Bad Practices in ML

Jan Brabec

Critiquing and Correcting Trends in Machine Learning - NeurIPS 2018 Workshop

- https://ml-critique-correct.github.io/

Current trends in ML scholarship:

1. Failure to distinguish between explanation and speculation.
2. Failure to identify the sources of empirical gains, e.g. emphasizing unnecessary modifications to neural architectures when gains actually stem from hyper-parameter tuning.
3. Mathiness: the use of mathematics that obfuscates or impresses rather than clarifies, e.g. by confusing technical and non-technical concepts.

4) Misuse of language

- **Suggestive Definitions:** thought vectors, consciousness prior, …
- **Overloading Technical Terminology:** deconvolution, generative models, …
- **Suitcase Words:** interpretability, generalization

What sort of papers best serve their readers?

1. provide intuition to aid the reader’s understanding, but clearly distinguish it from stronger conclusions supported by evidence
2. describe empirical investigations that consider and rule out alternative hypotheses
3. make clear the relationship between theoretical analysis and intuitive or empirical claims
4. use language to empower the reader, choosing terminology to avoid misleading or unproven connotations

Suggestions to authors

1. Ask: What worked? Why? Instead of just: How well?
2. Error analysis
3. Robustness checks (sensitivity to hyperparameters, randomness, …)
4. Ablation studies

Please Stop Explaining Black Box Models for High Stakes Decisions – Cynthia Rudin

- „The practice of trying to *explain* black box models, rather than creating models that are *interpretable* in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society“

https://arxiv.org/abs/1811.10154

Please Stop Explaining Black Box Models for High Stakes Decisions – Cynthia Rudin

1. „It is a myth that there is necessarily a tradeoff between accuracy and interpretability.“
2. „The preconceived belief that there is a tradeoff between accuracy and interpretability has led many researchers to forgo the attempt to produce an interpretable model. This problém is compounded by the fact that researchers are now trained in deep learning, but not in interpretable machine learning.“

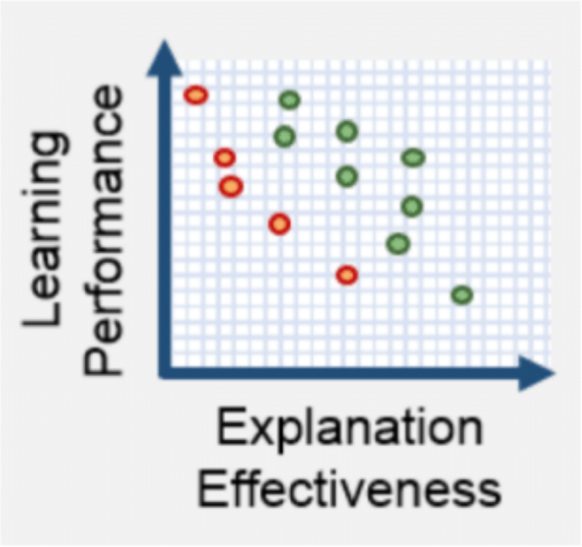


Figure 1: An imaginary illustration of the accuracy-interpretability tradeoff, taken from the DARPA XAI BAA.

https://arxiv.org/abs/1811.10154

Please Stop Explaining Black Box Models for High Stakes Decisions – Cynthia Rudin

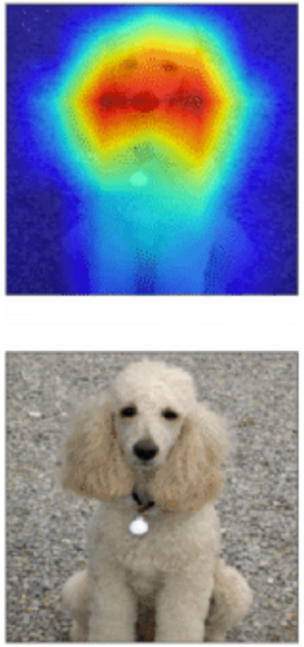
- **Explainable ML methods provide explanations that are not faithful to what the original model computes.**

