



Lessons learned when applying ML in Off. Stat.

Why it helps to be a survey statistician and a data scientist!

Piet Daas, Marco Puts
Statistics Netherlands

6 June 2023

Introduction

- Machine Learning
 - Work data driven (inductive)
 - Goal is to find generalizable patterns in data
- Official statistics
 - Work theory driven (deductive)
 - From a general theory to more specific (hypothesis) tests
- Both 'ways of working' are needed when applying ML in official statistics!



How to start?

- A colleague asked “Could you help us in identifying online platform companies?”
 1. How to answer that question
 - a) What is an online platform?
 - OECD definition:
 - “a digital service that facilitates interactions between two or more distinct but interdependent sets of users (whether firms or individuals) who interact through the service via the *Internet*.”



How to start?

- A colleague asked “Could you help us in identifying online platform companies?”

1. How to answer this question

- a) What is an online platform? [there is a definition]
- b) A lot of is unknown
 - Where’s the difference between online platforms and others?
 - How well can one discern between both? (and what is good enough?)
 - What is the reason for asking this question?
 - Have others studied this topic already?
 - Is there (already) data available?


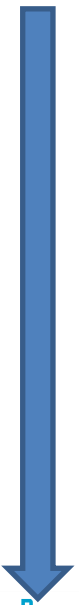



1. Start with a preliminary investigation!

- Test the idea
 - Is there data available? Use it!
 - Accept the fact that the data is not perfect (at the start)
- Luckily, a (first version of a) platform questionnaire had been answered by 1034 businesses
 - 926 (90%) had a website that could be scraped
 - 168 (18%) platforms, 197 (21%) non-platforms, 561 (61%) unknown
- Investigate if the website texts differ for plat. and non-plat.
 - 'Bag of words approach', SVM-model, accuracy of 73% (YES!!!)
 - Asked for a larger dataset (got more positive cases)



2. Retry with better & more data

- 
- Get more/better positive cases
 - *Risk* of expert bias (may miss unknown positive cases)
 - Get more/better negative cases
 - Usually a random sample of businesses in BR to which a website has been assigned
 - Check and remove any accidental included positive cases
 - What percentage of positives should be used for training?
 - Recommend somewhere between 20-50%
 - Build new models (iteratively)
 - Try different algorithms, vary %positives, different metrics, ...
 - Get feedback from experts
- 
- 

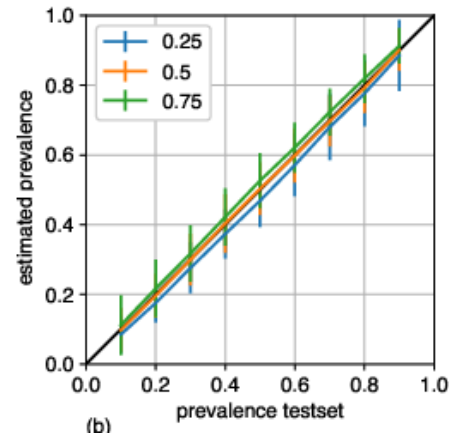
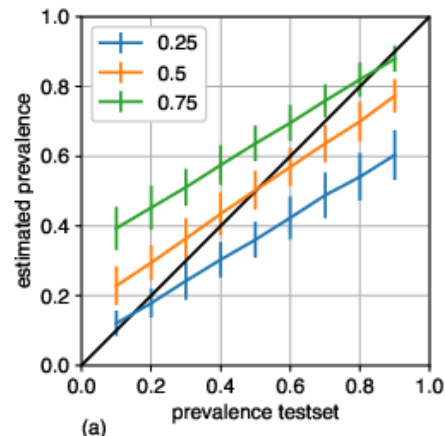
3. Internal and External validation

- Often a particular percentage of positive (and negative) cases is used for model development
- Usually an 80% random sample of this dataset is used to train the model. The performance of the model is determined on the (remaining, unseen) 20% test set. This is what we call the *Internal validation*.
- For official statistics the performance of the model on the target population is of interest. This is what we call the *External validation*.
 - What is the best way to obtain a model with a high external validity?
 - Currently looking at: Construct a representative training/test set with random samples, Iterative model development, New metrics, ...



4. Bias correction (model induced)

- Using a particular percentage of positive and negative examples in a training set, negatively affects the performance of the model
 - It introduces a *bias*
- A Bayesian correction method has been developed to correct for this
 - Code on: github.com/mputs/BayesCCal
 - Works really well!
 - Requires classification models that can produce 'probabilities'



Lessons learned so far

- Can ML assist in providing the answer?
 - Do a preliminary investigation (use what's available)
- Don't assume anything
- Preferably use random samples (of target population) for model development
 - But what about rare subpopulations? (such as platforms)
- Try to study the entire target population (census like approach)
- Deal with biases (various sources of bias exists)
 - Model induced, Expert bias, Scraped affected, Non-website part of pop, ...
 - We are (slowly) getting grip on these issues
- ...



Thank you

