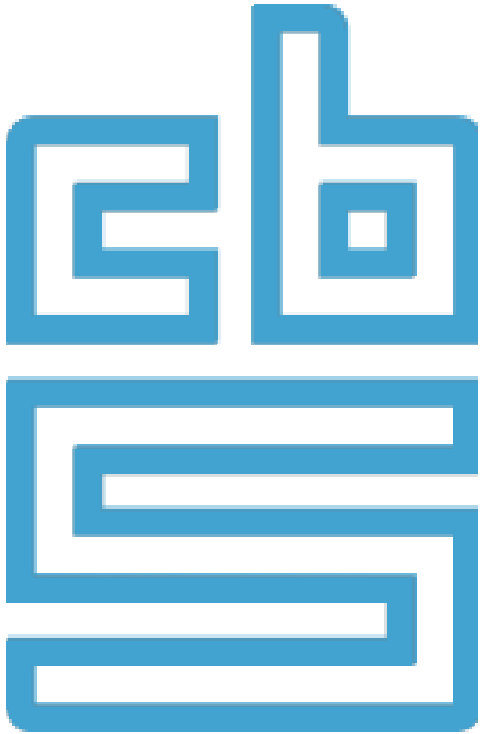


Machine learning in official statistics: towards statistical based machine learning

Marco Puts, Piet Daas
Statistics Netherlands

ISTAT workshop on methodologies for Official Statistics, December 7, 2023

Difference between official statistics and ML

Objectives

In official statistics:

Produce relevant, objective, and accurate statistics!

In Machine Learning (ML; data science in general):

Identify patterns in data using various statistical techniques

Methodology

- **Technique**: A way of carrying out a particular task.
- **Method**: a particular procedure for accomplishing or approaching something, especially systematic or established one.
- **Methodology**: A system of methods used in a particular area of study or activity.

Methodology

Claim

Methodology is not well defined for ML from a governmental point of view

Goal of this presentation

Illustrate the need for ML methodology

A case study: detecting Online Platforms

Online Platforms: websites

- Statistics Netherlands and Ministry of Economic Affairs want to produce statistics on Online Platforms

- What is an Online Platform?
  - A digital intermediary service facilitating interactions (and transactions) between two or more sets of users
    - Via a website (and app)

- Idea: 1) Identify Online Platform websites with ML
  2) Send a questionnaire to the businesses detected

Online Platforms: model development

- Asked experts for examples
    - 680 online platform websites
    - Only a few negative cases were given
    - Added a random sample of non-platform websites from Business Register
    - Created a balanced training set (50% pos., 50% neg.)
- Scraped and combined the text of multiple pages per website
    - up to 200 pages
- Trained a text-based ML model (simple bag-of-words approach)
    - Supervised classification task
    - Best result: Support Vector Machine, accuracy of 82%

Online Platforms: model evaluation

- Model gives the '*chance*' of being an online platform website
  - Value between 0 and 1 (but it is not a real probability)
  - U-shape distribution of test set (relative small set)

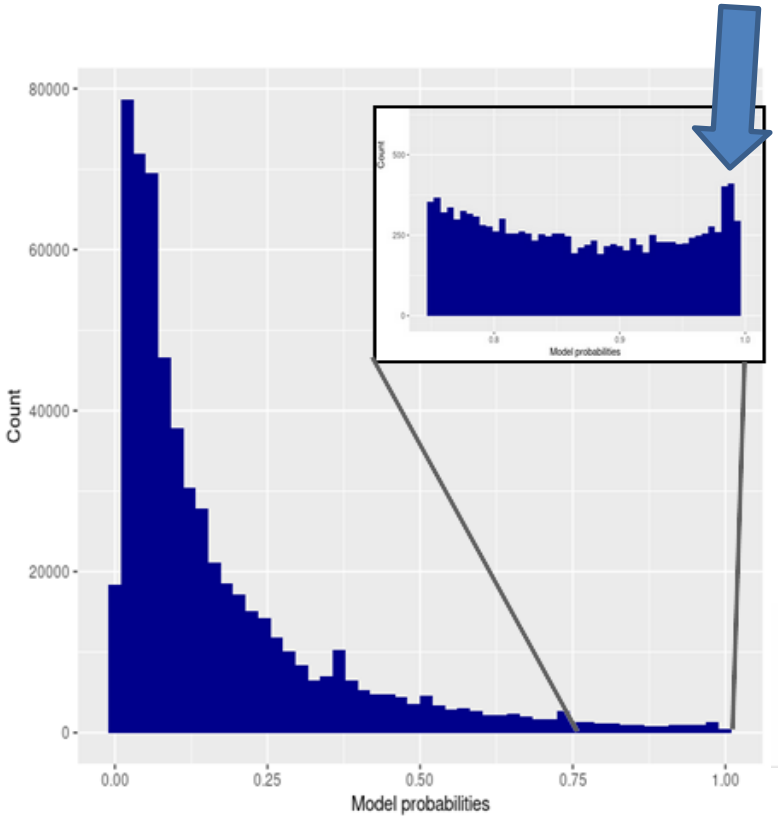
- Words positively associated with an online platform
  - **Register, login, *platform*, invest, sign up, …**
  - Negatively associated words are indicative for many other types of websites