- „An explainable model that has a 90% agreement with the original model indeed explains the original model most of the time. However [..] if a tenth of the explanations are incorrect, one cannot trust the explanations and thus one cannot trust the original black box.“

https://arxiv.org/abs/1811.10154

Please Stop Explaining Black Box Models for High Stakes Decisions – Cynthia Rudin

- **Explanations often do not make sense, or are incomplete.**



Cat

Figure 2: Saliency does not explain anything except where the network is looking. We have no idea why this image is labeled as a cat. Credit: Alexis Cook.

https://arxiv.org/abs/1811.10154

Please Stop Explaining Black Box Models for High Stakes Decisions – Cynthia Rudin

- **Interpretable models can entail significant effort to construct, in terms of both computation and domain expertise.**

https://arxiv.org/abs/1811.10154

Please Stop Explaining Black Box Models for High Stakes Decisions – Cynthia Rudin

- Currently the GDPR and other AI regulation plans govern “right to an explanation," where only an explanation is required, not an interpretable model

- „We propose to govern that, for high-stakes decisions, *no black box should be used if there exists an interpretable model with the same level of performance“*

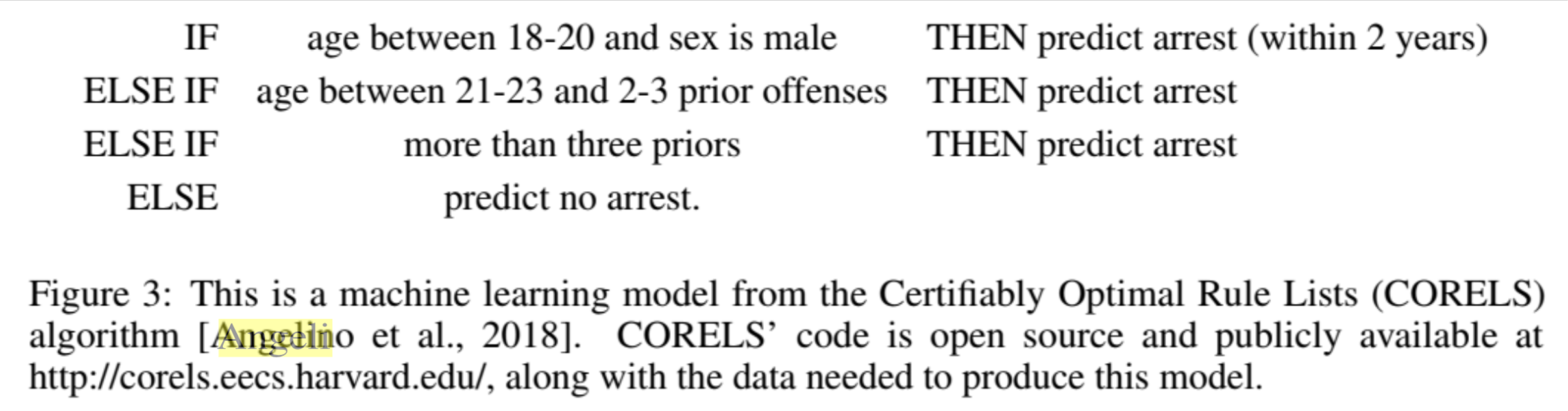
https://arxiv.org/abs/1811.10154

Please Stop Explaining Black Box Models for High Stakes Decisions – Cynthia Rudin

- There is a proprietary model called COMPAS in widespread use in the U.S. Justice System for predicting whether someone will be arrested after their release [Brennan et al., 2009]. COMPAS uses over 130 covariates.

https://arxiv.org/abs/1811.10154

Please Stop Explaining Black Box Models for High Stakes Decisions – Cynthia Rudin



```
     IF    age between 18-20 and sex is male      THEN predict arrest (within 2 years)
ELSE IF    age between 21-23 and 2-3 prior offenses  THEN predict arrest
ELSE IF    more than three priors                 THEN predict arrest
   ELSE    predict no arrest.
```

Figure 3: This is a machine learning model from the Certifiably Optimal Rule Lists (CORELS) algorithm [Angelino et al., 2018]. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data needed to produce this model.

