CSC 597-Statistical Learning
War Susceptibility Project Report
Berk Mankaliye & Isaac Attuah

Introduction

In history, few topics have persisted in our conversations: power, love, and the highly sought "happily ever after". While these terms might be synonymous in some respects, they share a common affiliation: war. Like all living organisms, humans have been particularly grim about war. History is riddled with wars for fortunes, territories, and riches, all for the aforementioned topics. This makes one think if our past is filled with so much violence, what is to say our futures do not hold the same fate.

Whether it is the Taliban enforcing their far-extremist propaganda on innocent civilians, Russian troops desecrating Ukrainian cities, racism reeking through cities of Europe, or gang activities terrorizing citizens, violence is very much present all around us. As long as this trend persists, war is imminent. If history tends to repeat itself, the next best thing we can do is to focus on high-risk areas and mitigate the escalation of war. This is exactly why it is of utmost importance to use the available technology and data from past events to predict the potential emergence of war and conflict, and if inevitable, how long it will take before it can come to an end.

Our project aims to answer the following questions:

- Is it possible to predict the emergence of a conflict or a potential war in 20 years based on the presence / absence of conflict / war in subsequent periods from a given year?
- How do wars end and what factors tend to influence the end?
- What factors influence the emergence of smaller conflicts or massive wars?

Background

Current studies revolving around this topic include UCDP (Uppsala Conflict Data Program) which is a project worked on by a team at Uppsala University, Sweden that lists all information provided in the dataset as an interactive map. Furthermore, an online project that we found approached the predicting of future wars and conflicts in 2019-2024 through another database that used variables such as corruption, debt, infant mortality, and many more to predict conflict. Lastly, another group used the Correlates of War database (the database that we were considering using initially) to analyze periods of peace between global-scale wars to determine a statistically significant period of peace. Our objectives stem from these projects, and we intended to pursue a project that builds on the ideas.

Methods

Dataset

We are utilizing the UCDP (Uppsala Conflict Data Program / Peace Research Institute Oslo) Onset Dataset version 19.1 and a modified version of the UCDP/PRIO Armed Conflict Dataset version 21.1 to perform our analysis. The former has 10, 203 instances with 13 attributes while the latter has 2,253 instances with 6 attributes. The tables below show each of the attributes, their content and types.

Onset Dataset Version 19.1

Variable name	Content	Type
abc	Country abbreviation	String
name	Name of location according to the government side in the conflict.	String
gwno_a	The Gleditsch and Ward country code	Integer
newconf	Coded 1 if the country-year contains a new	Integer

	conflict/conflict-dyad (not a new episode of conflict)	
onset1	Onset of an intrastate armed conflict, >25 battle deaths. Coded as 1 if this is a new conflict or there is more than one year since the last observation of the conflict	Integer
onset2	Onset of an intrastate armed conflict, >25 battle deaths. Coded as 1 if this is a new conflict or there is more than two years since the last observation of the conflict	Integer
onset3	Onset of an intrastate armed conflict, >25 battle deaths. Coded as 1 if this is a new conflict or there is more than three years since the last observation of the conflict	Integer
onset5	Onset of an intrastate armed conflict, >25 battle deaths. Coded as 1 if this is a new conflict or there is more than five years since the last observation of the conflict	Integer
onset10	Onset of an intrastate armed conflict, >25 battle deaths. Coded as 1 if this is a new conflict or there is more than ten years since the last observation of the conflict	Integer
onset20	Onset of an intrastate armed conflict, >25 battle deaths. Coded as 1 if this is a new conflict or there is more than twenty years since the last observation of the conflict	Integer
year	Year of observation	Integer
year_prev	Previous year of observation	Integer
***duration**	Length of time conflict persisted	Integer

