le cnam

About Interpreting and Explaining Machine Learning and Statistical Models

Gilbert Saporta

CEDRIC-CNAM,

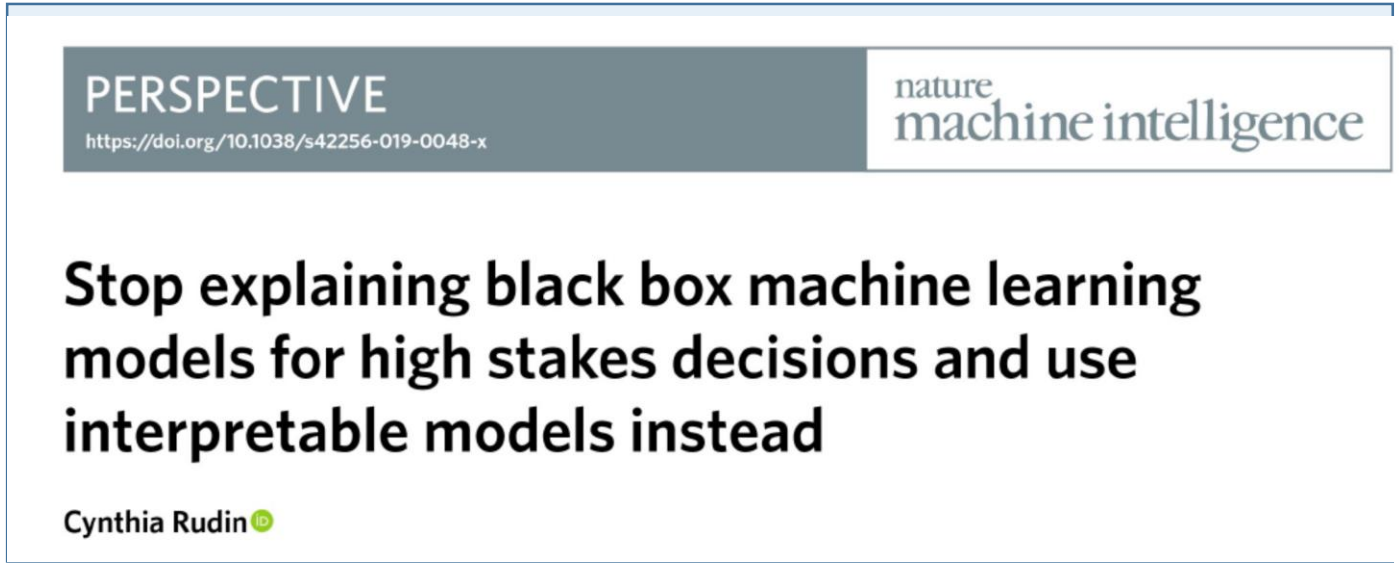
292 rue Saint Martin, 75003 Paris, France

Outline

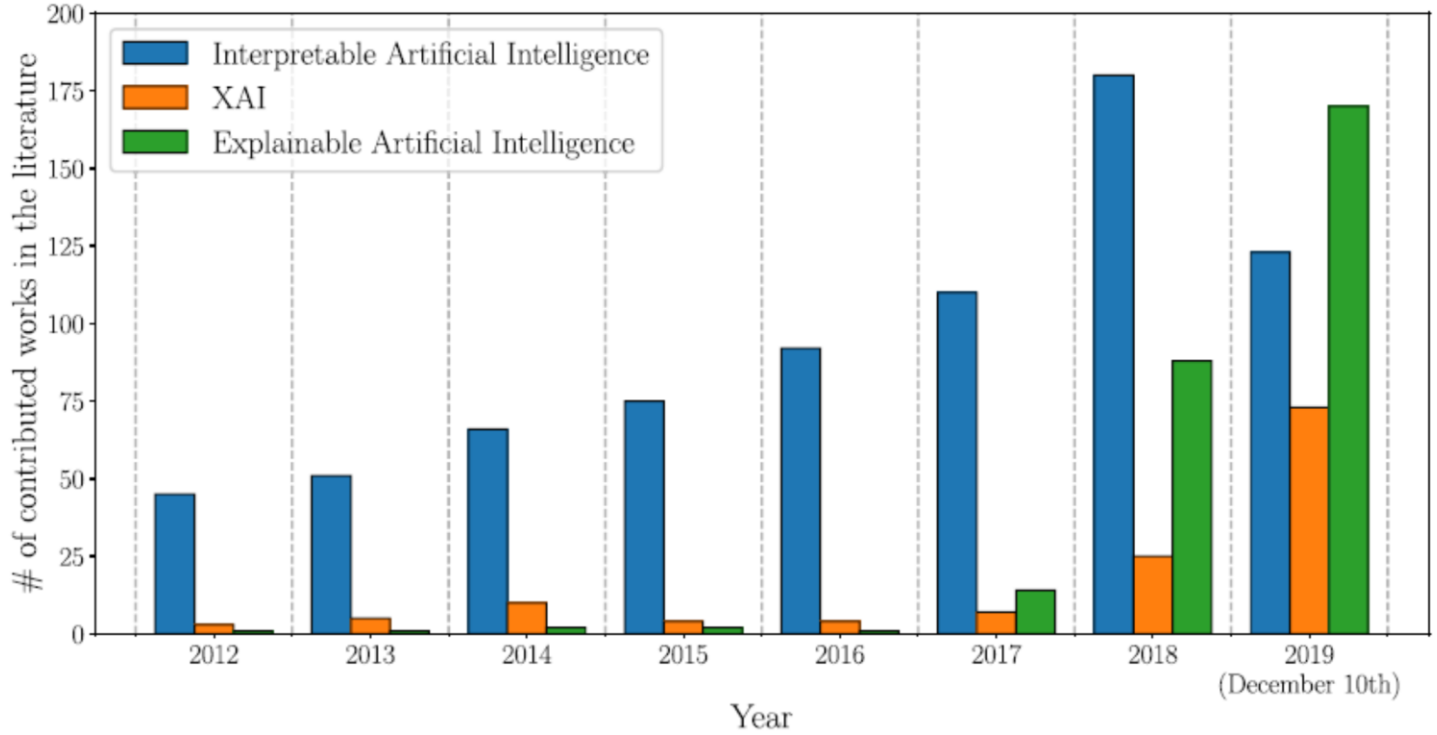
1. Explaining black box ML: the challenge
2. A long-standing debate in Statistics: explain or predict
3. Surrogate models
4. Variable importance
5. Causality and interpretability
6. Conclusions and perspectives

