Predicting the Occurrence of Civil Wars

By: Edward Piechowicz, Jonathan Taylor, Hwankyu Song

36-600 - Overview of Statistical Learning and Modeling

Introduction

The dataset contains information about whether a civil war was occurring at a particular time in a particular country, as well as pertinent data of potential importance.

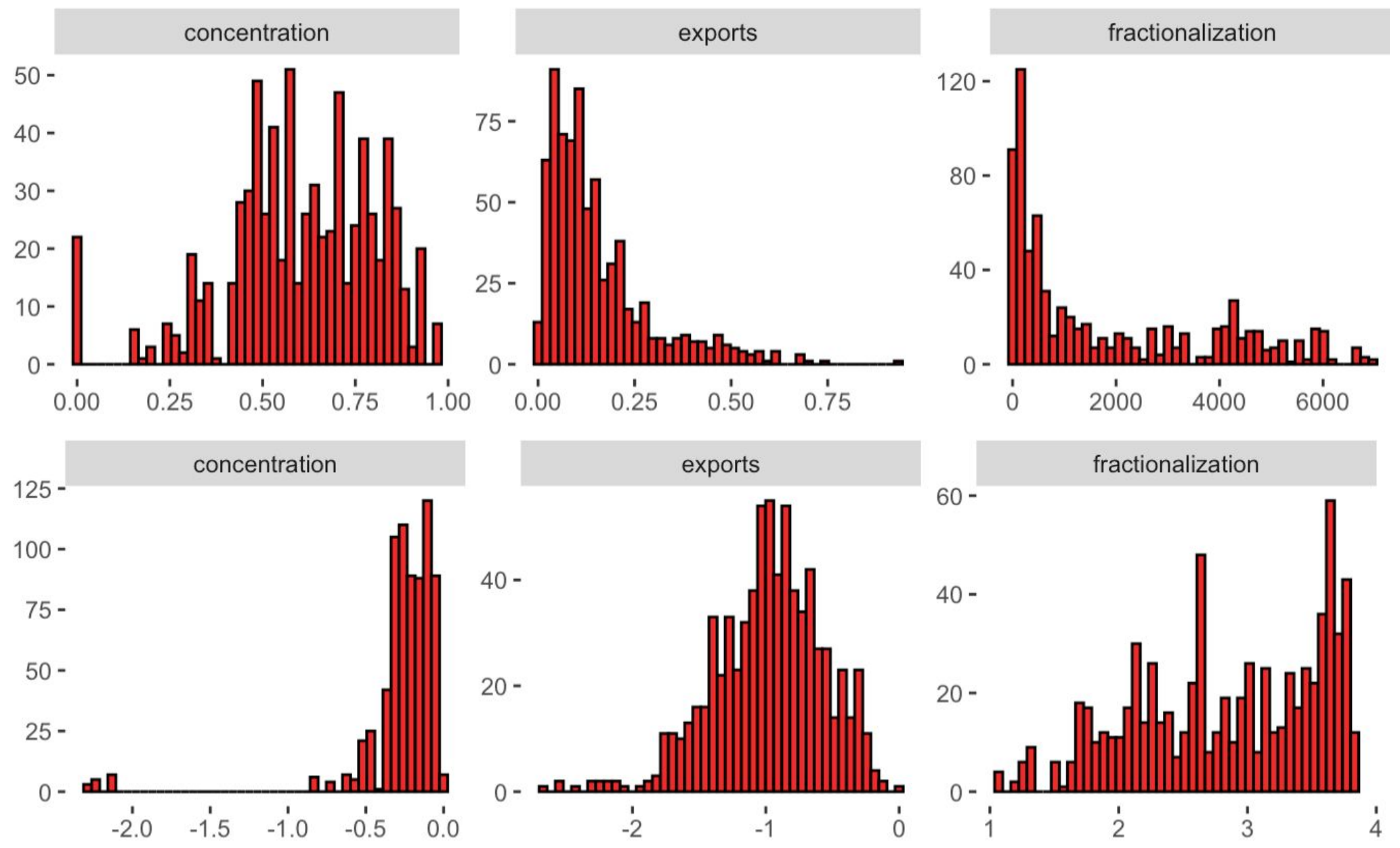
The goal of the project is to learn possible associations between factors such as population, schooling and growth, etc. and whether a civil war was occuring at the point in time in which the data were gathered.

