Seven suggestions for developing predictive models

Philip A. Schrodt, Ph.D.
Parus Analytics LLC and Open Event Data Alliance
Charlottesville, Virginia USA
http://philipschrodt.org
https://openeventdata.org

Presented at the PREVIEW Prediction and Early Warning Workshop
German Federal Foreign Office, Berlin
29-30 January 2019

These will all be brief...

1. Simple models are good
2. [Mostly] Forecast at the monthly level with a 6 to 24 month horizon
3. Evaluate accuracy metrics and recalibration intervals on long time series
4. Data quality is more important than quantity: beware fake news and false positives
5. Data generation methods and software need to be transparent: make your models open source and open access
6. Utilize the software singularity to inexpensively experiment
7. Cooperate with the news media and create the "must-have app"

1. Simple models are good!

- This result goes back to the 1950s. For example, a recent study on predicting criminal recidivism showed equivalent results could be obtained from
  - A proprietary 137-variable black-box system costing $22,000 a year
  - Humans recruited from Mechanical Turk and provided with 7 variables
  - A two-variable statistical regression model
- For forecasting political conflict, there is a widely-recognized “speed limit” on predictive accuracy of around 80% and multiple methods can achieve this.
- Political Instability Task Force maintains a set of 2,700 variables but absolutely exhaustive evaluations show 20—if that—are sufficient for their models

Recidivism source: Science 359:6373 19 Jan 2018, pg. 263; the original research is reported in Science Advances 10.1126/sciadv.aao5580 (2018).

A few more thoughts on simple models

- Complex models—large numbers of independent variables—may be useful for *classification* but still not useful for *prediction*
  - This assumes classification and prediction problems are distinct: in classical statistical modeling this is unquestionably the case but it is less clear in machine learning
- Overfitting is a huge potential problem. But you know that
  - Also see later remarks on the availability of new long time-series event data sets
- Trigger models are probably a cognitive illusion: structural factors dominate.
  - Though I would be happy to be proven otherwise: this is an empirical question.
  - If you aren't seeking subtle triggers—fruit vendors self-immolating in obscure regional market towns—your data requirements are considerably reduced

2. Forecast at the monthly level with a 6 to 24 month horizon

- This is the general "policy relevant forecast interval"
- Surprisingly, most conflict forecasting models do not show a substantial decline in accuracy with increasing horizons: this is additional evidence for the importance of structural factors
- In some applications, short time intervals—6 to 24 *hours*—may be important, and social media may be useful here.
- Educating users to accept the limitations of quantitative models—which are substantially more accurate than all but the very best human forecasters but can't buy you a beer when they fail—is critical. This includes estimating and explaining accuracy "speed limits" imposed by working with open complex social systems.

3. Evaluate accuracy metrics and recalibration intervals on long time series

Two new long-time-series event data sets:

- TERRIER (Univ of Oklahoma): every Lexis-Nexis news article in English, Spanish and Arabic 1980-2015, obtained legally
- Cline Center (Univ of Illinois): *New York Times* (1945-2005) and BBC World Monitoring Summary of World Broadcasts (1979-2015)

  Both of these are currently coded with the less-than-optimal PETRARCH-2 coding systems but could be recoded in the future

Rare events prediction has a number of methodological issues not found in other common forecasting problems such as election and economic predictions

Determining the optimal frequency for re-estimating models is still poorly explored, and may depend on the type of model

4. Data quality is more important than quantity

- We have apparently passed the "inversion" where more information on the web is computer-generated than human-generated. Some of it is harmless; much of it is not
- The social media platforms clearly will not police themselves: anger, greed and delusion are their business model
- Various event data issues that need to be seriously addressed
    - False positive rates
    - Duplication rates
    - Urban and other geographical/socioeconomic biases
    - "Media fatigue," particularly in conflict zones
    - Biases in existing event data sets primarily focusing on violent conflict

5. Data generation methods and software need to be transparent: open source and open access

- Many new models have numerous hyperparameters which are often set in rather ad hoc fashions (or, at best, optimized for a specific set of data)
- In rare events analysis, continuity is critical: we need to be able to generate comparable data across many years and many different projects
- There is increasing concern about assessing the implicit biases in models
- We owe transparency to the people who actually are funding us, whether directly through taxes or indirectly through the extortionate oligopoly profits accruing to foundations and philanthropists

Academics: remember that publishing in paywalled journals is equivalent to burying your work beneath a pile of radioactive sludge in Antarctica

6. Utilize the software singularity to inexpensively experiment

- We don't need "one model to rule them all"
- Follow the approach of hurricane and snow forecasters who triangulate results of multiple independently developed models which have different assumptions and strengths
- Cloud-based machine learning systems (Google, Amazon, Microsoft) can now automatically evaluate a series of standard methods
    - You need to interpret these models skeptically and in light of existing theories, but with these tools now available you might as well experiment with them

The software singularity

- Code for doing almost anything you want is now available for free and has an effective support community on Stack Overflow: things that once took months now can be done in hours
  - `newspaper3k` downloads, formats and updates news scrapping in 20 lines of Python
  - universal dependency parses provide about 90% of the information required for event coding
  - easily deployed data visualization dashboards are now too numerous to track
- This enables development to be done very rapidly with small decentralized "remote" teams rather than the old model of large programming shops
  - In the software development community in Charlottesville, our CTO group focuses on this as the single greatest current opportunity, and doing it correctly is the single greatest challenge
- That other singularity?: no, sentient killer robots are not about to take over the world, and you're going to die someday. Sorry.

7. Cooperate with the news media and give us the "I've got to have one of those on my desk!" application

- Publicize your success stories
- Collaborate with the news media
    - kudos to ACLED, which is now regularly featured in *The Economist*
- Patiently educate policy-makers and other stakeholders (NGOS, IGOs) on understanding the strengths, weaknesses and appropriate applications of models
    - This is *really hard!!*—if you are having difficulties, it is not because you are stupid or they are stupid: statistical reasoning is just really hard, for everyone (including statisticians)
    - Investing is user-friendly dashboards/visualizations and thoroughly documenting how to use these is time well spent

Thank you

Email: schrodt735@gmail.com

Blog: asecondmouse.org

Slides: http://eventdata.parusanalytics.com/presentations.html

Links to open source software: https://github.com/openeventdata/

Links to open access data: http://openeventdata.org/datasets.html