**World Development Indicators to Predict Armed Conflict**

Ethan Engel

University of Denver

COMP4448 – Data Science Tools II

Dr. Neba Nfonsang

18 March 2023

**World Development Indicators to Predict Armed Conflict**

**Purpose**

The World Bank's WDIs (World Development Indicators) will be used to predict whether an armed conflict will occur in a country in a specific year.

**Significance**

According to the United Nations’ unstats.org, one quarter of the world’s population is affected by armed conflict, and in May of 2022, 100 million people on earth were displaced because of such conflicts and/or persecution. If armed conflict could be more consistently predicted, interventions could be initiated by the global community to prevent as much human death and misery as possible.

**Research Question**

Which binary classification algorithm will best predict armed conflict in a nation in a specific year, given several world development indicator statistics? The algorithm candidates are: Logistic Regression, Random Forest, K-Nearest Neighbors, Naïve Bayes and XGBoost.

**Datasets**

The data for the project comes from two different sources. The attribute data is provided by the World Bank. As there are 1441 WDI’s recognized by the World Bank, some attributes need to be chosen prior to the beginning of building models. Nineteen will be chosen at first. Several of these are: percent of a nation’s population living in rural areas (in a given year), percent of a nation with access to electricity, life expectancy, and food insecurity index. The label data is provided by the Peace Research Institute of Oslo (PRIO). It is a list of every armed conflict on Earth beginning in 1989. Each instance will be a nation-year, such as Myanmar2010 or Mozambique2015. If there was an armed conflict in that nation during that year, the label is coded as 1, if not, it is coded as 0.

**Data Preprocessing**

**Data Preparation**

Since the attribute and label data come from different sources, the csv files needed to be merged. As they are structured differently, this was especially time consuming. The WDI attribute variables were renamed to make them less wordy. Finally, many of the attribute columns were especially sparse, so they were either dropped or had null values replaced with imputed values. The imputations were based upon distributions from the exploratory analysis.

**Exploratory Data Analysis and Visualization**

The remaining attributes were electricity access, GDP per capita, birth rate, life expectancy, and percent of the population living in rural areas, all quantitative and continuous. Descriptive statistics were examined, as well as quantile plots and histograms, to see whether the distributions of the columns were approximately Gaussian. Since all the attributes were heavily skewed, they were imputed with median values and scaled with the MinMax Scaler. A correlation matrix was generated to check for collinearity. Some of the attributes were somewhat collinear, but not to the point of invalidating a model.

**Data Splitting**

A three-way train-validate-test split was employed with a 60-20-20 ratio. The purpose for this scheme was to eliminate certain model candidates after the validation stage, prior to parameter tuning.

**Model Building and Evaluation**

**Model Building**

Base model objects were instantiated for the five candidates. Confusion matrices were generated for each candidate’s training and validation sets, as well as the classification reports with the associated metrics. 92.7% of the label data was coded as 0. Therefore, this was the minimum threshold accuracy for a model being at all productive. Particular attention was paid to recall, as false negatives, predicting no conflict when one arises, would be especially costly.

Graphical user interface, text

Description automatically generated

**Model Optimization and Selection**

During the training-validation stage, the Random Forest and XGBoost models performed significantly better than the others by all metrics. These two had their parameters tuned for the final test stage.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Model Comparison**

While resulting metrics were comparable for the final two models, the tuned Random Forest produced accuracy, recall and F1 values that were slightly higher.

|  |  |  |
| --- | --- | --- |
|  | **Random Forest** | **XGBoost** |
| **Test Set Confusion Matrices and Accuracies** |  |  |
| **Test Set Classification Reports** |  |  |

**Conclusion**

**Lessons Learned**

Using the remaining five attribute variables, it was possible to construct productive binary classification models which predict armed conflict in a nation during a given year. They can yield weighted recall values of .96 and accuracy scores greater than .927, which is the proportion of label class 0.

**Recommendations**

Since all the classification metrics for the tuned Random Forest model were the metrics for the tuned XGBoost model, it will be used going forward. It could be used by the World Bank, United Nations, Doctors Without Borders or other global philanthropic organizations. As the WDI datasets become more complete, the model should be retrained to improve predictive power.