# Civil War Prediction using Statistical Learning Methods

*By: Brandon Wallace, Gautami Kant, Ishita Malhotra, Phu Hai Le, Saksham Sarwari, Zulkifli Palinrungi*

## Background

Large-scale civil conflicts can cause unprecedented loss in a country. Accurately predicting the occurrence of civil conflict can reduce economic and social costs. In 2017, the global economic impact of violence was estimated at 12.4% of world GDP, in addition to loss of life [1]. Socio-economic data may be informative for predicting conflict. In this poster, we attempt to learn an association (if it exists) between measures of schooling, exports, population, etc., and whether a civil war was occuring at the point in time when the data was gathered.

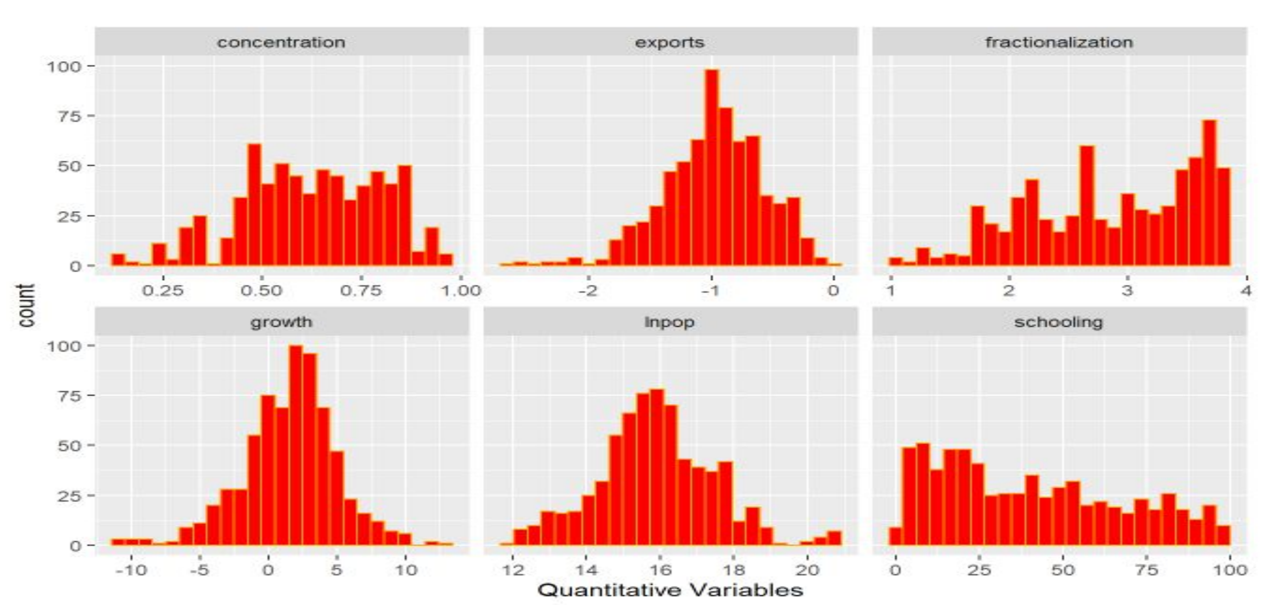
## Data Summary

The dataset consists of 741 observations and seven predictor variables as shown below. The column `civil.war` is the target variable.

| Name | Description | Variable Type |
|---|---|---|
| Exports | A measure of the dependence of a country on commodity exports | Numerical |
| Schooling | Percentage (school enrollment rate for males) | Numerical |
| Growth | Annual GDP growth rate | Numerical |
| Concentration | Population concentration (from 0 to 1, all in one city) | Numerical |
| lnpop | Natural logarithm of a country's population | Numerical |
| Fractionalization | Index measuring divides on ethnic/religious lines | Integer |
| Dominance | YES if one ethnic group dominates the country, NO otherwise | Factor |

Data processing and exploratory data analysis:

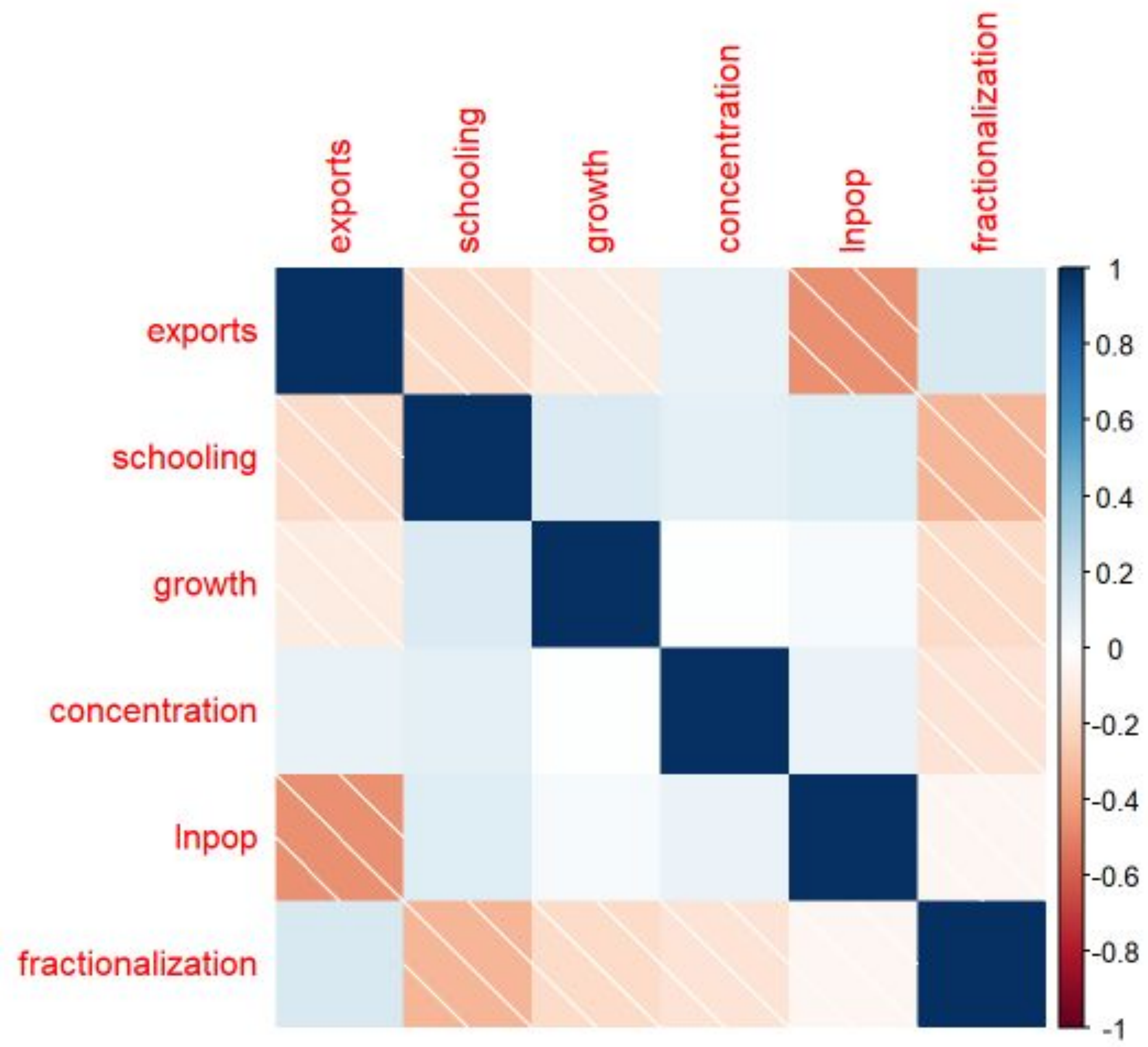
- `Schooling` has values in the form of percentages. Values greater than 100 have been removed since they could potentially be typos or errors
- Removed outliers from concentration (values < 0.1)
- Applied log transformation to two columns `exports` and `fractionalization` to make skewed distributions normal




**Variables distribution after cleaning and transformation**

**Correlation analysis** shows that:

1. High multicollinearity in any of the variables is not observed.
2. Dependency on commodity exports and a country's population are moderately (positively) associated with each other.
3. The risk of civil war can be related to such economic factors, as implied by the correlation between `exports` and `lnpop` variables.




**Correlation Analysis**

The threshold used to map probabilities to classes is 14.4%.

## Methods

This project utilized four binary classification models to evaluate the objective: Logistic Regression, Best Subset GLM, Decision Tree, Random Forest. 80% of the data was used for training and 20% was used for testing the models.

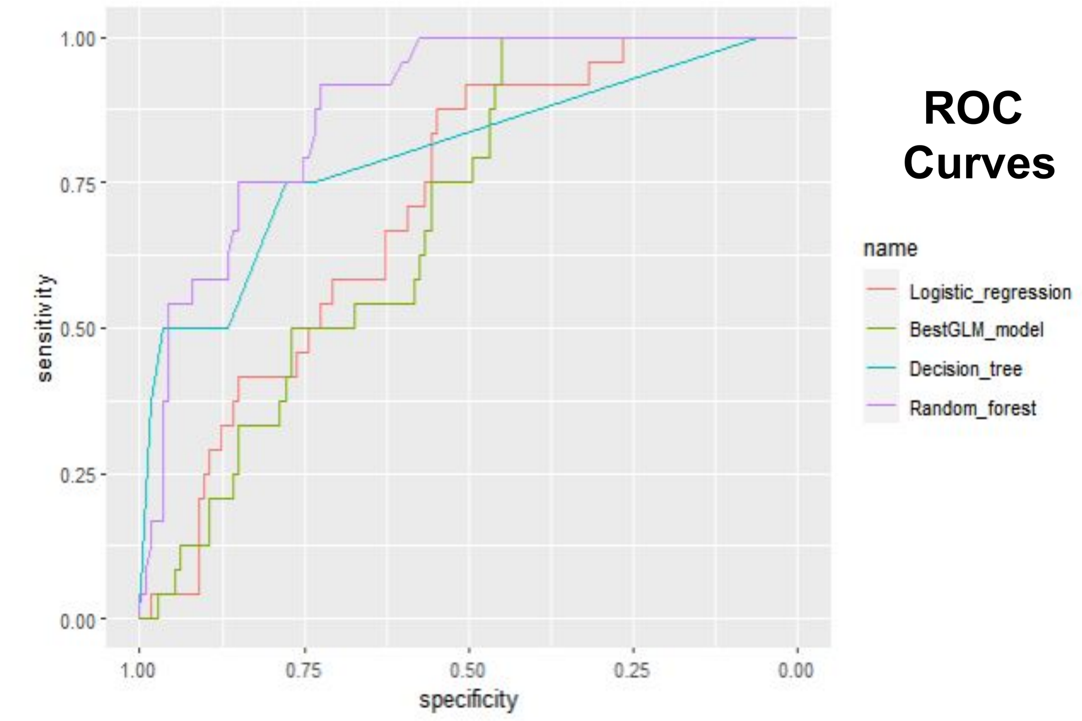
Model evaluation: Misclassification Rate, Area under the Curve, Accuracy, and Recall were used as evaluation criteria for the best model.

