Forecasting Social Unrest: A Machine Learning Approach

Redl, Hlatshwayo (2021)

Discussion by Martin Baumgärtner

THM Business School

2021-11-01

Overview

> "(...) can we reliably predict outbreaks of social unrest before they occur? Of course, many factors may potentially predict social unrest including economic, financial, or other social variables. We leave investigation of these questions for other authors"
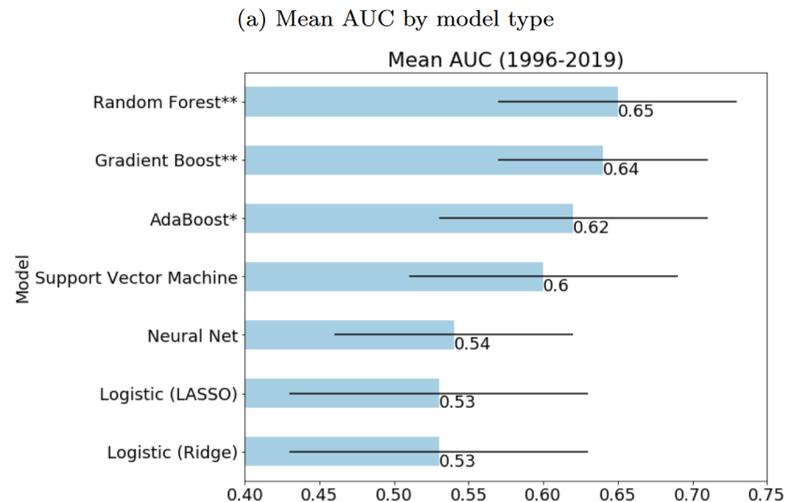>
> (Barrett et al. (2020) p.34)

Data Science Perspective

Aim: Predict RSUI with

- 340 features
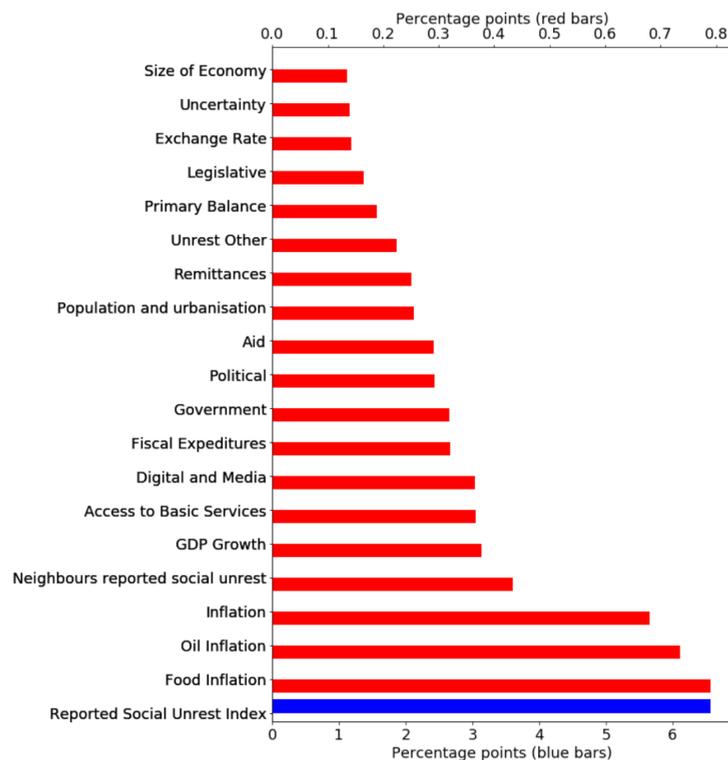- 125 countries x 29 years = 3625 observations?
- 7 Algorithms

Evaluation

(a) Mean AUC by model type

Redl, Hlatshwayo (2021) p.10

Shapley Values

(a) Average Shapley value contribution, 1996-2019

Results and Open Questions

- Predictors are in line with the literature
    - Social unrest
    - (Oil-/Food-) Inflation
    - Unrest in neighbour countries

- Does the model find deviations from the literature?
- Which features are not relevant although they are important in the literature?

Interactions

> "however, it is likely that the drivers of unrest are disparate and interact with socioeconomic conditions in complex and non-linear ways that are difficult to enumerate."
>
> p. 3

- random forest is most successful
- lasso and ridge are unsuccessful

➡ hint to non-linear connections

I would love to see more of this line of story!

> Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888 (2018).

Breakdown of digital and media

The data set ranges from 1990 to 2019

At the same time, media use has shifted from television to the Internet.

> Hypothesis: Media use has become a more important over time as the use is more interactive

Perhaps it is worth looking at the breakdown of SHAPE values in the digital and media category.

Index uncertainty?

Uncertainty in the RSUI index (~5% events mis-indentified)

- article refers to past events (8 of 31)
- article describes wrong country (12 of 31)

This uncertainty is not taken into account during the integration into subsequent models.

How do we deal with measurement errors in an index in subsequent models?

Summary of Comments

1) I like this paper!

2) It has an idea of how Machine Learning can approach the question of why.

3) I expect more similar applications of machine learning / feature selection in the next years

4) Some open questions:

- Interactions?
- What happens inside the Media cluster?
- How should you deal with index uncertainty?

(a) Area under receiver operating characteristic curve by year