Online Platforms: results on ‘real world data’

- **Results**

Totally scraped (with text)	629,284	(100%)
Score ≥ 0.5	41,881	(6.7%)
Score ≥ 0.8	9,387	(1.5%)
Questionnaire	4,385	(0.7%)
Response	2,997	(0.5%)
Online Platforms	±1,400	(0.22%)

- *Manually* checking samples in various score ranges revealed that 0.8 and higher contained online platforms
- After a rigorous *manual* check 4,385 of those companies received a questionnaire
  - Response is (also) used to validate the model
- Approach has been applied for 4 years now
  - Model is **stable**



What have we learned from this?

There is a need for ML methodology

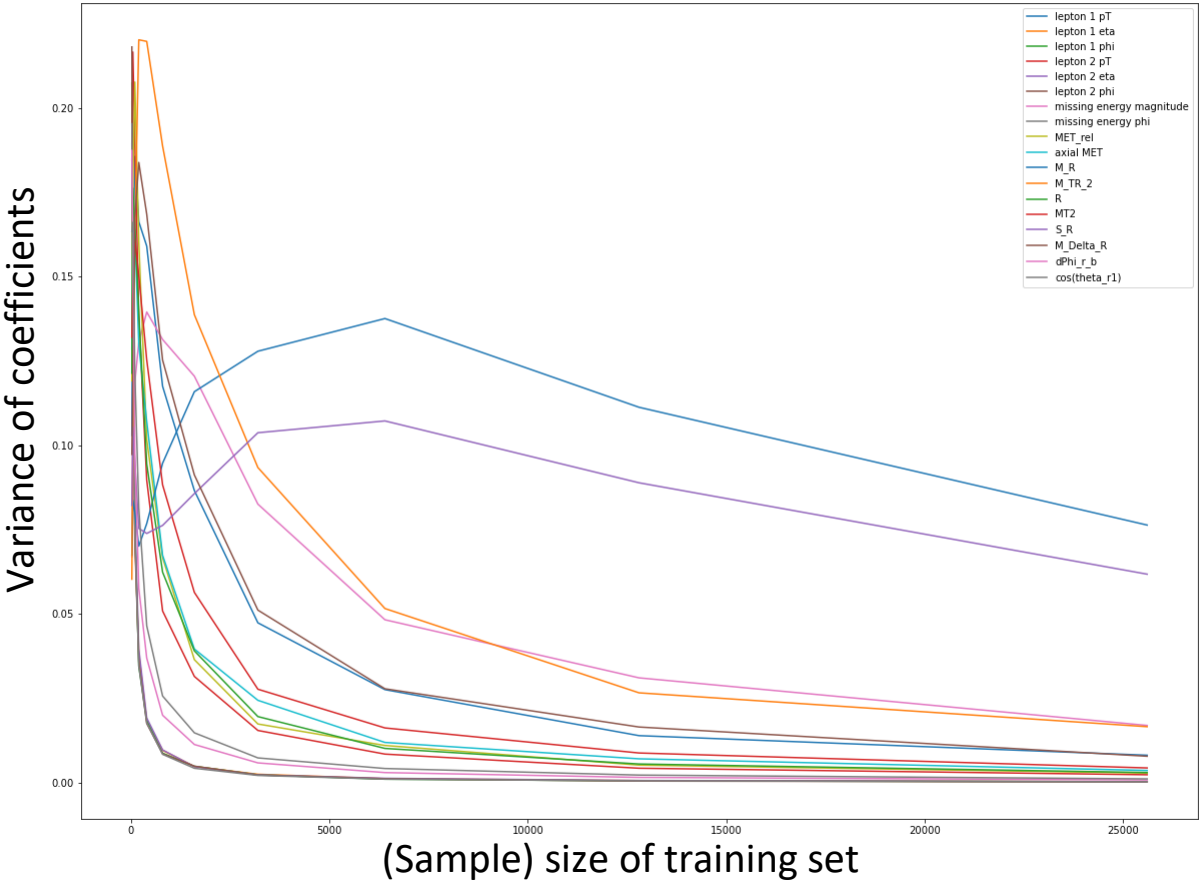
Methods are needed for:

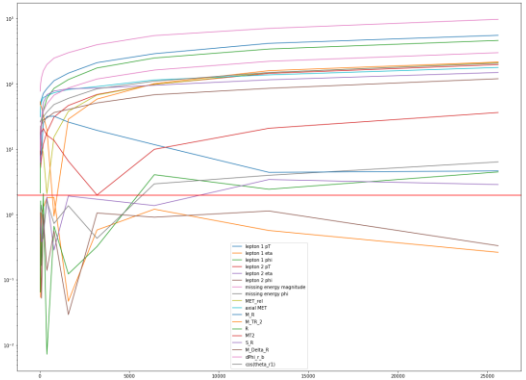
- Creating a representative training and test set
- Determining the optimal size of the training set
- Selecting the relevant features
- Developing ML-models that are externally valid
- Correcting the pseudo-probabilities of some ML-classifiers
- Dealing with intrinsic prevalence of ML-models (bias correction)
- Reducing bias (in general)
- ….

Methods are needed for:

- Creating a representative training and test set
- ***Determining the optimal size of the training set***
- Selecting the relevant features
- Developing ML-models that are externally valid
- Correcting the pseudo-probabilities of some ML-classifiers
- ***Dealing with intrinsic prevalence of ML-models (bias correction)***
- *Reducing bias (in general)*
- ….