Data

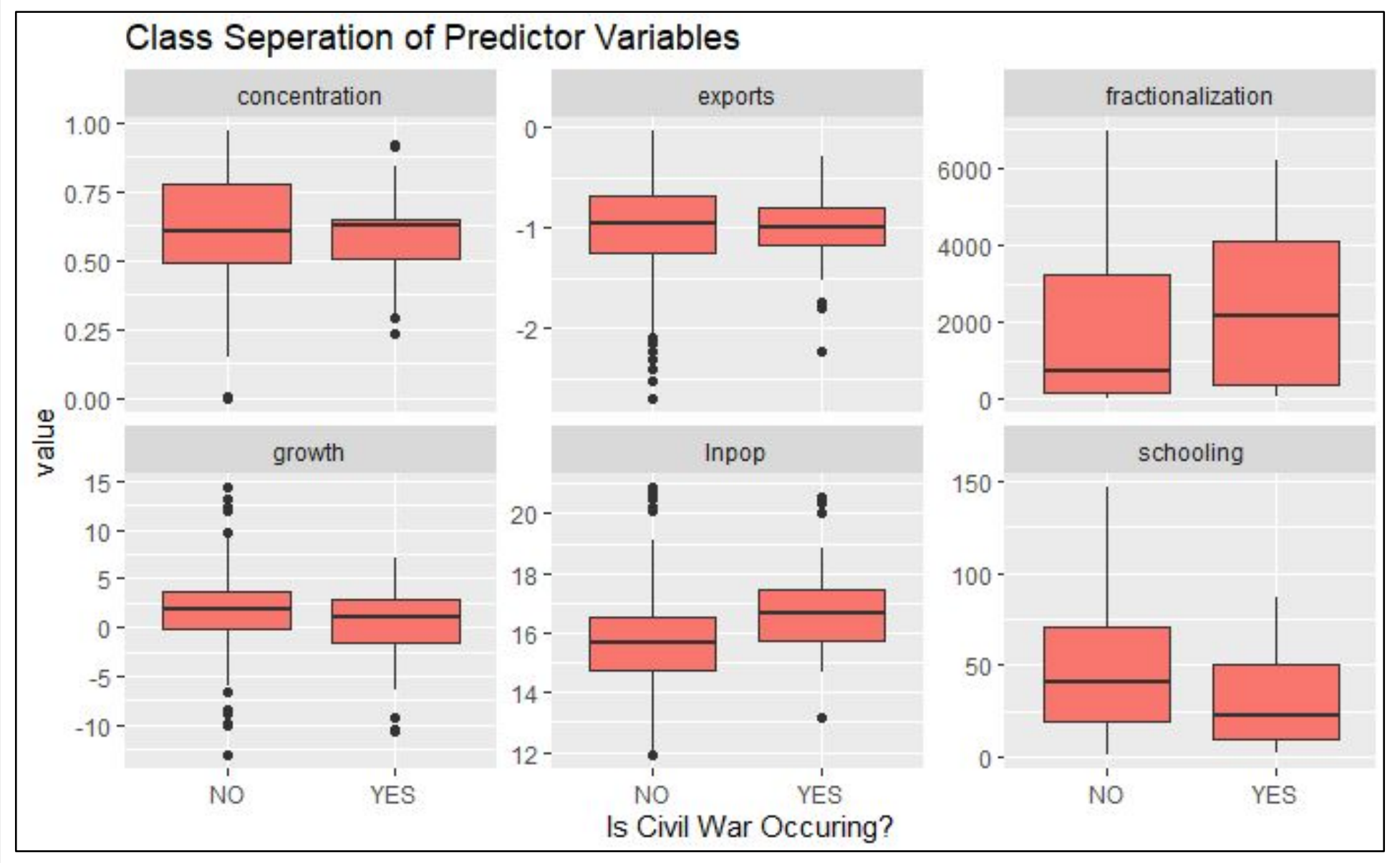
Our dataset consists of 741 instances and 8 variables. The class separation is very imbalanced with 642 no's and 99 yes's in the response variable.

Variables that appear right-skewed and have positive minimum values are considered for log-transformation. Exports is more symmetric post-transformation, so its column in the dataset was replaced.

