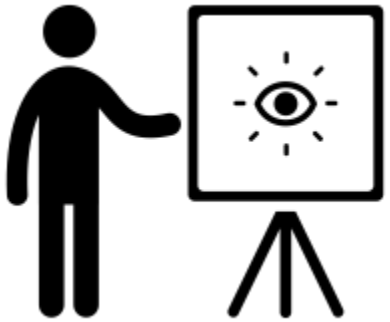


Explainability of statistical algorithms

UNECE HLG-MOS

Machine Learning for Official Statistics

Quality Framework for Statistical Algorithms

Joep Burger (Statistics Netherlands) and InKyung Choi (UNECE)

Webinar November 16, 2020

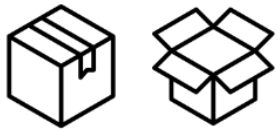
Quality Framework for Statistical Algorithms

Quality dimensions

1. Explainability ←
2. Accuracy
3. Reproducibility
4. Timeliness
5. Cost effectiveness

What

The degree to which a human can understand how a prediction is made by a statistical algorithm using its input features



Not

- mechanical working
- interpretability

Trade-off

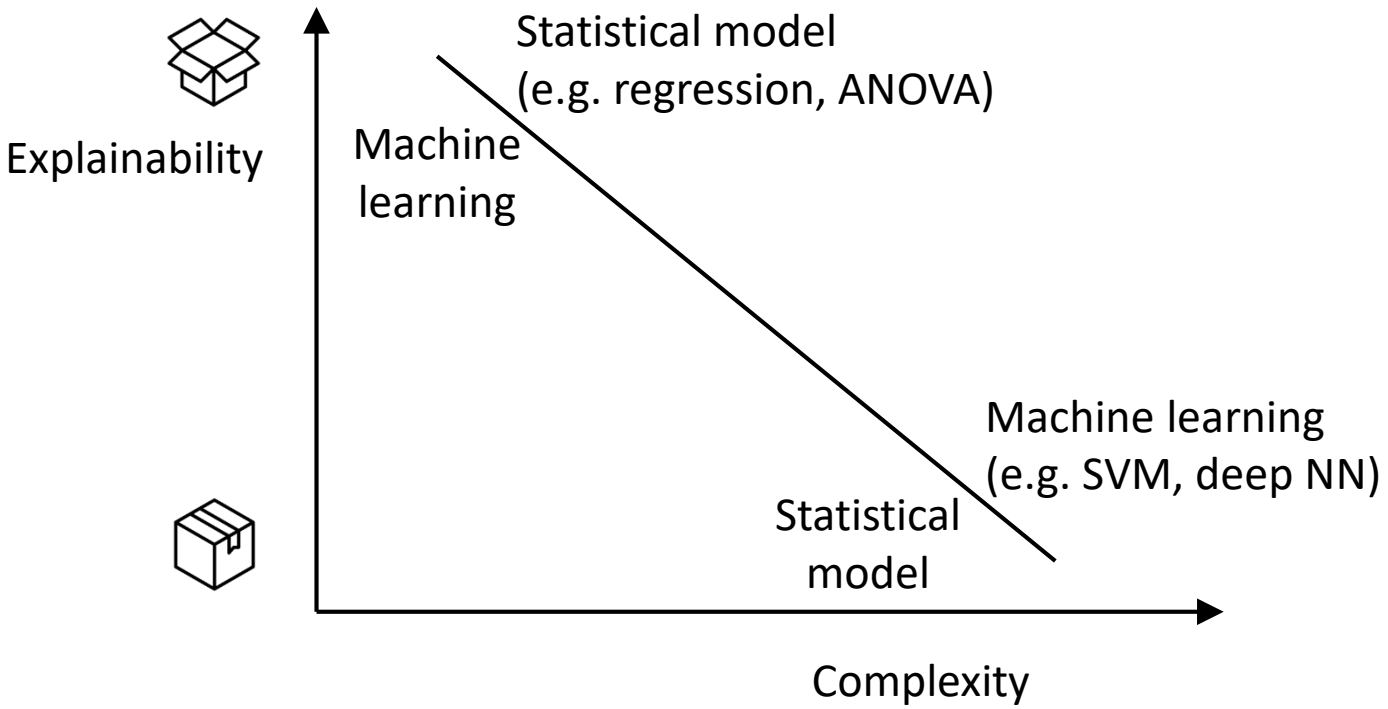
More data

More complex algorithms

Lower prediction error

Less explainable

Nuance



- Deep decision tree
- Deep NN with few features
- GLMM with transformed features and higher-order interactions

Why

Computer says

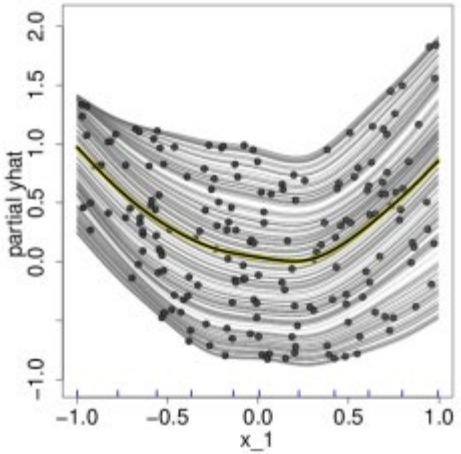
- no loan
- no insurance
- not a pedestrian
- no carcinoma
- it’s your face
- buy this
- read this
- check this record
- use this imputation
- …

- Trust
- New insights
- Safeguard
- Fair, Accountable, Transparent, Ethical (FATE) AI

Little Britain

How

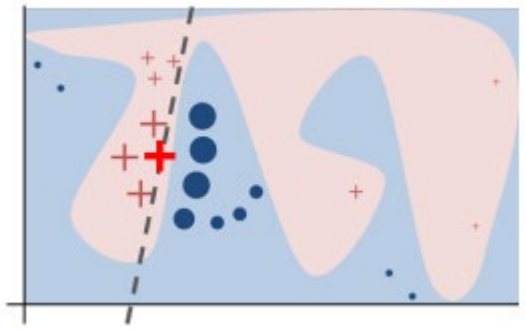
- Feature importance
- Individual conditional expectation
- Partial dependence plot



Goldstein et al. 2014

How (continued)

- LIME
- Shapley value
- Counterfactual
- Adversarial example
- Influential instance
- …



Ribeiro et al. 2016

Sum up

- Explainability as important as prediction error
- Need
  - less driven by use of ML
  - more by big data allowing for increased complexity
- Active field of research

References

- Arrieta, B.A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115, doi:10.1016/j.inffus.2019.12.012.
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F. and Eckersley, P. (2020). Explainable machine learning in deployment. arXiv:1909.06342.
- Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2014). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. arXiv:1309.6392.
- Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/.
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144, doi:10.1145/2939672.2939778.
- Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: a systematic review. arXiv:2006.00093.