1. Explaining black box ML: the challenge

- "Unveiling secrets of black box models is no longer a novelty but a new business requirement" https://appsilon.com/please-explain-black-box/
- General Data Protection Regulation (EU GDPR)
- A proliferation of tools
  - LIME, SHAP, DALEX etc.
- And a controversial:



- A hot topic

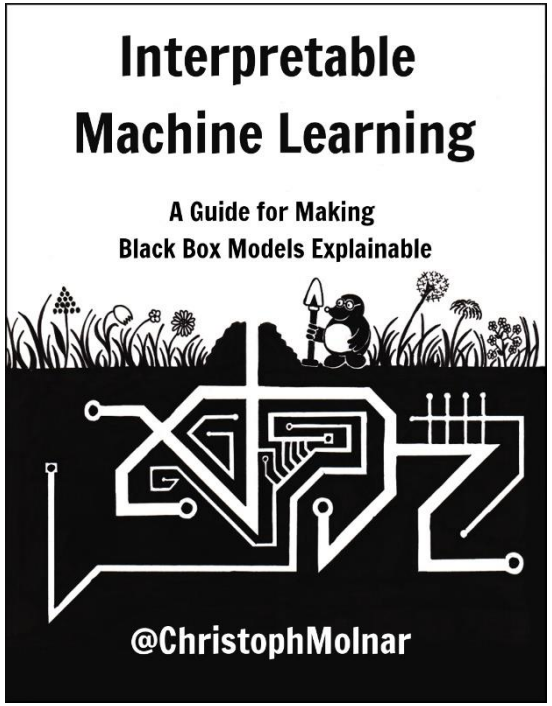


« eXplainable AI (XAI) proposes creating a suite of ML techniques that
1) produce more explainable models while maintaining a high level of learning performance (e.g., prediction accuracy), and
2) enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners. »

(Arrieta et al, 2020)

Explainable versus Interpretable

- Explainability
  - The ability to explain or to present in understandable terms to a human
  - Generally post-hoc (open the black box)
    - Local or Global interpretability
    - Specific or Agnostic
    - Variable importance measures
- Interpretable models: simplicity, sparseness
  - Logical models (trees, …)
  - Linear models (sparse, …)
  - Case based



BETA

- **B**lack Box **E**xplanation through **T**ransparent **A**pproximation
  - Simultaneously optimizing for fidelity to the original model and interpretability of the explanation. Lakkaraju et al (2017)

