Enhancing civil war onset-prediction using terrorism data and discrete-time derivatives of features

Grégoire Molas, Pål F. Austnes, Jean Maillat *

December 18, 2020

Abstract

It is important to write a nice abstract.

1 Introduction

Studies predicting civil war onset have made use of different features to create the best model. Traditional models using logistic regression show relatively poor performance. [?][?][?]. More modern approaches has shown to considerably increase prediction-power using Random Forest algorithms. [?] However, the papers show that the important parameters for prediction remain similar across different methods. GDP per capita, GDP-growth, population-size etc. are some of the important features resulting from logistic regression and Random Forest classifiers. In this work, the authors aims to study if the performance of the models can be enhanced by addition of terrorism-data and transformation of some key features to exploit their temporal derivative.

2 Related work

Maybe move related work here.

3 Datasets

3.1 Civil war dataset

The civil war dataset (CWD) is stored in a csv-format file called SambanisImp.csv. It contains data from 1945 to 2000 for 157 countries. In Muchlinski et al. 91 features from this dataset was considered when training models. This article considers the same 91 features.

3.2 Terrorism dataset

The terrorism dataset is provided by the Global Terrorism Database (GTD)[?]. It contains data on more than 150,000 terrorist attacks between 1970 and 2019. The data has 135 columns of parameters per event and contains as such, a multitude of potential research items. In this study, only the total amount of attacks per country, per year was considered to train the model.

4 Methods

Random forest is a set of machine learning algorithms to construct a forest of decision trees for classification and regression problems. They differ from traditional decision-forest in their capability to limit overfitting by randomly constructing multiple instances of sub-forests of the feature-space and combining them into one model. This limits the variance by accepting a higher bias, resulting in a better generalization [?]. The implementation of this algorithm is done in Scikit-Learn [?]. To evaluate its performance, two methods are considered:

*All authors at: École Polytechnique Fédérale de Lausanne

4.1 Model performance using Receiver Operating Characteristics

Receiver operating characteristics (ROC-curves) is a two-dimensional plot showing the correlation between the false positive rate: $FPR = \frac{FP}{FP+TN}$ with FP: false positive and TN: true negative and the true positive rate: $TPR = \frac{TP}{TP+FN}$ with TP: true positive and FN: false negative. In other words, the ROC-curve shows how good the model is at choosing the correct label. The main parameter studied here is the area under the curve (AUC). An AUC-score of 1 means the model perfectly classifies the correct labels, while 0 means it perfectly classifies the wrong labels. An AUC-score of 0.5 means the model does not perform better than chance.

4.2 Feature importance using Gini-impurity

To assess the importance of each feature, the Gini impurity is calculated for the whole tree. It is defined as the sum of the total decrease in node impurity for a feature, where the impurity for a single node is defined as:

$$G_{node\ k} = 1 - \sum_{i=1}^{N} p^2(i|k) \tag{1}$$

Where $p(i|k)$ is the probability of an outcome i with feature k.

4.3 Discrete-time derivative

Data-wrangling was performed to study the impact of temporal evolution of the features in the dataset. This was done by calculating a new set of features which corresponds to the discrete-time derivative of the original features. The procedure is summarized in the following equation.

$$X_{ij,evol} = \frac{X_{ij} - X_{(i-1)j}}{X_{(i-1)j}} \tag{2}$$

With $i \in \{\text{nb years}\}$ and $j \in \{\text{nb features}\}$

Exceptions: -First year value is zero -If previous year value is zero we consider a big gain of 10 in the direction of the current year.

5 Results

5.1 Comparing Gradient Boost and Random Forest classifier

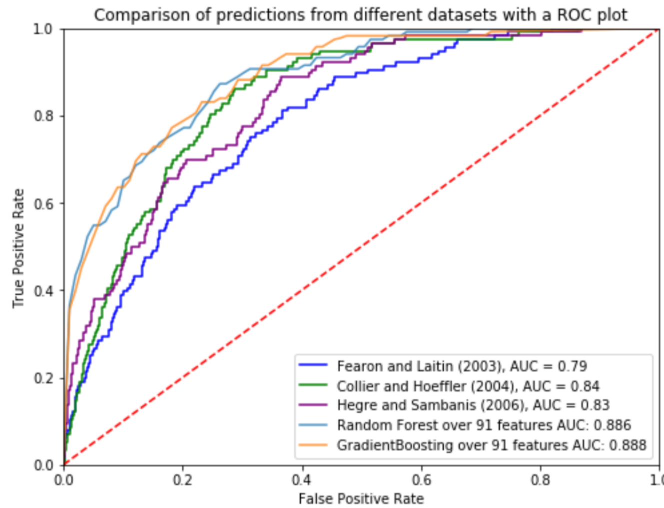

Figure 1: Comparing Gradient Boost and Random Forest classifier on predicting civil war onset. For comparison, three models using logistic regression [?][?][?]