https://arxiv.org/abs/1811.10154

Expanding search in the space of empirical ML – Bronwyn Woods

- **Target.** The specification of the problem to be solved, including the scope of the relevant population, assumptions about environment, form of the expected output, and measure of success.
- **Data.** An actual dataset instantiated from a particular sampling scheme, measurement methodology, and (frequently) labeling effort.
- **Algorithm.** The model formalism, hyperparameter selection, and training methodology.

https://arxiv.org/pdf/1812.01495.pdf

Expanding search in the space of empirical ML – Bronwyn Woods

- **Redefine novelty**
- **Make space for synthesis –** new viewpoints, challenge long-held ideas
- **Incentivize publishing open world experimentation**

https://arxiv.org/pdf/1812.01495.pdf

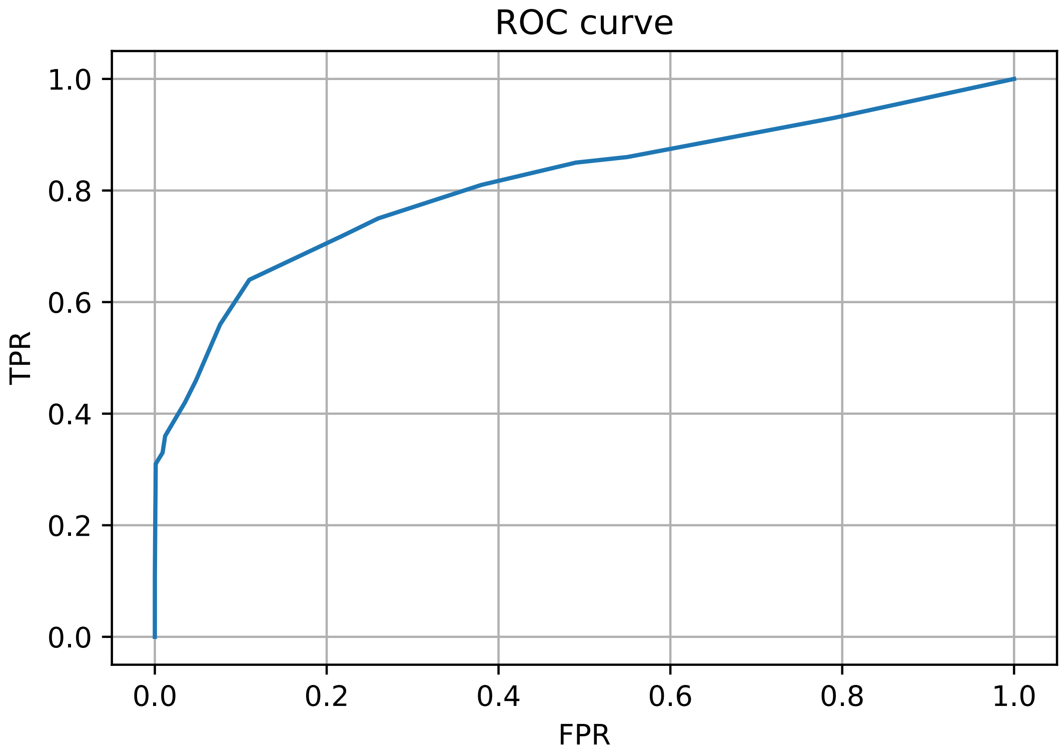
Bad practices in evaluation methodology relevant to class-imbalanced problems