DALEX

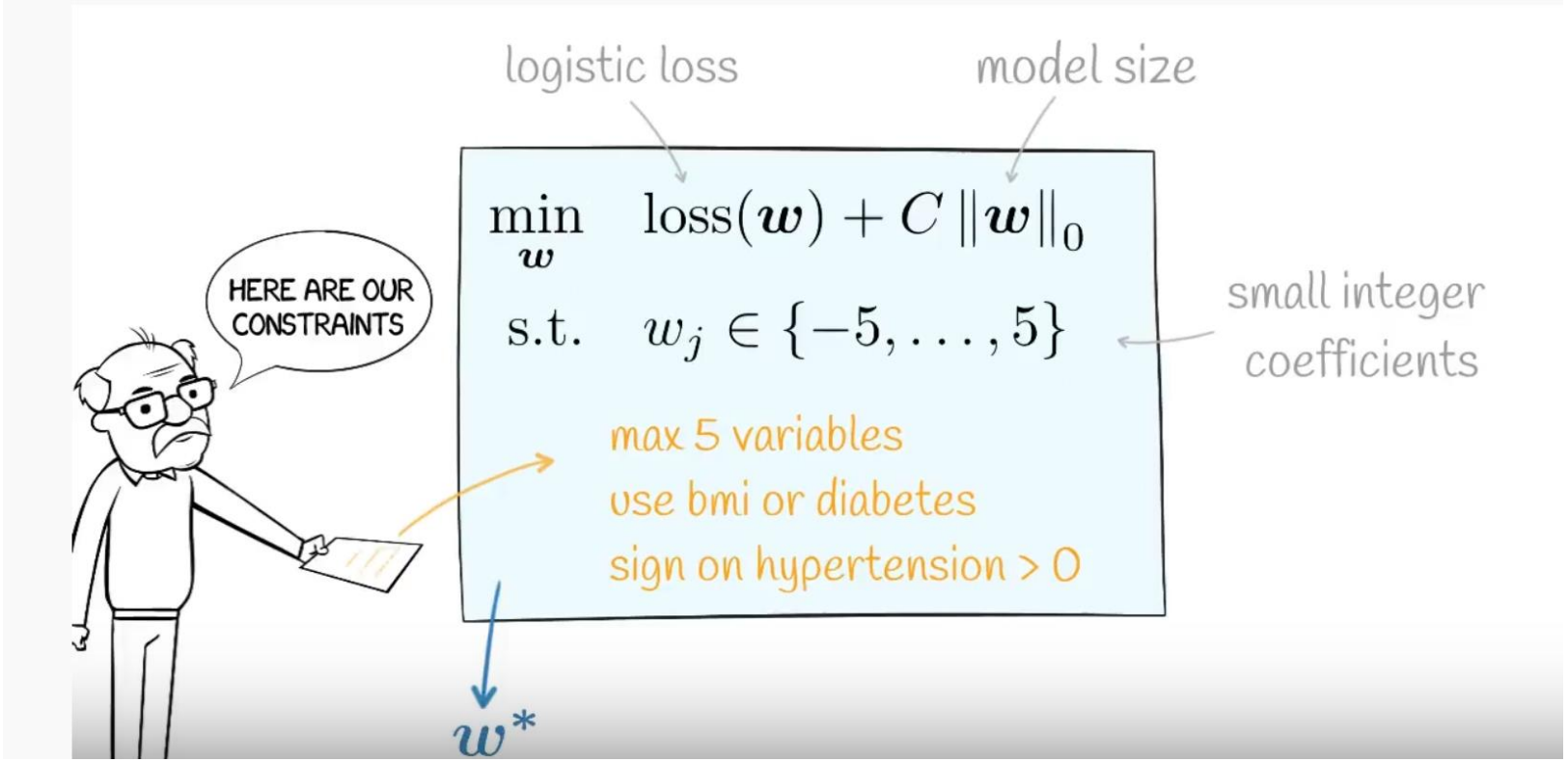
- **D**escriptive m**A**chine **L**earning **Ex**planations. Biecek (2018)

GLASS-BOX

- Explaining AI decisions with counterfactual statements through conversation with a voice enabled virtual assistant. Sokol & Flach (2018)

Etc.

Risk-calibrated Supersparse Linear Integer Model (RiskSLIM)



Ustun, B., & Rudin, C. (2017)

CORELS (Certifiable Optimal RulE ListS)

- An example rule list that predicts two-year recidivism for the ProPublica data set, found by CORELS.
  - if (age = 18 - 20) and (sex = male) then predict *yes*
  - else if (age = 21- 23) and (priors = 2 - 3) then predict *yes*
  - else if (priors > 3) then predict *yes*
  - else predict *no*
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017)

- Similar to class association rules CARs introduced by Liu et al., (1998)

2. A long-standing debate in statistics: explain or predict

- Breiman (2001), GS (2008), Shmueli (2010), Donoho (2017)

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

To explain (or understand) or to predict?

Donoho (2017):

- The generative modelling culture
  - seeks to develop stochastic models which fits the data, and then make inferences about the data-generating mechanism based on the structure of those models. Implicit (…) is the notion that there is a **true model** generating the data, and often a truly `best' way to analyze the data.
- The predictive modelling culture
  - is silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only **accuracy of prediction** made by different algorithm on various datasets. **Machine Learning** is identified by Breiman as the epicenter of the Predictive Modeling culture.

