Utilizing socioeconomic factors to predict the occurence of a civil war in a country

Carnegie Mellon University

Sungwon Lee, Usama Tash, Sandeep Selvarajan, Nitish Chinta, Sarvesh Paradkar

Background and Information

Using a dataset consisting of various socioeconomic features of a country, we attempt to predict whether that country is undergoing a civil war. The benefits of this project are that it provides:

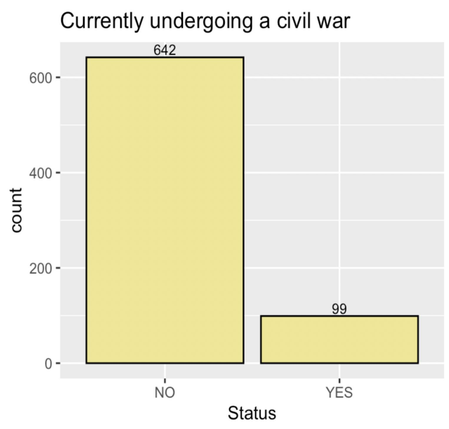
- The ability to build foundational knowledge about causes of war.
- Useful information to decision making bodies such as the United Nations.
- Information about occurrence of a war, which has major effects on calculation of other projections for a nation, such as gross domestic product or annual budgets.

Exploratory Data Analysis

- The dataset contains 741 observations and 6 predictors. The response variable is a factor, `YES` if the country is undergoing a civil war and `NO` if not.
- Most of our predictors are quantitative variables, with `dominance` being the only predictor that is a factor.

`exports`	a measure of the dependence of a country on commodity exports
`schooling`	percentage, school enrollment rate for males
`growth`	annual GDP growth rate
`growth`	annual GDP growth rate
`lnpop`	natural logarithm of the country's population
`fraction-alization`	index measuring divides on ethnic/religious lines
`dominance`	`YES` if one ethnic group dominates the country, `NO` otherwise

Description of each of the predictors

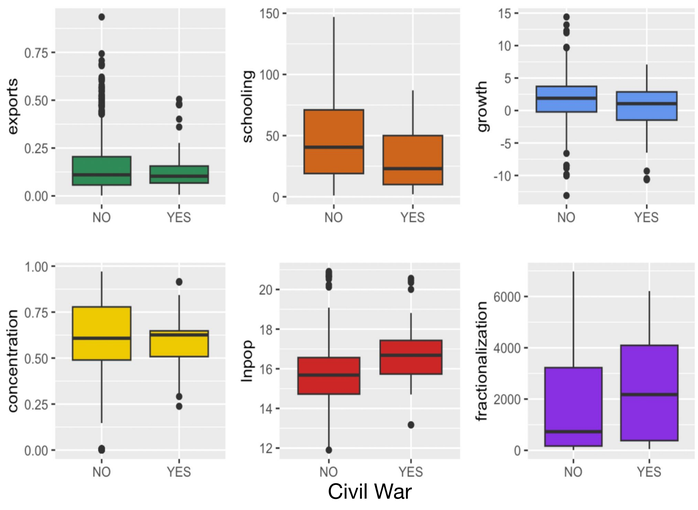


Distribution of occurrence of civil war

```
   exports          schooling
Min.   :0.0020   Min.   :  1.00
1st Qu.:0.0580   1st Qu.: 18.00
Median :0.1100   Median : 39.00
Mean   :0.1546   Mean   : 43.68
3rd Qu.:0.2020   3rd Qu.: 66.00
Max.   :0.9350   Max.   :147.00

    growth         concentration
Min.   :-13.088   Min.   :0.0000
1st Qu.: -0.265   1st Qu.:0.4910
Median :  1.833   Median :0.6140
Mean   :  1.563   Mean   :0.6034
3rd Qu.:  3.557   3rd Qu.:0.7630
Max.   : 14.409   Max.   :0.9710
    lnpop        fractionalization   dominance
Min.   :11.90    Min.   :  12        NO :408
1st Qu.:14.89    1st Qu.: 176        YES:333
Median :15.82    Median : 900
Mean   :15.81    Mean   :1833
3rd Qu.:16.78    3rd Qu.:3375
Max.   :20.91    Max.   :6975
```

Summary statistics of the predictors



The medians for `schooling`, `lnpop`, and `growth` appear to differ, indicating possible associations with occurrence of wars

Analysis

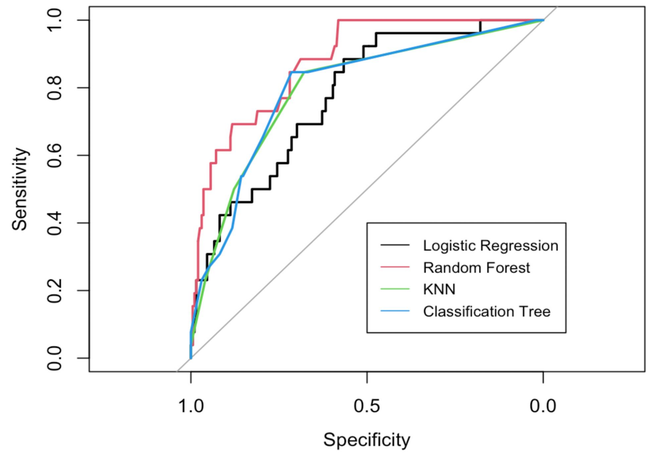
- The data was split into a training and test set. 70% of the data was used for the training model.
- Using best subset selection with AIC, the predictors `exports`, `concentration` and `fractionalization` were removed.
- The classification models take into account the fact that the classes are unbalanced and are designed so that the models classify instances of both classes as efficiently as possible at the same time

Classifier	Area Under ROC Curve	Misclassification Rate
Logistic Regression	0.776	0.401
Random Forest	0.878	0.369
K Nearest Neighbors	0.796	0.302
Classification Tree	0.798	0.270
Logistic Regression with Best Subset Selection	0.778	0.329

- From the table above, we find that Classification Tree model is the most accurate since it has the lowest misclassification rate. We illustrate the performance of the model below.

		Actual	
		`NO`	`YES`
Predicted	`NO`	140	4
	`YES`	56	22

Classifications from Classifier Tree Model



ROC Curve for each model

Conclusion

- The factors that are most important in predicting the presence or absence of a civil war in a country are the school enrollment rate of its males, its annual GDP growth rate, the natural logarithm of its population and whether one ethnic group dominates the country. A low school enrollment rate, low annual GDP growth rate, a high population and the dominance of a single ethnic group can all lead to an ongoing civil war.
- If we were to predict `NO` for every war, we would obtain an accuracy of 86.4% but a 0% accuracy in cases where a war actually occurs. Since our goal is to be able to predict wars, the Classification Tree model sacrifices accuracy (86.4% to 73%) in order to achieve the ability to predict actual wars. Thus, we designed a more conservative model that has an 84.6% accuracy when predicting the occurrence of a civil war when a war actually occurred.