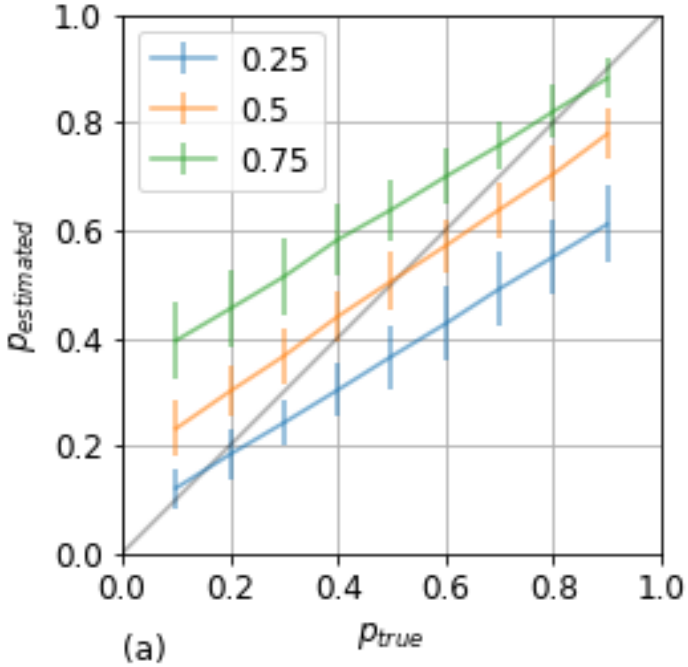
Size of the training set





Significance of features

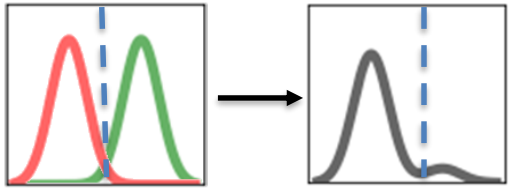
Dealing with intrinsic prevalence

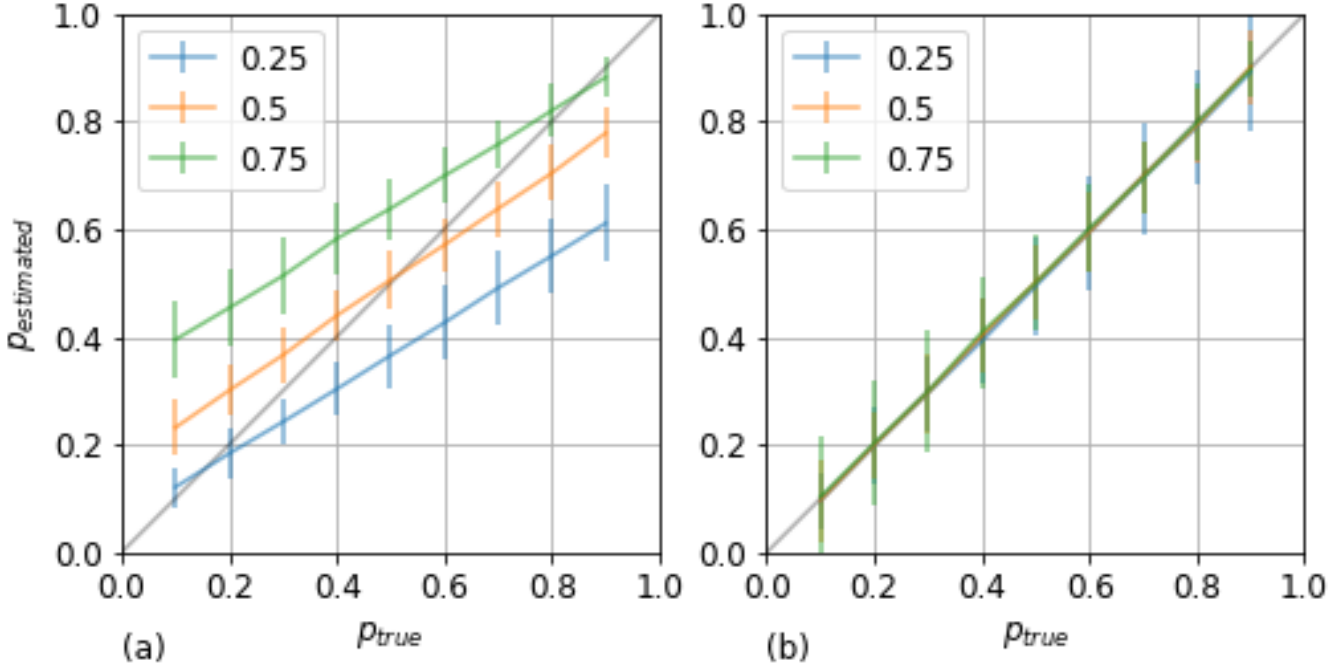


ML-based classifications models are biased towards the pos./neg. ratio they are trained on

A correction method has been developed!

Dealing with intrinsic prevalence





Calibration method uses probabilities https://github.com/mputs/BayesCCal

Online Platforms revisited

Online Platform detection

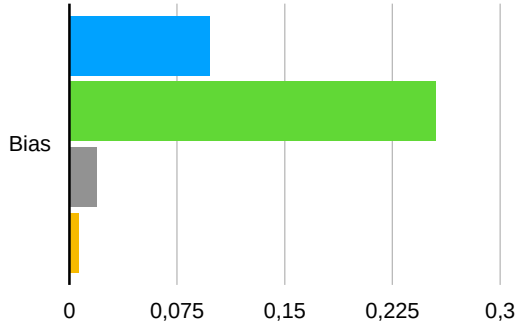
Results

		Proba's	Corrected proba's (BayesCCal)	Multiple models	TP	Pos. Est	Bias	Acc.	Bal. Acc.
1.	**SVM model**	No	No	No	69	2991	0,098	0,9012	0,09071
2.	**SVM model, proba's**	Yes	No	No	69	7657	0,255	0,9012	0,09071
3.	**SVM model, corrected proba's**	Yes	Yes	No	69	637	0,019	0,9848	0,7610
4.	**Multiple SVM models, corrected proba's**	Yes	Yes	Yes	69	306	0,007	0,9925	0,6275

Online Platform detection

Results

		Proba's	Corrected proba's (BayesCCal)	Multiple models	TP	Pos. Est	Bias	Acc.	Bal. Acc.
1.	**SVM model**	No	No	No	69	2991	0,098	0,9012	0,09071
2.	**SVM model, proba's**	Yes	No	No	69	7657	0,255	0,9012	0,09071
3.	**SVM model, corrected proba's**	Yes	Yes	No	69	637	0,019	0,9848	0,7610
4.	**Multiple SVM models, corrected proba's**	Yes	Yes	Yes	69	306	0,007	0,9925	0,6275



Conclusions

Conclusion

- ML provides us with a bunch of new techniques
- ML would greatly benefit from a more methodological approach
  - We can learn from Survey Methodology (TSE)
  - We need to think about the target population and not (only) about the data in the training and test set
  - Scores of (many) ML-classifiers are pseudo-probabilities
- This, and other issues, are very important when applying ML in the context of official statistics

Facts that matter