- Standard conception (models for understanding)
  - Provide some comprehension of data and their generative mechanism through a parsimonious representation.
  - A model should be simple and its parameters interpretable for the specialist : elasticity, odds-ratio, etc.
- In « Big Data Analytics or ML » one focus on prediction
  - For new observations: generalization
  - Models are merely algorithms

  Cf GS, compstat 2008
- “Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data” (Breiman, 2001).
- “Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms” (Vapnik, 2006).

ML and economics

Symposium: Big Data

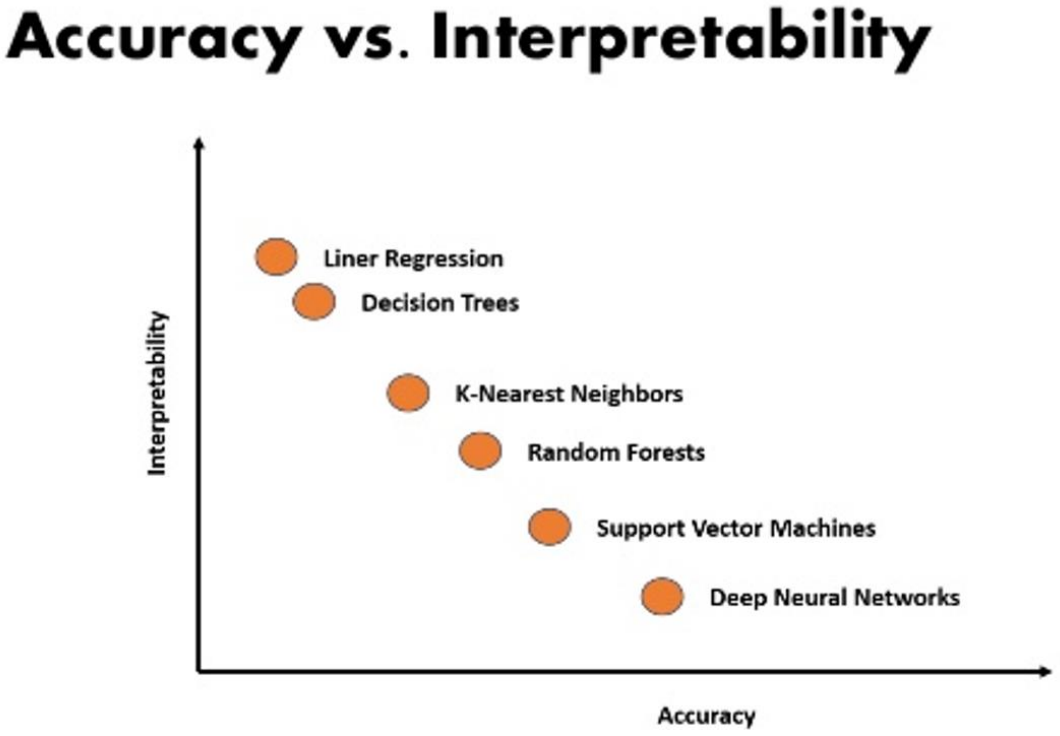
Big Data: New Tricks for Econometrics (pp. 3-28)
Hal R. Varian

AMERICAN ECONOMIC ASSOCIATION
AEA American Economic Association
AEA | Journals | Annual Meeting | EconLi

Journal of Economic Perspectives: Vol. 28 No. 2 (Spring 2014)

- "Data manipulation tools and techniques developed for small datasets will become increasingly inadequate to deal with new problems.
- Researchers in machine learning have developed ways to deal with large datasets and economists interested in dealing with such data would be well advised to invest in learning these techniques."

A cliché



- More complex models are supposed to have better accuracy
- This is often not true, particularly when the data are structured (Rudin, 2019)
- Statistical Learning Theory proves the existence of an optimal complexity (Vapnik, 2006)

https://medium.com/swlh/the-great-ai-debate-interpretability-1d139167b55

3. Surrogate models

“A surrogate model is an interpretable model that is trained to approximate the predictions of a black box model” (Molnar, 2020)