Right skewed variables before (top) and after (bottom) transformation

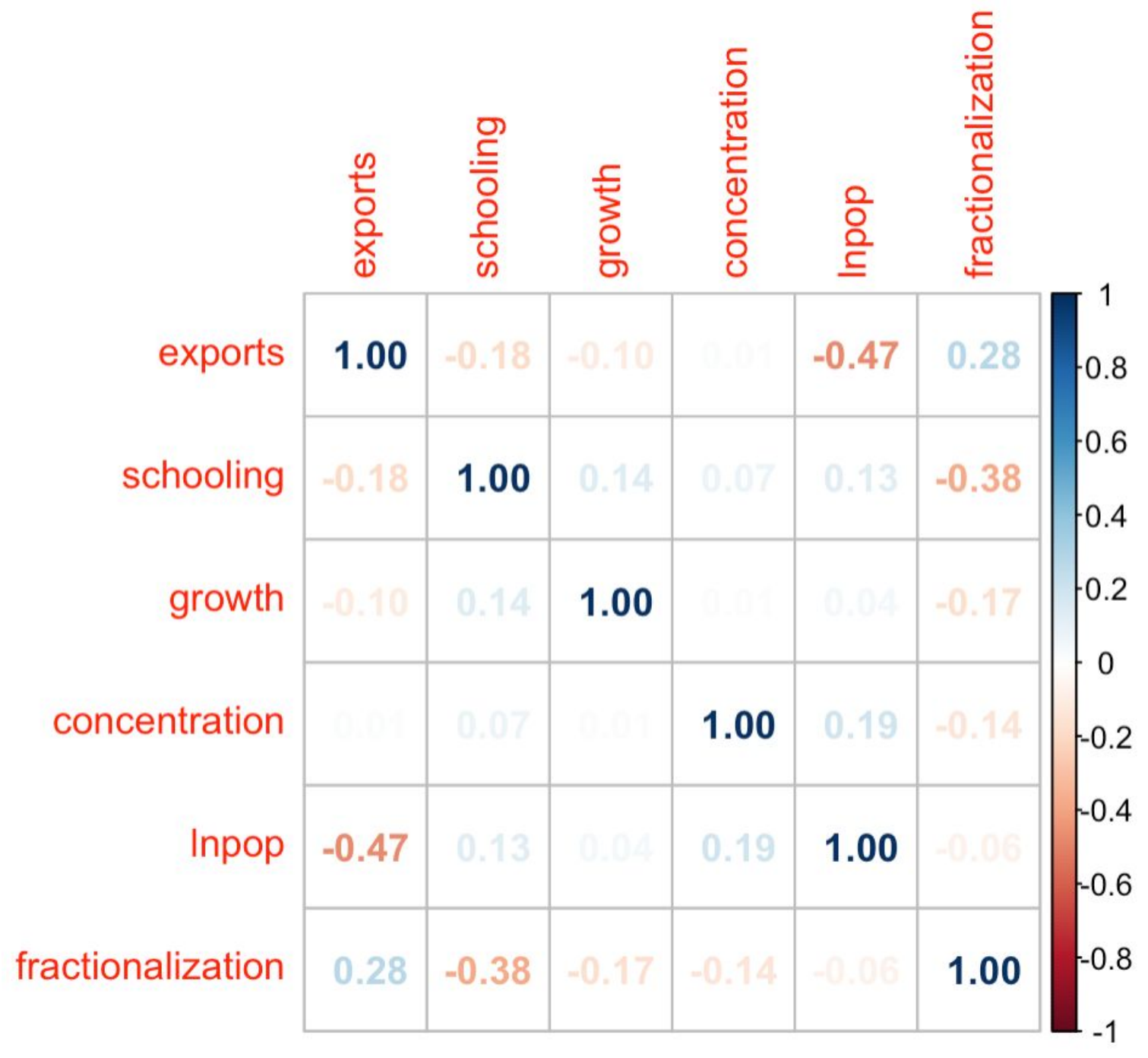


Box plot showing association between predictor and response variables



It can be seen that the variables are not heavily skewed and do not contain major outliers.

Correlation plot



There is some correlation between Inpop and exports (-0.47) and fractionalization and schooling (-0.38). VIFs using logistic regression were calculated, and confirmed that multicollinearity is not present.

Models

As the response variable `civil.war` is binary, logistic regression, classification tree, random forest, best GLM were all tested. The AUC values for all models were calculated, and random forest was found to be best. Random forest's class separation was then optimized by maximizing Youden's J statistic (specificity + sensitivity -1), and then feature importance was calculated.

