was severelly pruned, from over 60 variables to 35. This prunning was based on three things, the number of null values that a variable had associated with it, if the vraiable was replicated or represtend in some way else where in the data, and finally based on research conducted by other political scientist on what variables are important in determining wheteher or not the groups were prone to conflict. For the first two reasons the resonong was based off of prinviples of machine learning, particularly that reducing "noise" in data can lead to improved results. While the perceptron should be able to do this, on a such a large problem, that also had limited data it made sense to try to reduce the problem as much as possible for the machine. The other cutings were based on preserving the values that polititical scientists such as Gurr himself hasve idtified as particuarly important in determing the position of groups in civil socities. The end result of this cutting is documented 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 of doing so were dramatic. Not only did the accuracy shoot up to over 80%, but the consistance of the result was alos immesurable improved. At 2000 iterations there was less than a 0.01 % change in the accuracy of the data on any one, a drastic change from the results seen previously. Therefore, a windowing process occured where variaou learning rates an training iterations were attempted in an effort to find the ideal set of constraints for this data. This process was quite successful, and the results of it are displayed below: //

NIterations / LearningRate	50	100	150	200	250	300
0.05	82.94	83.64	84.08	84.41	84.25	84.22
.10	82.53	83.91	84.08	84.10	84.41	84.44
.15	82.87	83.83	83.89	84.36	84.30	84.20
.20	82.79	83.59	84.26	83.99	84.04	84.20
.25	82.90	83.37	83.79	84.16	84.46	84.36

As can be seen from the chart the idea rate seems to have been 250 iterations with a laerning rate of .25, which in effort to maximize the resukts offered by this model, was adopted for all subsequent experiments. While the accuracy of the perceptron at 84 % is directly comparable to the results in terms of accuracy to the Basyian classifier, the results in terms of recision and specificatio are around 10% higher: a fact demonstrated by this chart below of an average confusion matrix

	Labeled False	Labeled True
Predicted False	215	20
Predicted True	17	32

As can be clearly seen from this chat, for a siimilar accuracy result to that demonstrated in the , the precision was 65%, and specification was 60%. Once again this means that I correctly identified political violince 65 % of the time and that if the algorithm identified a case of political violince there was a round a 60% chance that it was true. As these results are significantly better than the niaeve approach, it seemed reasonable to progress into creating a full multi-layer Neural network

4.3 The Full Neural Network

Coming soon....

5 Discussion

The results of this experiment are while not outstanding are positive, and should be considered an intresting look forward to how these techniques can be used in the context of political science. The fact is that while 85 % is hardly a spectacuarlar result, as there is not other information on the topic, it is infintly better thann any other measure that political scientist have to predict these trends.

Futhermore there are two factors that point to the abiluity to extend this project, so that it is able to be used as a policy and research tool. The first of these is that I am by no means an exper in Machine Learning. While I am familiar and comfertable with the tools that I used in the context of this project, there are far more potent tools currently avaiable in Machine Learning, and some may be more applicable to this project. In addition, I do not have much experience in the pruning or contol of data, and some one with more experience may be able to produce significatntly better result from the Data that I had access to.

The second piece of hope for the project is that the data set in question, the Minorities at Risk Project, is in the midst of major revison effort that will approximatly quadrouple its size, an exciting prospect as it will provide much more data for the machine to train on.

This however may actually harm the overall of efficay of this project, as it encounters one of the mai issues that I faced. While these machines are great at what they do, they work much batter in the context of equall distrubuited data, epecially on predictions that are more complex such as the one tattempted in this paper. with a distribution of just under 80% coded as there being no violince, these models may simply resort to replying no on every option, as by doing so they can achivieve a relitivly high rating set. While this paper was able to create a higher accuracy than tha, and the Perceptron in particualar actually was fariy decent at distibuting its data, 80% is a higher accuracy than can be achieved on many data sets. Therefore, if 900 additional samples are added to the MAR data, most of them negative, it may significatly effect if not the accuracy the precicion and sentitivity of the data by encouraging the models to "cheat". Not including these groups however aslo proses problems, most of these grups were selected for bing "politically relevant", a term that the authors of the minoritite at risk have admmittted is mostly associated with violence, introducing a particarlt problematic bit of selction bias (Gurr et. al 4). When taken together these pose reasonable questions to the singifcancy and accuracy of the findings here, one to which there are no easy answers to.

The bigger problem however is that this model as it is now is not partcuarly useful as it currently stands. Acedemics are primarily intrested in uderstanding why these outbreaks of violince occur, not necessarily with when ore where they will. On the other hand the accuracy of the information is hopefully not eneought to tempt policy makers. The rise of the Right To Protect (R2P) doctrine has significantly increased international pressure to stop these atrocities (Western and Goldstein 364). However, despite the questionable judgment displayed in both the invasion of Iraq as well as out president elect, basing regime change off of a model with only 90% accuracy is hopefully unlikly. Where this could be useful is for internatioanal aid workers, and NGO's who could respond to areas of concern, basing their response on the other work that has been done on this data set.

None of this is to say however that this experiment is not important: in fact it may be critical. As local crises and etbnoc conflicts become international issues, as evidenced by the Syrian war, having the tools and abiity to be able to predict and counteract these events may be crucial. This particular expirement and moel may not offer the ncessary precision or accuracy to do so, but it does prove that it is possible. Other fields of machine learning such as image recognition tht have exploded over the course of the last decade where in general not solo projects by undergrads. As more time, expertise and rescources are devoted to this, adn similar problems, there is nothing to suggest that similar advances cannot be made. Political science may be a field that historically is based in qualitiative anyalsisi, but this project may be the final piece of proof that nothing can escape the coming revolution offered by machine learning.

6 Acknolwedgments

This project would not have been possible without the rescoures that the machine lrearn ing community has generously placed online. In particular the work of Steve Masey of the University of New Zealand was particuarly instremental in helping to understand how the neural networks work, and most of the basic code used in that segment is either his or based off of it. In addition, the website" " which provided the basis for my code in the Bayesian Classifier.

7 Refrences