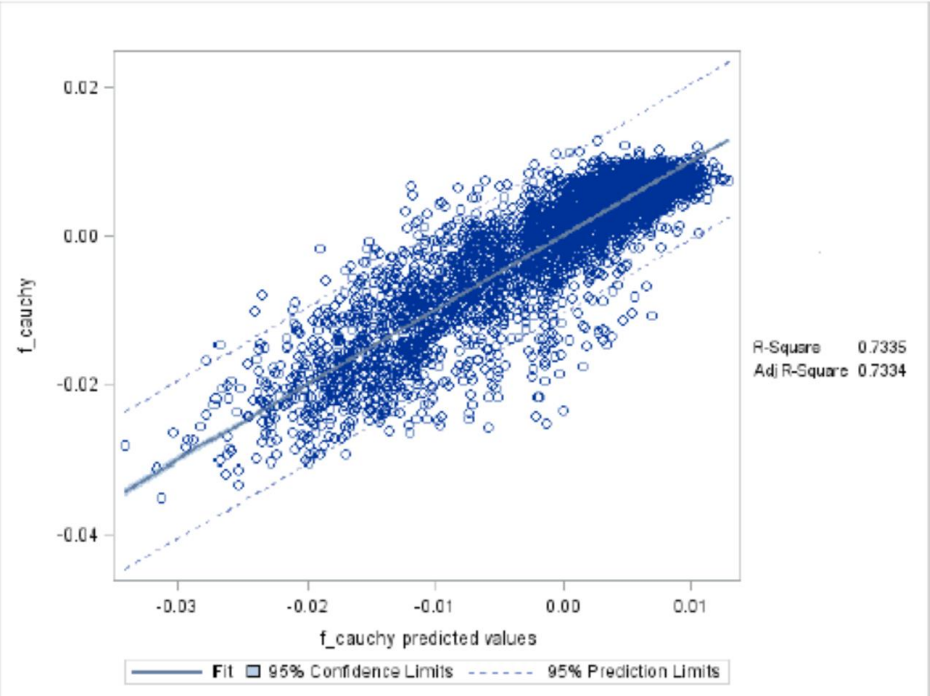
3.1 Global surrogate models

- Trees, linear models are among the favorite surrogate models
- Easy to use
- Note: the surrogate model tries to fit the black box model not the data

- Example: Linearizing a kernel classifier (Liberati, Camillo, Saporta, 2017)
  - A credit scoring example: 75 000 « good » and 10 000 « bad » small businesses asking an italian bank for a credit
  - The best classifier was a SVM with a Cauchy kernel

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + \mathbf{b} \quad k(\mathbf{x}_i, \mathbf{x}) = \frac{1}{1 + \|\mathbf{x}_i - \mathbf{x}\|^2}$$

  - Difficult to use. Professionals prefer an additive scoring rule
- Solution: Reconstruction of the kernel discriminant function through a linear regression where $f(\mathbf{x})$ is the target and the original variables are the predictors



Discriminant rules	Correct classification rates		
	Bad class	Good class	Overall
Cauchy	71.48	74.78	73.86
Logistic regression	54.91	89.88	79.90
FLDA	61.88	56.57	58.08
Reconstructed	73.80	74.51	74.30

r=0.85

A paradox

- Since the vector space linearly spanned by the input variables is embedded in the feature space, there should be no gain to approximate **y** by the kernel classifier and then approximate $f(\mathbf{x})$ by a linear combination of the input variables, instead of a direct projection onto the x's .
- This paradox disappears if we notice that the SVM classifier does not correspond to the orthogonal projection onto the feature space, or, in other words, to the least squares approximation of the binary response. It consists in maximizing the margin around the boundary and not in minimizing the sum of squares of residuals.

3.2 Local surrogate models

- Local surrogate models are interpretable models that are used to explain individual predictions of black box machine learning models.
- A popular approach: **LIME** (**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations) introduced in the paper: *"Why Should i Trust You?" Explaining the Predictions of Any Classifier* (Ribeiro et al., 2016)
- Model-agnostic means that it can be applied to any machine learning model. The technique attempts to understand the model by perturbing the input of data samples and understanding how the predictions change.

- Every complex model is linear on a local scale
- Two very similar observations are expected to behave predictably even in a complex model. LIME fits a simple model around a single observation that will mimic how the global model behaves at that locality. The simple model can then be used to explain the predictions of the more complex model locally.

1. For each prediction to explain, permute the observations n times.
2. Let the complex model predict the outcome of all permuted observations.
3. Calculate the distance from all permutations to the original observation.
4. Convert the distance to a similarity score.
5. Select m features best describing the complex model outcome from the permuted data.
6. Fit a simple model to the permuted data, explaining the complex model outcome with the m features from the permuted data weighted by its similarity to the original observation.
7. Extract the feature weights from the simple model and use these as explanations for the complex models local behavior.

https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html

4. Variable importance

Also known as feature importance in ML. Seems essential to explain a prediction-decision.

4.1 Specific methods

- A common belief is that simple models, like linear or logistic regression are easily interpretable
- Generally untrue!
- Except in case of orthogonal designs, parameter values hardly reflect variable importance

- More than 14 methods of quantifying variable importance in linear models! (Grömping, 2015, Wallard, 2015)
  - Based on coefficients, correlation, explained variance etc.

b's (not normalized)
Joint contribution (not normalized)
Squared semipartial correlations
Squared raw correlations
Squared standardized b's
Sequential SS, from left to right
Sequential SS, from right to left
Pratt
CAR scores/Gibson
Green et al.
Fabbris
Genizi/Johnson
LMG
PMVD