Armed Conflict Dataset Dataset Version 21.1

Variable name	Content	Type
year	The year of observation (1946-2020).	Integer
intensity_level	The intensity level in the conflict per calendar year. The intensity variable is coded in two categories: 0 or 1 0. Minor: between 25 and 999 battle-related deaths in a given year. 1. War: at least 1,000 battle-related deaths in a given year	Integer
cumulative_intensity	This variable takes into account the temporal dimension of the conflict. It is a dummy variable that codes whether the conflict since the onset has exceeded 1,000 battlerelated deaths. For conflicts with a history prior to 1946, it does not take into account the fatalities incurred in preceding years. A conflict is coded as 0 as long as it has not over time resulted in more than 1,000 battle-related deaths. Once a conflict reaches this threshold, it is coded as 1.	Integer
ep_end	Onset of an intrastate armed conflict, >25 battle deaths. Coded as 1 if this is a new conflict or there is more than one year since the last observation of the conflict	Integer
gwno_a	Onset of an intrastate armed conflict, >25 battle deaths. Coded as 1 if this is a new conflict or there is more than two years since the last observation of the conflict	Integer

An inner join was performed between these two datasets (using the gwno_a and year) at

to produce a dataset with 877 instances and 17 attributes.

A duration attribute was also computed to show how long war persisted

For additional information, kindly reference the Onset dataset codebook or the Action dataset codebook. There are no missing values in the dataset, and the absence of conflict is denoted by a 0. New Conflict (newconf), onset values (onset2, onset3, onset5, onset10), intensity values, and episode end values were used for the analysis to predict onset20, war episode end, or the intensity of war if it persists. The digit assigned to onset"x" represents the year duration of a conflict. onset1 was excluded since it did not have enough variation for any of the models.

Techniques Applied

Prior to testing each of the techniques below, we performed an 80/20 split on the onset data set where 80% is the training data and 20% is the testing data. The columns "abc", "name", "year", "gwno_a" , "onset1", and "year_prev" were removed from the data set since we wanted to perform analysis on solely categorical variables.

i. Logistic Regression

We used glm.fits() method to analyze logistic regression. The code then creates a binomial family model using glm() to predict whether there is an increase in conflict duration after each of these events. Below is a confusion matrix to see how well the model is performing.

ii. Linear Discriminant Analysis (LDA)

We used lda.fit to determine the linear discriminant analysis. The code uses the input variables described above to predict onset20 variable found in the testing set. The code checks for the confusion matrix to see how well the model is performing.

iii. Decision Trees

We created a tree to make predictions on the test set. The code checks for the confusion matrix to see how well the model is performing.

iv. Regression Trees

We created a tree object to analyze the training set. We also did prune.tree() to prune the tree. We then proceed to generate a tree object which will be trained and analyzed using cross-validation with cv.tree().

v. Random Forests

The code uses the randomForest() function to create a model.

vi. KNN

K-Nearest Neighbor function is a great supervised non-parametric model used in regression and classification problems. We thought that using k-NN to make classification-based predictions would yield a high accuracy rate based on the dataset we were using.

vii. SVM

Support Vector Machines are also great supervised models used in classification and regression problems. It comes in two forms, where the **linear** model uses a linear boundary that might or might not be effective in determining the boundary between two classes depending on the layout of the points. When the linear model is not sufficient, running the **polynomial** model might help accommodate more points appropriately. We wanted to see both cases.

viii. Partial Least Square

Lastly, we used the partial least square model to reduce the number of dimensions in our dataset in order to obtain better accuracy in predictions. PLS similarly works well with classification and regression problems and is a supervised learning model.

Evaluation/Results

Phase 1: Predicting Outcome In 20 Years (onset20)

This phase involved predicting the occurrence of conflict in 20 years based on conflict, intensity level in previous conflict and non-conflict years.