Analysis

Table of AUC Values for Each Model

	AUC
Logistic	0.729
Classification Tree	0.802
Random Forest	0.889
Logistic with Variable Selection	0.673

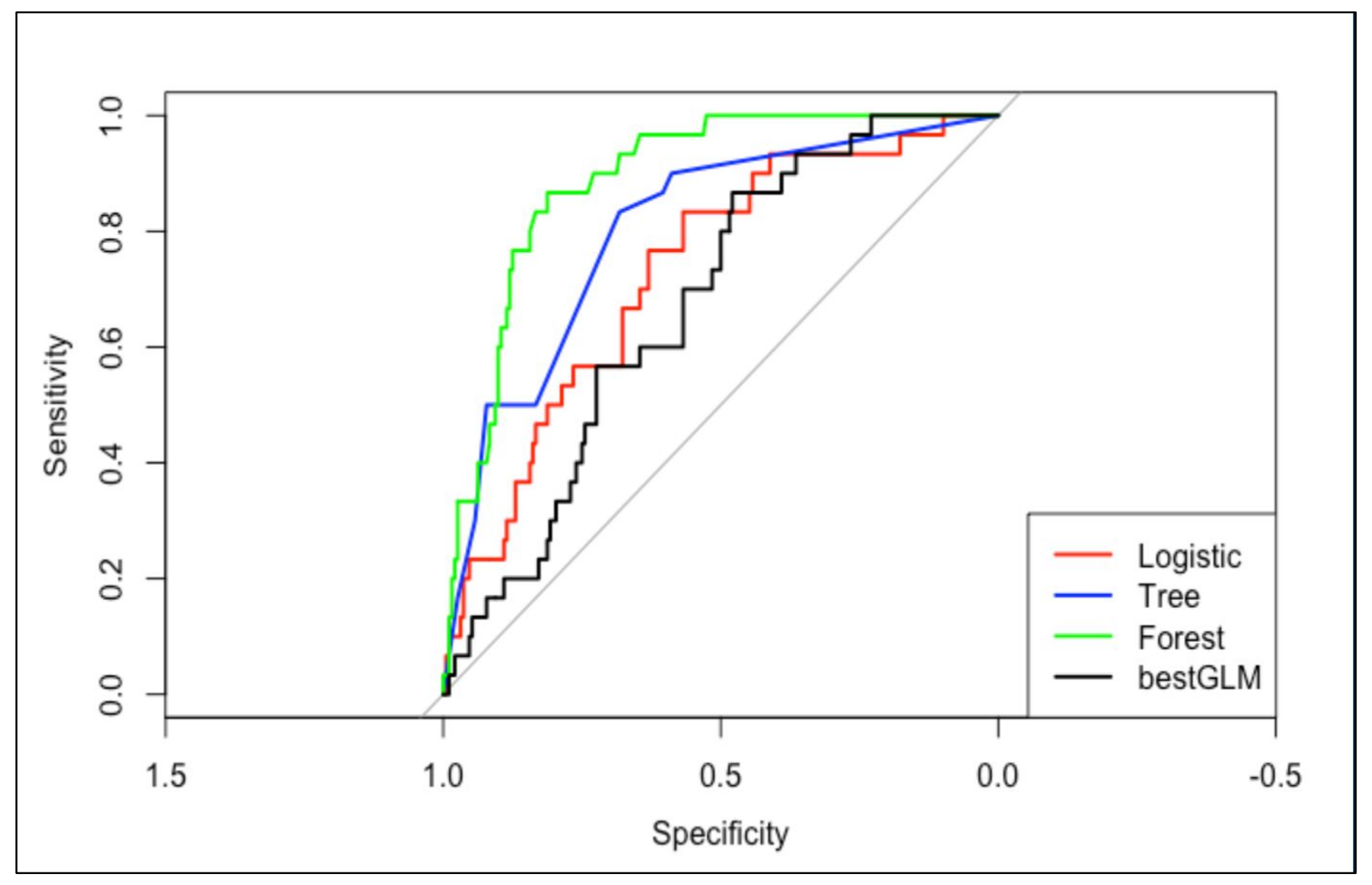
- AUC values were calculated for all models. Random forest had the highest AUC value.

Confusion Matrix for Optimal Random Forest Model

	Actual = NO	Actual = YES
Predict = NO	155	4
Predict = YES	37	26

- The optimized randomized forest did a better job at classifying both classes. The MCR for this model was 0.185.