`R package relaimpo`

- A related problem: variable selection
  - Discard unnecessary variables
  - Fight the curse of dimensionality (p>>n)
  - Provide sparse solution (Lasso)
- But:
  - Leaves unsolved the problem of highly correlated variables; why choose x_1 rather than x_2 if $|r(x_1, x_2)|$ is high?
  - " *Statistical significance* plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, *even if they are statistically significant*, results in improved prediction accuracy"
  - "A researcher might choose to retain a causal covariate which has a strong theoretical justification *even if is statistically insignificant*. "

(Shmueli, 2010)

4.2 Agnostic methods

May be applied to any model

4.2.1 Permutation Variable Importance

- Introduced by Breiman (2001b) for random forests as the increase in the model's prediction error when permuting randomly (shuffling) the values of the predictor.
- Easy to understand approach, takes into account interactions
- Importances are not additive
- May lead to physically impossible pairs of units, and outliers
- Should be repeated and averaged

4.2.2 Shapley value

Inspired by game theory (Lundberg & Lee, 2017)

- Local model (for one unit)
  - Prediction task= game
  - Predictor= player
  - Subset of predictors= coalition
  - Prediction= payout
  - Gain= prediction-average prediction for all units
- The Shapley value is the weighted average of all possible differences when a predictor is added or not, across all possible coalitions

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p-|S|-1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right)$$

- Like LIME, allows an explanation of a decision: eg which were the most important features in classifying an unit in some class.

- Nice mathematical properties
  - Including additivity and uniqueness under some conditions
- Shapley global importance of a predictor
  - Obtained by averaging the Shapley (local) values over all n units
- Exponential number of coalitions : 2^p
- R and Python packages

4.3 Importance-Performance Analysis

- Data visualisation is necessary
- IPA is a simple graphical tool used in marketing studies for customer satisfaction studies
  - Performance: average level of a driver
  - Variable Importance in a regression model