I. Logistic Regression

Logistic regression served well for determining outcome in 20 years. With an error rate of 0.48%, this model worked near perfectly.

```
predictions   0   1
          0 101   1
          1   0 106
Test Error Rate
[1] 0.004807692
```

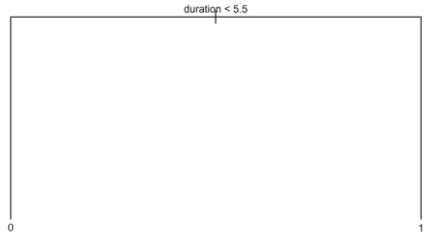
II. Linear Discriminant Analysis

LDA functioned similarly to glm model. It had the same accuracy rate.

```
predictions   0   1
          0 101   1
          1   0 106
Test Error Rate
[1] 0.004807692
```

III. Regression Trees

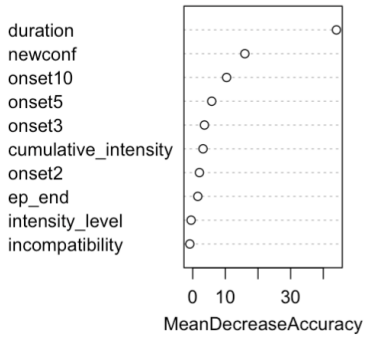
Similar to logistic regression, decision trees achieved 99.52% accuracy.

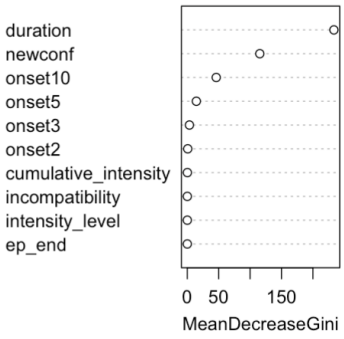


```
predictions   0   1
          0 101   1
          1   0 106
Test Error Rate
[1] 0.004807692
```

IV. Random Forests

Random forests also yielded 99.52% accuracy in predicting onset20. Furthermore, the model allowed us to see which variables *duration* and *newconf* were the highest predictors of mean squared error.






```
predictions   0   1
          0 101   1
          1   0 106
Test Error Rate
[1] 0.004807692
```

V. K-Nearest Neighbor

KNN worked no differently than the rest of the models. It yielded 99.52% accuracy.

```
predictions   0   1
          0 101   1
          1   0 106
Test Error Rate
[1] 0.004807692
```

VI. Support Vector Machine

In the future phases, we saw that linear SVM and polynomial SVM models have differences in their outputs. However, when determining outcome in 20 years, the output was same with each other and elsewhere.

a. Linear

```
predictions   0   1
          0 101   1
          1   0 106
Test Error Rate
[1] 0.004807692
```

b. Polynomial

```
predictions   0   1
          0 101   1
          1   0 106
Test Error Rate
[1] 0.004807692
```

VII. Partial Least Square

Lastly, PLS had a similar outcome to all other models.

```
predictions   0   1
          0 101   1
          1   0 106
Test Error Rate
[1] 0.004807692
```

Summary:

Method	Accuracy	Error Rate
Logistic Regression	99.52%	0.48%
LDA	99.52%	0.48%
Regression Trees	99.52%	0.48%
Random Forests	99.52%	0.48%
KNN	99.52%	0.48%
SVM (Linear)	99.52%	0.48%
SVM (Poly)	99.52%	0.48%
PLS	99.52%	0.48%

Phase 2: The End of War (ep_end)

This phase involved predicting the end of conflict. Primarily identifying the various factors that lead to the end of conflict.

I. Logistic Regression

We see a decrease in accuracy when trying to predict the end of conflict. GLM performs the same as LDA.

```
glm.pred   0   1
       0 127  55
       1  10  16
Test Error Rate
[1] 0.3125
```