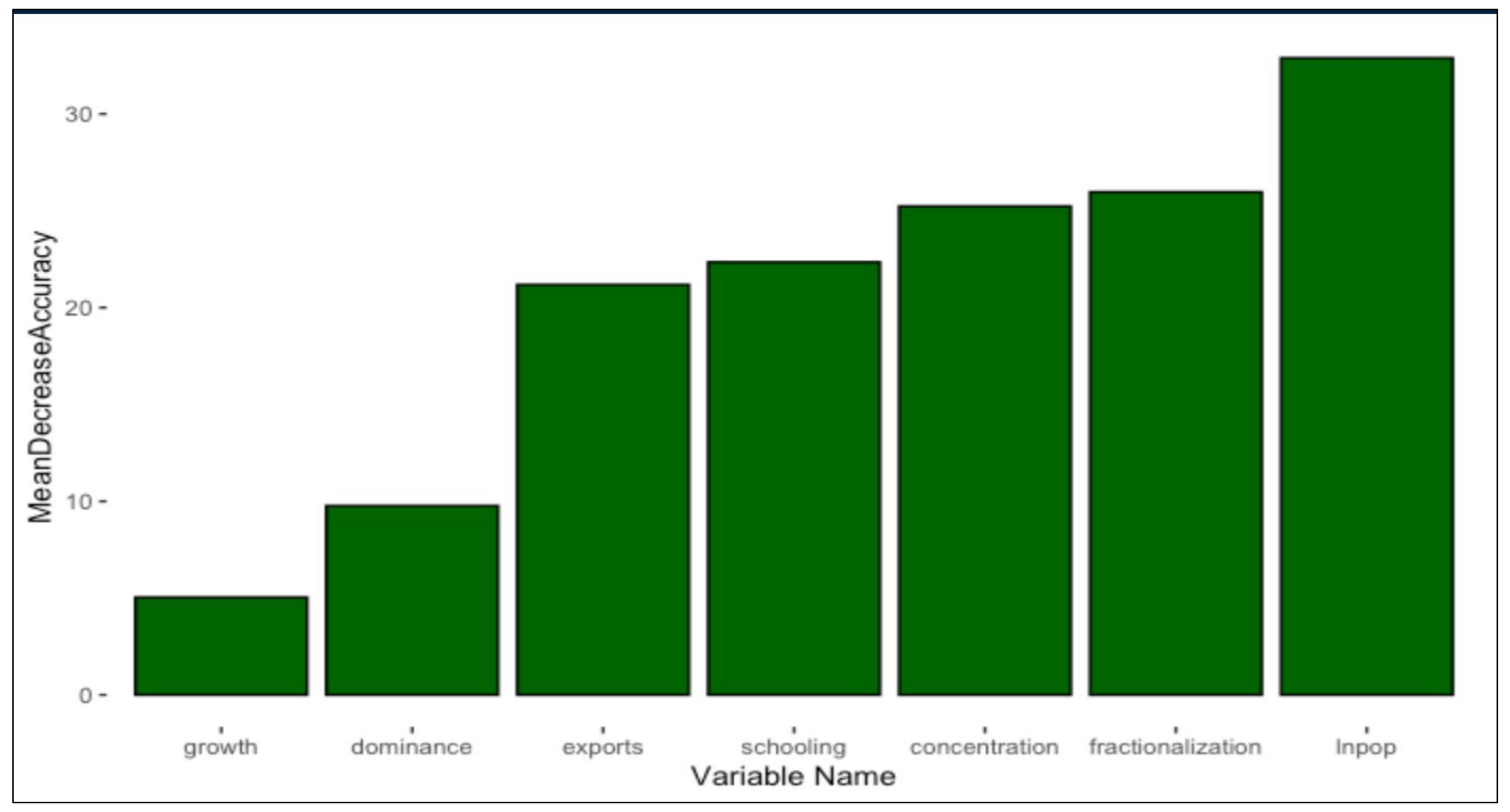
ROC Curves for Each Model



- Random Forest had the best ROC Curve among all the models.

Conclusion

Feature Importance using Mean Decrease Accuracy



- MDA is a measure of how much accuracy is lost by excluding each variable.

Higher MDA values indicate higher variable importance. The MDA graph shows that `Inpop` is the most important variable. The remaining variables have somewhat equal importance, with the exception of dominance and growth, which are substantially less important.

This project shows how statistical and machine learning models can be applied to learn what causes civil wars and therefore, how they can be foreseen.