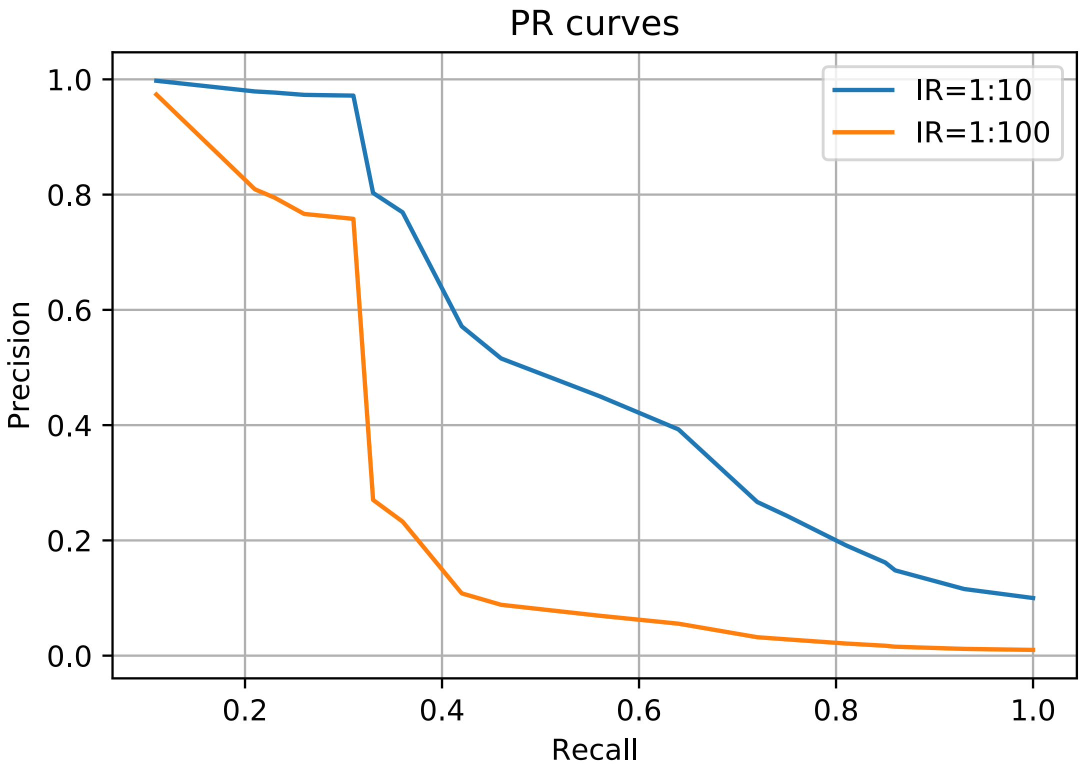
https://arxiv.org/pdf/1812.01388.pdf

1) Class imbalance ratios encountered in the wild should always be discussed and addressed in applied papers.

https://arxiv.org/pdf/1812.01388.pdf

2) ROC curves alone do not contain information about imbalance ratio of the test dataset.





https://arxiv.org/pdf/1812.01388.pdf

3) Ideally, the test dataset should originate from the same pipeline as the data in the production environment to reflect the distribution in the wild.

https://arxiv.org/pdf/1812.01388.pdf

4) Precision and F-score on datasets with artificial class distribution are almost always too optimistic and measuring them usually does not make sense.