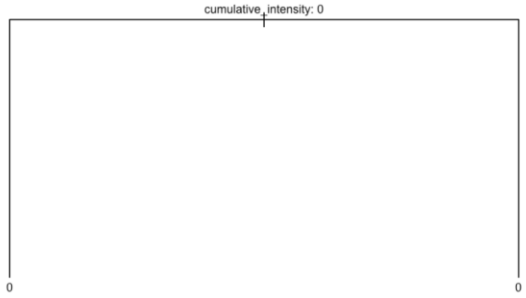
II. Linear Discriminant Analysis

LDA appears to retain constant accuracy when run multiple times compared to the latter models. It had an accuracy of 31.25%.

```
glm.pred   0   1
       0 127  55
       1  10  16
Test Error Rate
[1] 0.3125
```

III. Regression Trees

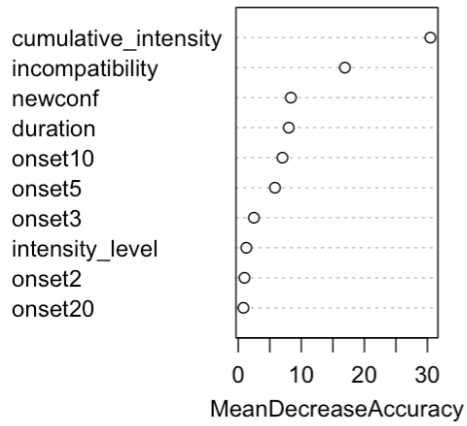
The tree model perhaps is the least accurate model we have. The accuracy is around 49%.

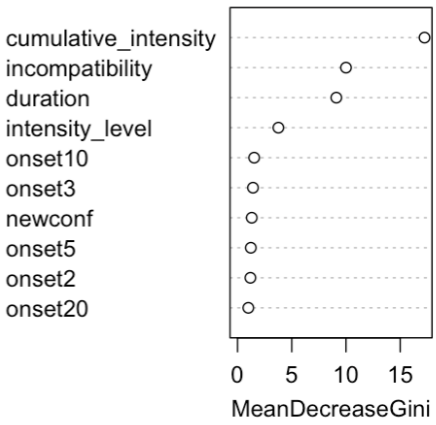



```
tree.ep_end  0  1
           0 70 31
           1 67 40
[1] 0.5288462
```

IV. Random Forests

Random Forests model shows that the cumulative_intensity and incompatibility play a big role in calculating mean squared error. In this run, this model performed similar to partial least square model.






```
predictions   0   1
          0 130  57
          1   7  14
Test Error Rate
[1] 0.3076923
```

V. K-Nearest Neighbor

KNN performs in the 60~70% accuracy range.

```
predictions   0   1
          0 129  55
          1   8  16
Test Error Rate
[1] 0.3028846
```

VI. Support Vector Machine

The polynomial SVM model performs similarly to the SVM linear model. Depending on the run, one model performs better than the other. In this particular run, the linear SVM model performed poorer than the polynomial SVM.

a. Linear

```
predictions   0   1
          0 137  71
          1   0   0
Test Error Rate
[1] 0.3413462
```

b. Polynomial

```
predictions   0   1
          0 129  58
          1   8  13
Test Error Rate
[1] 0.3173077
```

VII. Partial Least Square

Similar to previous models, PLS varies on accuracy depending on the run of the code, but it performs within the range of 60~70% accuracy.

```
predictions   0   1
          0 132  59
          1   5  12
Test Error Rate
[1] 0.3076923
```

Evaluation/Results

Method	Accuracy	Error Rate
Logistic Regression	62.02%	37.98%
LDA	68.75%	31.25%
Regression Trees	65.87%	34.13%
Random Forests	66.35%	33.65%
KNN	69.71%	30.29%
SVM (Linear)	65.87%	34.13%
SVM (Poly)	68.27%	31.73%
PLS	69.24%	30.76%

Phase 3: War Susceptibility (intensity_level)

This phase involved predicting whether a predicted future conflict will remain minor (between 25 and 999 battle-related deaths in a given year) or escalate into what is classified as war (at least 1,000 battle-related deaths in a given year).