The Gradient Boost and Random Forest classifiers show similar performance. Comparing their AUC-scores, they show considerable better prediction-power compared to the logistic regression models.

5.2 Compare GINI-scores of random forest and Gradient boosting

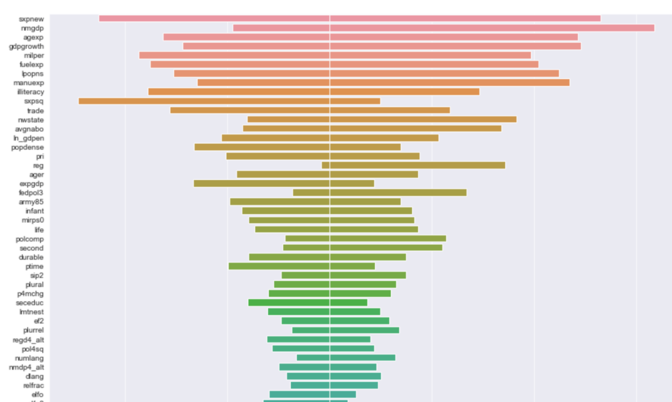

Figure 2: Comparing the importance of features. Left: Random Forest. Right: Gradient Boost

A similar importance of features is observed, consistent with the prediction and the AUC-score.

5.3 Data-wrangling on terrorism dataset

The terrorism dataset contains one event per row. To adapt it to the needs of this study, the number of terrorist attacks per year per country was calculated and stored in a single column. Then, the column was merged into the Sambanis CWD.

5.4 Evaluation of the terrorism-feature

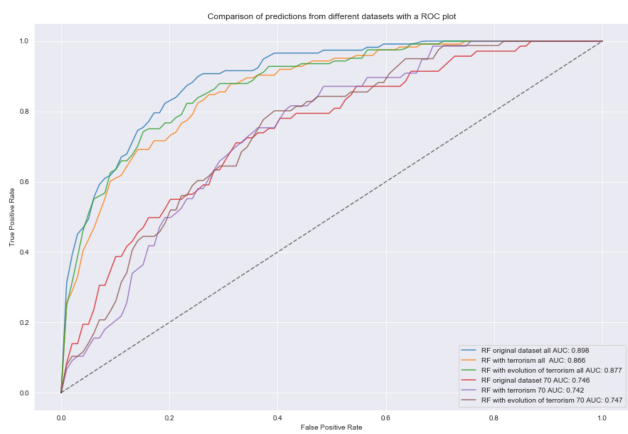

Figure 3: Comparing the ROC-curves for different classifiers with and without terrorism feature

Since the terrorism data only spans from 1970-2000, two models were trained. One, using all the data from 1945 (setting all values of terrorism before 1970 to 0), and one using only data from 1970-2000. It is observed that the model using data from 1945 in generally performs much better. There is no significant difference between the models with and without the terrorism feature.

5.5 Effects of slicing the data into 5-year batches

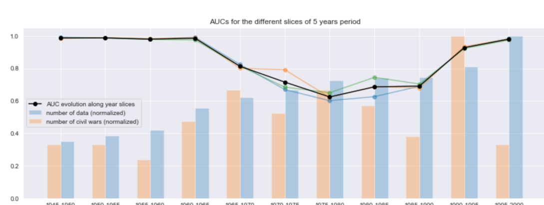

Figure 4: Comparing the AUC-scores for 5-year slices of the data

Performing temporal splicing on the data helps evaluating if time and evolution of features is relevant for the prediction-power. For this reason, the data was split into 5-year batches and evaluated independently. The plot shows the AUC-score of each model on the 5-year batch. A significant dip in the scores is observed between 1970-1990. A similar dip is observed when considering 11-year slices. Further studies has determined that the reason for this is neither the dataset size, nor the relative occurence of civil war, as the ratio of the labels remain relatively constant [1].

5.6 Discrete-time derivative of features (evolution)

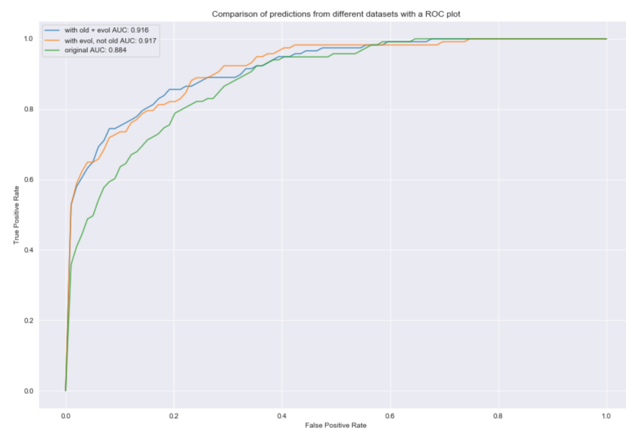

Figure 5: Comparing the ROC-curves for original data and evolution data

The results in 5.5 led to the development of a temporal modification of features to study possible improvements to the predictions. A discrete-time derivative of key features was performed to assess the predicting-power.