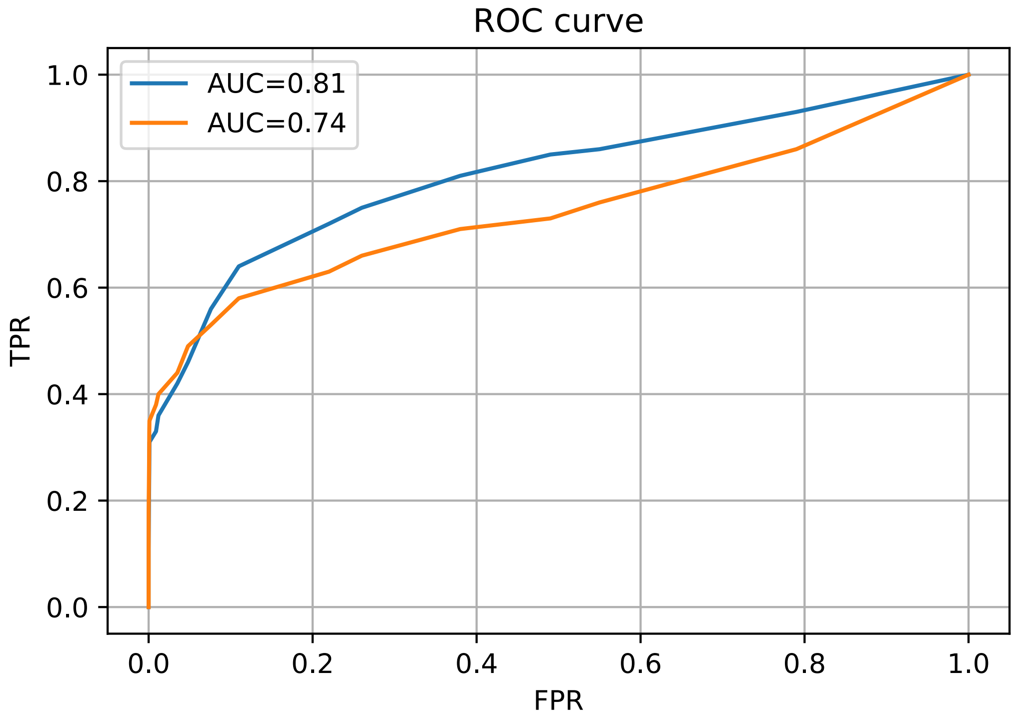
https://arxiv.org/pdf/1812.01388.pdf

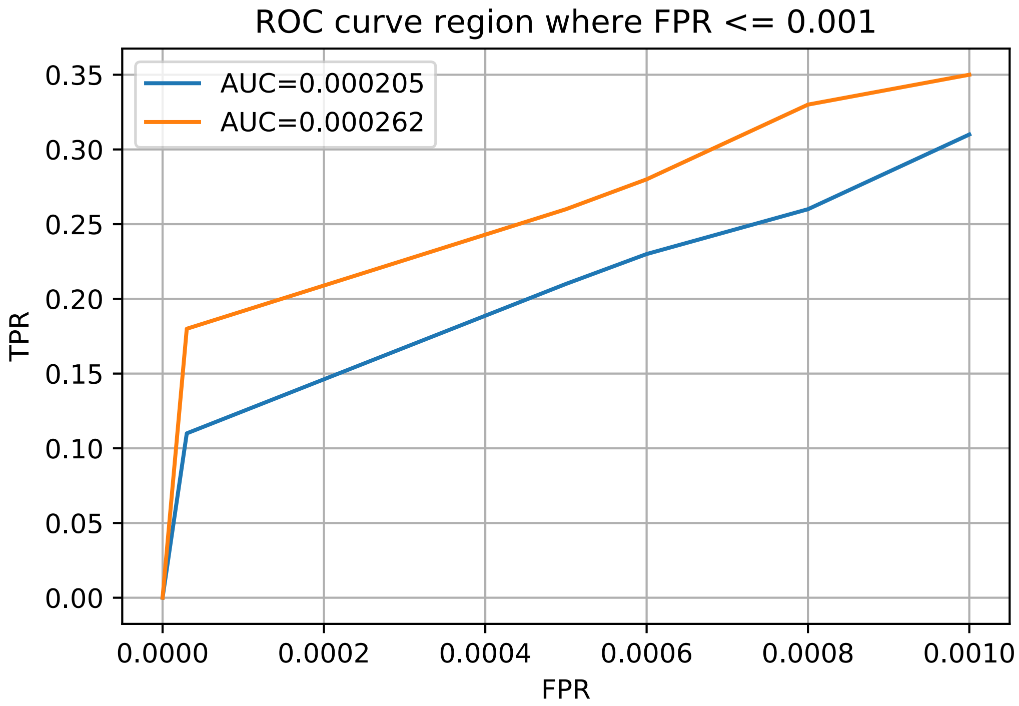
5) Precision can be adjusted to different imbalance ratios with the following formula if it is not possible to obtain a representative test dataset:

$$adjusted\_precision = \frac{\frac{p_{real}}{p_{test}} \cdot \mathrm{TP}}{\frac{p_{real}}{p_{test}} \cdot \mathrm{TP} + \frac{1-p_{real}}{1-p_{test}} \cdot \mathrm{FP}}$$

https://arxiv.org/pdf/1812.01388.pdf

6) The regions of no interest may represent most of the area under the curve, having dominant influence on the value of AUC. AUC should only be compared on the regions of interest.





https://arxiv.org/pdf/1812.01388.pdf