I. Logistic Regression

GLM model was relatively good compared to other models when calculating the intensity. It had an error rate of 13.94%.

```
predictions   0   1
          0 142  14
          1  15  37
Test Error Rate
[1] 0.1394231
```

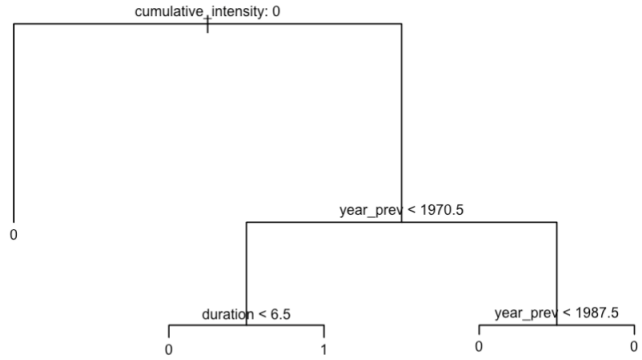
II. Linear Discriminant Analysis

LDA performed relatively well compared to the other models. It yielded the same outcome as GLM.

```
predictions   0   1
          0 142  14
          1  15  37
Test Error Rate
[1] 0.1394231
```

III. Regression Trees

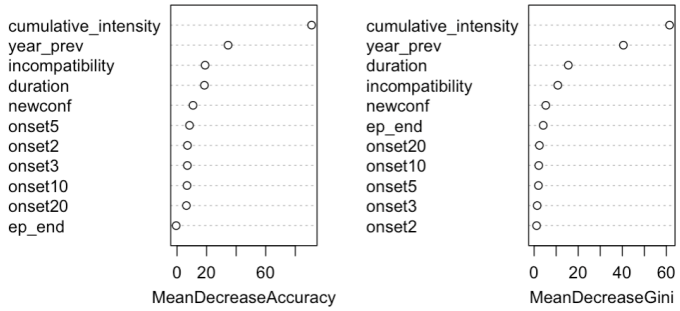
The decision trees created the following tree based on our code. RTs were one of the lesser-accurate models in predicting intensity.




```
tree.pred   0   1
        0 142  15
        1  15  36
[1] 0.8557692
```

IV. Random Forests

RFs was one of our better performing models. We saw that cumulative_intensity and year_prev played the biggest role in determining mean squared error.




```
predictions   0   1
          0 136   7
          1  21  44
Test Error Rate
[1] 0.1346154
```

V. K-Nearest Neighbor

KNN was one of our mediocre models. It remained stable in accuracy rates unlike Phase 2.

```
predictions   0   1
          0 142  15
          1  15  36
Test Error Rate
[1] 0.1442308
```

VI. Support Vector Machine

For our SVM model, the polynomial model performed better than the linear model. The accuracy values remained constant, unlike Phase 2.

a. Linear

```
predictions   0   1
          0 142  15
          1  15  36
Test Error Rate
[1] 0.1442308
```

b. Polynomial

```
predictions   0   1
          0 142  14
          1  15  37
Test Error Rate
[1] 0.1394231
```

VII. Partial Least Square

PLS was the worst model to use for Phase 3. It had the lowest accuracy rate at 22.12%

```
predictions   0   1
          0 154  43
          1   3   8
Test Error Rate
[1] 0.2211538
```