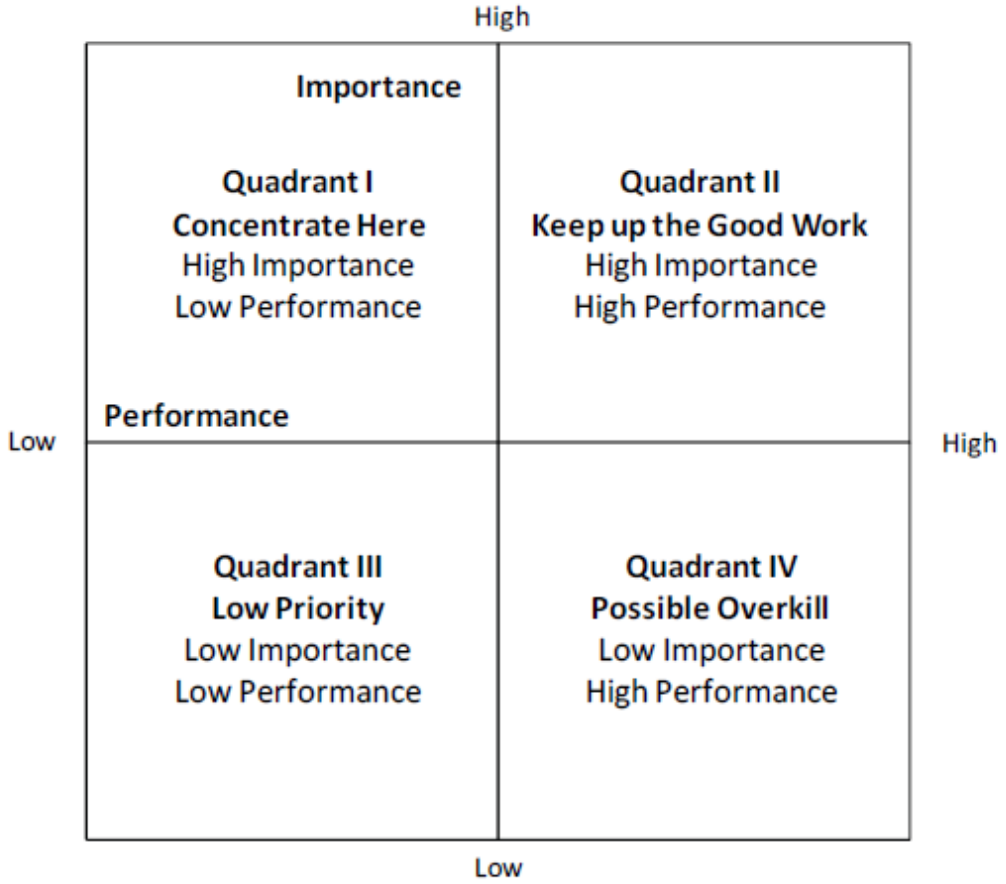
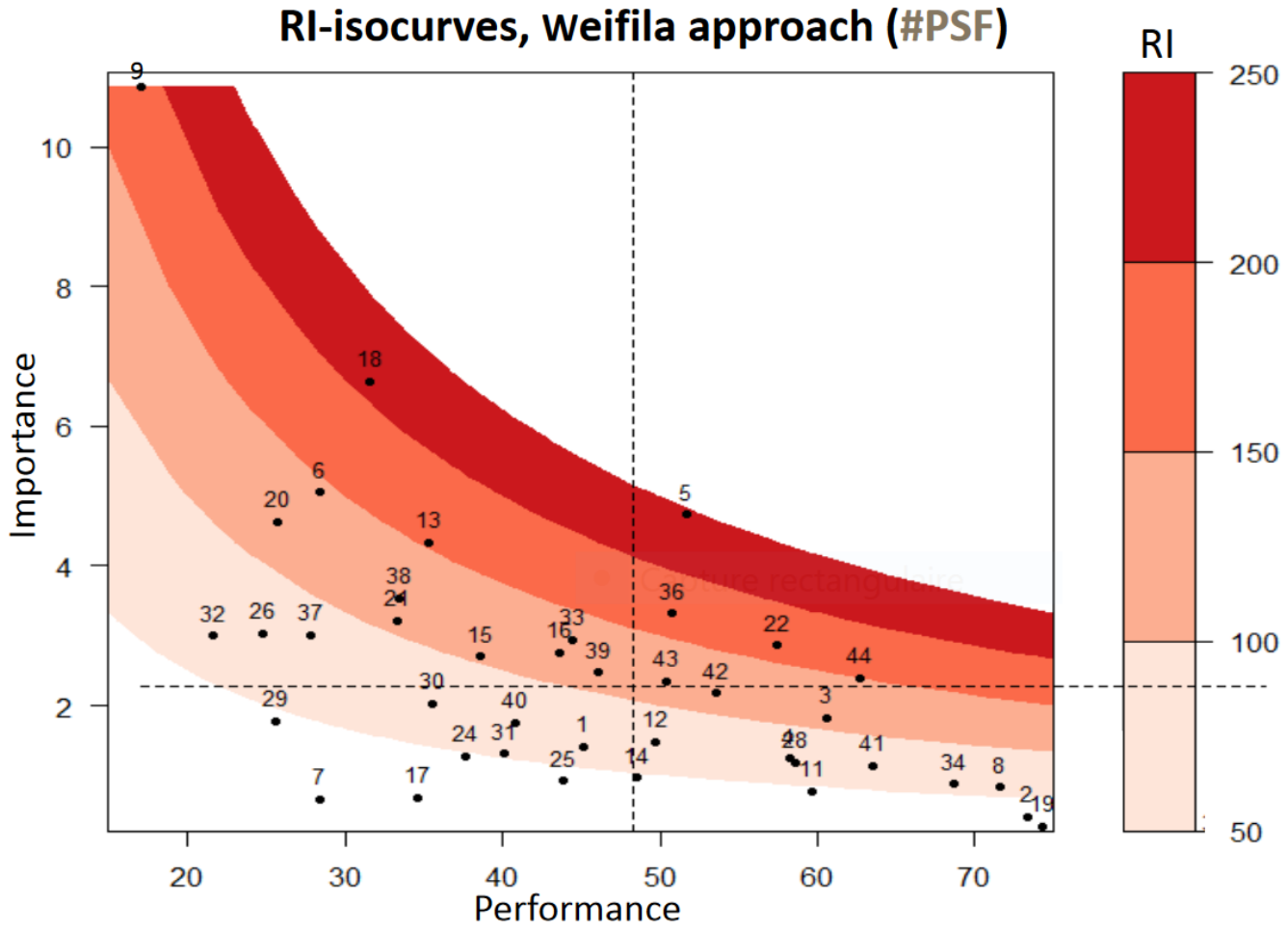


Figure 1. Importance-Performance analysis grid

An application to work-related psychosocial factors (PSFs) impacting mental health (Daouda, Temime, Saporta, Hocine 2019)

- Sample of 3200 individuals, representative of the French workers
- Mental health status measured by GHQ 28 «General Heath Questionnaire with 28 items». Variables : 44PSFs
- Performance measurement : prevalence of exposure to each PSF.
- Importance measurement : strength of association between mental health and PSF



Rank	#PSF Weifila	PSF description
1	**5**	Unsatisfactory communication at work
2	**18**	Inability to depend on work collaborators
3	**9**	Imbalance private and professional life
4	**36**	Emotional demands at work
5	**22**	No good career prospects
6	**13**	Not feeling valued or recognized at work

Ranking Index RI = importance x performance

4.4 The multiplicity of good models and the Rashomon effect

- *"A wonderful Japanese movie in which four people, from different vantage points, witness an incident in which one person dies and another is supposedly raped. When they come to testify in court, theyall report the same facts, but their stories of what happened are very different."*
- "*What I call the Rashomon Effect is that there is often a multitude of different descriptions [equations f(x)] in a class of functions giving about the same minimum error rate*". (Breiman, 2001a)
- Variable importance should not be measured in one single model, but taking into account the set of almost-equally-accurate predictive models. Hence the concept of Variable Importance Cloud (Dong & Rudin, 2019).