## Results and Analysis

The table below shows results for the different models that were used for prediction and performance metrics for the test data set.

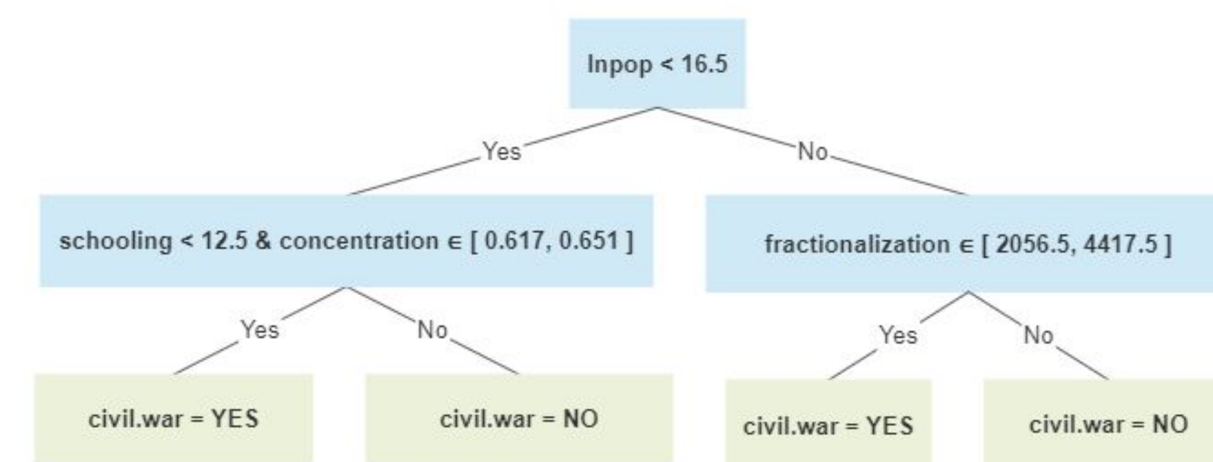
| Model | MCR | AUC | Accuracy | Recall |
|---|---|---|---|---|
| Logistic Regression | 0.343 | 0.713 | 0.657 | 0.583 |
| Best Model | 0.35 | 0.692 | 0.65 | 0.542 |
| Decision Tree | 0.226 | 0.798 | 0.774 | 0.750 |
| Random Forest | 0.233 | 0.879 | 0.766 | 0.750 |

**Model performance metrics**




**ROC Curves**

- With regard to AUC, Random forest has the best performance.
- From accuracy and misclassification rates (MCR), the decision tree performs slightly better than Random Forest and thereby outperforms the other 3 models.
- Random forest and Decision tree are similar in performance in terms of recall.




**Decision Tree**

The depth 2 tree gives a good training accuracy (90%). It is easily interpretable and gives us an idea of important features which are key in the decision making.

| | Actual Labels | |
|---|---|---|
| Prediction | NO | YES |
| NO | 88 | 6 |
| YES | 25 | 18 |

**Decision Tree - Confusion matrix**

| | Actual Labels | |
|---|---|---|
| Prediction | NO | YES |
| NO | 87 | 6 |
| YES | 26 | 18 |

**Random Forest - Confusion matrix**

Decision tree predicts one additional record correctly, when compared to Random Forest. That is why accuracy and misclassification rates are varying slightly.

## Conclusion

- Overall, random forest is the best model
- The population, its urban concentration, and the degree to which a population is fractionalized across religious or ethnic identities were the best predictors for identifying a civil war
- This research was limited to a selection of variables which may not fully capture the drivers of civil war. Additional predictors such as country, discrimination, external funding to non-state actors, and poverty could be introduced for future analysis.

## References

1. *Prevention Is Better Than Cure: Machine Learning Approach to Conflict Prediction in Sub-Saharan Africa*