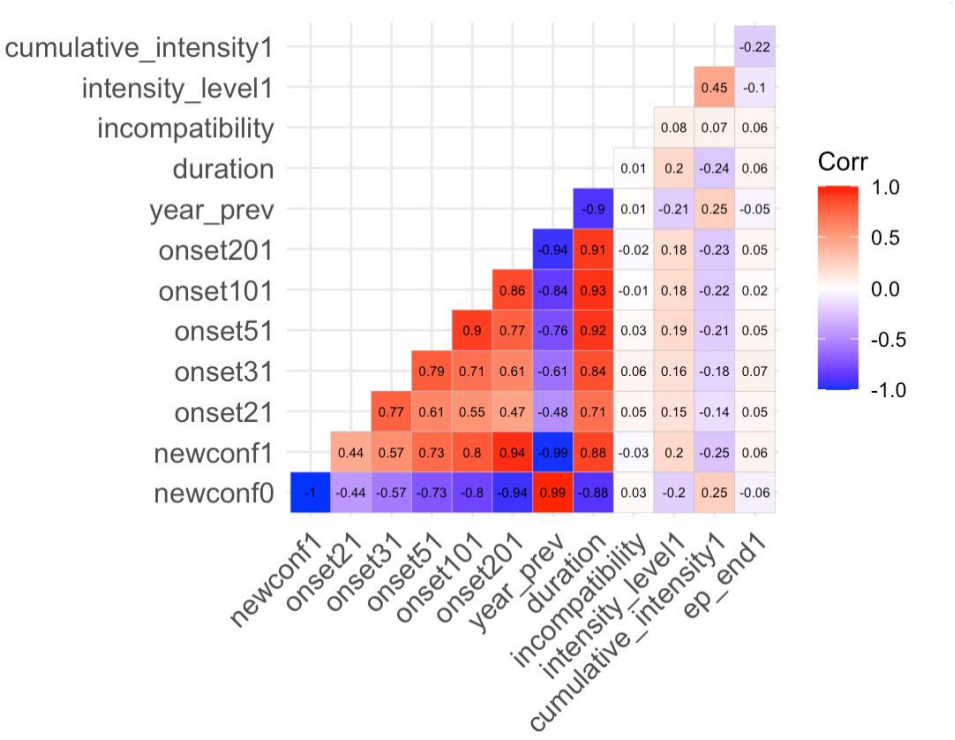
Evaluation/Results

Method	Accuracy	Error Rate
Logistic Regression	86.06%	13.94%
LDA	86.06%	13.94%
Regression Trees	85.58%	14.42%
Random Forests	86.54%	13.46%
KNN	85.58%	14.42%
SVM (Linear)	85.58%	14.42%
SVM (Poly)	86.06%	13.94%
PLS	77.88%	22.12%

Conclusions and Future Goals

Phase 1

Since the variables in this phase are highly correlated, performing cross-validation before running the models yields the same accuracy rates across all models. Diversification or increasing the sample size might resolve this issue.



Phase 2

This phase introduces highly varying datapoints and variables, which is why we observe a great decrease in accuracy levels compared to Phase 1. We noticed that the more advanced models starting from KNN onwards start having different accuracy/error rates when run multiple times,

yet these accuracies remain in the relative range. KNN and PLS had the highest accuracy rates compared to the rest of the models.

Phase 3

We wanted to see whether intensity will remain minor or will increase in upcoming years. With this research question, more variables were accounted compared to Phase 1, but they weren't as varying as Phase 2. We noticed that the best model to predict intensity was Random Forests and worst model was Partial Least Square.

Future Goals

This project showed that many questions could be answered using the database that we used. Future goals include focusing on specific time periods and regions, incorporating more datasets to enlarge sample size, and using model diagrams that are visually pleasing to appeal to an audience that might be unfamiliar with the process of running these statistical analyses.

References

- Uppsala Conflict Data Program Department of Peace and Conflict Research Website
- Gleditsch, Nils Petter; Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg & Håvard Strand (2002) Armed Conflict 1946–2001: A New Dataset. Journal of Peace Research 39(5): 615–637.
- Pettersson, Therese; Stina Högbladh & Magnus Öberg (2019). Organized violence, 1989-2018 and peace agreements. Journal of Peace Research 56(4): 589-603.
- Pettersson, Therese, Shawn Davis, Amber Deniz, Garoun Engström, Nanar Hawach, Stina Högbladh, Margareta Sollenberg & Magnus Öberg (2021). Organized violence 1989-2020, with a special emphasis on Syria. Journal of Peace Research 58(4).
- Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand (2002) Armed Conflict 1946-2001: A New Dataset. Journal of Peace Research 39(5).arme