Figure 6 show that the addition of the new features makes the model stronger. In both cases, the reported AUC score is around 0.92, compared to 0.88 for the original model. There is no signifcant difference between adding the new features or replacing them with the corresponding ones.

5.7 Ranking gain of adding discrete-time evolution features

It is observed that several of the new features are important for the classifier. However, for certain features there is no gain in adding them to the model.

[1]Details of calculations are provided in the Jupyter Notebook

5.8 Next steps

The results show that there is no gain between using gradient boost above random forest classifiers. The terrorism feature shows to have a medium importance among the 91 features in the model. The performance is not considerably improved by using the temporal evolution of terrorist attacks. In conclusion, the terrorism data might prove useful in civil war onset prediction.

The most promising results concerns the temporal evolution of the features. Indeed, different ways of calculating the discrete-time derivative considering different weights would be an interesting continuation.

6 Discussion

7 Conclusion

References

2

8 Annex

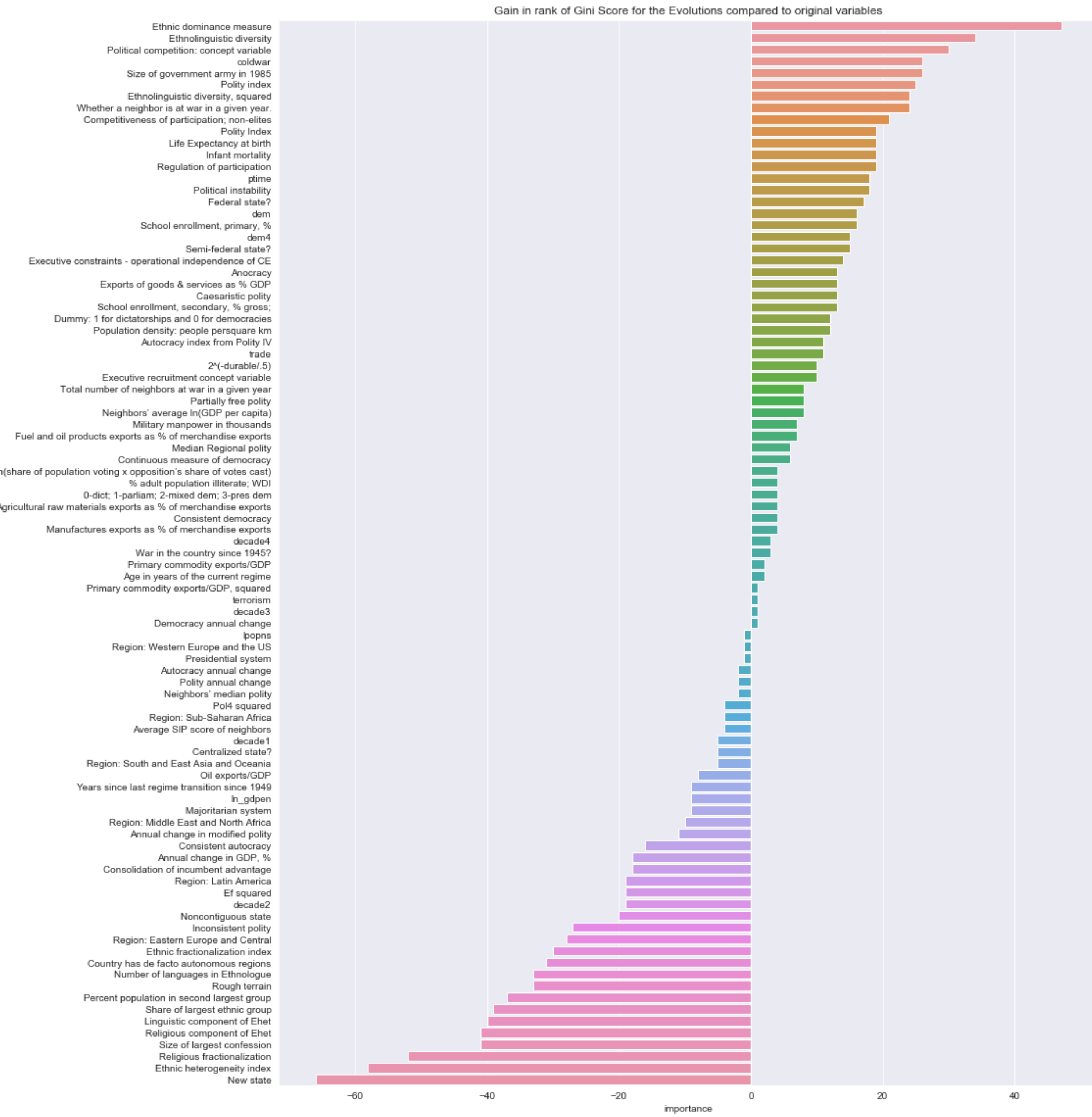

Figure 6: Comparing the ROC-curves for original data and evolution data