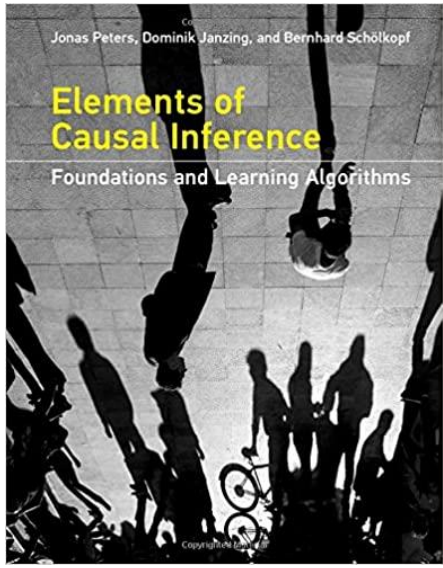
5. Interpretability and causality

- Measuring variable importance cannot answer this question: what would be the response if one or more predictors were changed intentionally or unintentionally?
- Changing x_j may change the values of other predictors if they are connected in a causal way
- In marketing, practitioners need to know what would be the effect of an intervention. Predictors should be actionable: “drivers”.

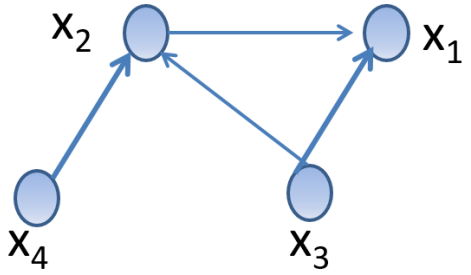
- Regression, ML models are not causal, but are often used as if they were, hence a lot of disappointment.
  - Measuring the effect of a variable “all things being equal” is often absurd.
- Seeing is not doing (Pearl & Mackenzie, 2018)

$$P(Y \mid X = x) \neq P(Y \mid do(X = x))$$

- Looking for causal models
  - Randomized control trials , AB testing in web advertising
  - Propensity score matching
  - Learning Bayesian networks
- Many works in progress (cf Peters et al, 2017; Ke et al, 2019)



- Towards hybrid models?
  - A causality graph for predictors



  - Followed by a (hopefully interpretable) model for the response

$$\hat{y} = f(\mathbf{x})$$

Conclusions

- Take home:
  - The demand for interpretable models is leading to an abundance of new methods.
  - Simple models are not really simple
  - Agnostic approaches are useful for measuring variable importance. Often better to consider a set of models
- Feature Importance does not imply causal effect
  - But actionability needs not only causality but also ease of modification.

References

- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, *18*(1), 8753-8830.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.
- Biecek, P. (2018). DALEX: explainers for complex predictive models in R. *The Journal of Machine Learning Research*, *19*(1), 3245-3249.
- Breiman, L. (2001a). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199-231.
- Breiman, L. (2001b). Random forests. *Machine learning*, *45*(1), 5-32.
- Daouda, O. S., Temime, L., Saporta, G., & Hocine, M. (2019). How to prioritize work-related psychosocial factors impacting mental health? Regression and random forest approaches. In *ISCB 40*.
- Dong, J., & Rudin, C. (2019). Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models. *arXiv preprint arXiv:1901.03209*.

- Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.
- Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(2), 137-152.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42.
- Hall, P., & Gill, N. (2019). *Introduction to Machine Learning Interpretability*. 2nd edition, O'Reilly Media, Incorporated.
- Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Pal, C., & Bengio, Y. (2019). Learning Neural Causal Models from Unknown Interventions. *arXiv preprint arXiv:1910.01075*.
- Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.
- Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094-1111.

- Liberati, C., Camillo, F., & Saporta, G. (2017). Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. *Advances in Data Analysis and Classification*, *11*(1), 121-138
- Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In *KDD* (Vol. 98, pp. 80-86).
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765-4774).
- Molnar, C. (2020). *Interpretable machine learning , A Guide for Making Black Box Models Explainable*, https://christophm.github.io/interpretable-ml-book.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.
- Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.
- Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.
- Saporta, G. (2008). Models for understanding versus models for prediction. In *COMPSTAT 2008* (pp. 315-322). Physica-Verlag
- Shmueli, G. (2010). To explain or to predict?. *Statistical science*, *25*(3), 289-310.
- Sokol, K., & Flach, P. A. (2018, July). Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. In *IJCAI* (pp. 5868-5870).
- Ustun, B., & Rudin, C. (2017). Optimized risk scores. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1125-1134).
- Vapnik, V. (2006) *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer
- Wallard, H. (2015). Using explained variance allocation to analyse importance of predictors. In *16th ASMDA conference proceedings* (Vol. 30), 1043-1054,