Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Sreya Dutta Roy ; Ziqian Lin

Introduction :

Black box ML Models are being Deployed in High Stakes domains these days. Some examples of such high stakes domains would be Criminal Justice, Healthcare, Energy Reliability and Financial Risk Management. These black box models often lead to catastrophic issues due to lack of interpretability/comprehensibility/transparency.

Types of Black box Models include :

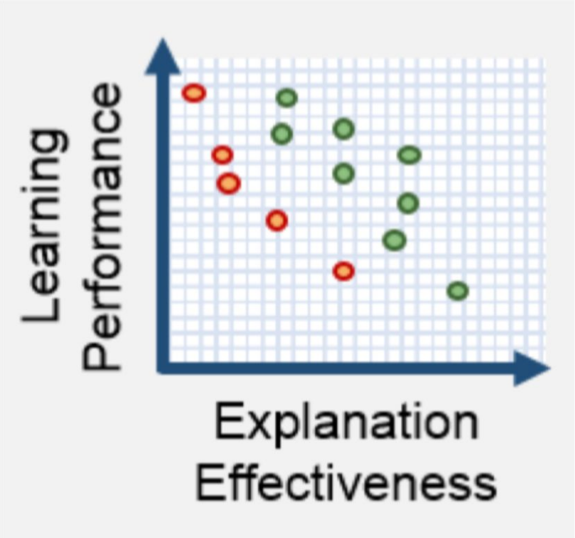
- Models Difficult for humans to comprehend (Eg. Deep Learning models)
- Proprietary Models (Eg. COMPAS)

Useful Definitions :

Explainable ML : Post-hoc Model to explain first black box model
Interpretable ML : Inherently interpretable and provides its own reasonings.

Key Issues with Explainable ML :

- Myth about Accuracy Vs Interpretability Trade-off



This was presented at DARPA XAI (Explainable AI) Board Agency Announcements showing a smooth negative correlation between Learning Performance and Explanation Effectiveness. However is this fair / informative enough ?
The x-axis and y-axis is unknown here and this was probably based on some static dataset where it's unknown how the ML Algorithms were selected for the comparison. Is it fair to even compare a highly interpretable 1984 CART model to a 2018 Deep Learning model ?

It's also important to consider the role of data here. It's actually seen that the performance of both Black box models and Interpretable models is quite similar in the case of structured, clean data. It's also well known that with multiple iterations of data processing based on identified issues, possibly with the help of interpretable models could increase the accuracy of the model manifold. This completely contradicts the myth of Accuracy and Interpretability Trade-off !

- Explainable ML Faithfulness to Original Model

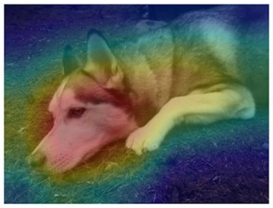
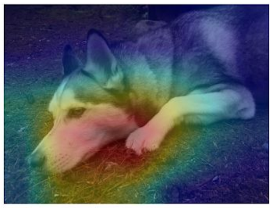
The primary reason for explanations is to evolve a sense of trust on these “Black box” models. However, these explanation models are different from the original models and might sometimes make mistakes wherein the explanations for predictions might not make much sense even if the predictions might not be that wrong. This could create a notion of “Distrust” on the original Black Box model and not serve the purpose.

Consider the case of criminal recidivism decisions where COMPAS, a proprietary model is used by the US Justice system broadly to make parole/bail decisions. Pro-Publica Analysis accused COMPAS of Racism and showed a linear dependency of COMPAS Results on Race. They approximated COMPAS predictions and provided an explanation saying “***This person is predicted to be arrested because they are black.***“. Is this explanation at all helpful to a judge ? Should this even be called an “explanation” ? This completely ignores the fact that the features used by this separate model could be completely different from the original ones used by COMPAS. The linear assumptions were also not accurate as COMPAS is apparently a nonlinear model created by experts. The fact that primary features used in criminal recidivism cases is generally age and criminal history which could possibly have a correlation with race in the datasets is also something to consider which underscores the fact that COMPAS might not be taking race as a feature explicitly at all. Hence making this “explanation” a “summary statistics” at best !

If this was an interpretable model instead, such bias/unbias would have been much clearer making the debate easier.

- Explanations might not always make sense

Suppose, the original model predicts correctly and the explanation model also tries its best to explain as per its algorithm. One might need to look at the informativeness / enoughness in order to be able to trust the Black Box Model results. Consider saliency maps, used in multiple low stakes cases.

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

The explainable model is given a test image and asked to show the region the neural network looked to predict it as a Siberian Husky. This result seems promising as in the above figure. However, when the neural network is asked to show where it looked for the class Transverse Flute, it still shows the same region. This is really confusing and makes us want to disbelieve the black box ML model results due to sheer incompleteness of the explanation.

- Low compatibility of black box models with new information based revisions

With Interpretable models, it might be very easy to show the reasons for a particular decision. For example a judge would know if the seriousness of a crime was factored in while making bail decisions. In case of any change of circumstances as well, factoring in this change becomes really difficult in the case of Black Box models where there are a lot of unknowns.

- Overly Complicated Decision Pathway ripe to Human Error

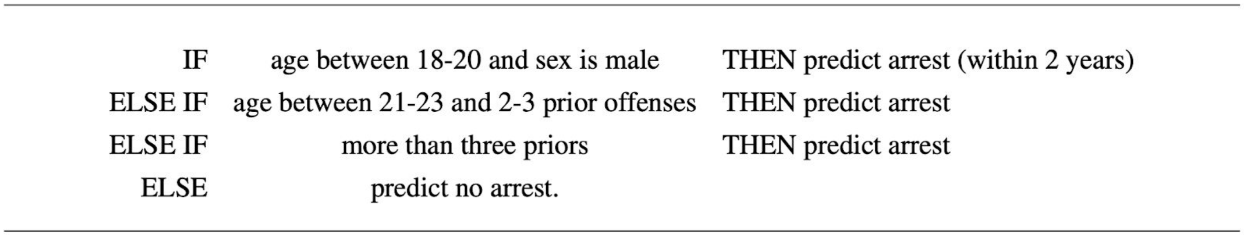
  Consider the COMPAS example, where it depends on ~130+ factors , some of which are human filled surveys. These surveys are highly susceptible to typographical errors. It is these kinds of errors that could lead the model to randomly / incorrectly provide decisions of bail. This is looked at as “procedural unfairness”.

Also, imagine troubleshooting such scenarios !! We would have a black-box model providing random predictions and an explanation model giving incomplete explanations making troubleshooting a complete nightmare.

Key Issues with Interpretable ML :

- Profit afforded to Black Box Models are not in favour of Interpretable Models

Consider CORELS (Certifiably Optimal Rule Lists) :



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.

This simple algorithm has similar accuracy to COMPAS which is a sophisticated model with 130+ factors. But, would any corporate be able to get paid for such a simple model ?
There is a conflict of interest presented wherein such companies like COMPAS, Breezometer don't directly suffer due to their incorrect predictions and wish to profit with black box models lacking interpretability to avoid catastrophic failures.

- High Efforts required to construct Interpretable Models

Constructing interpretable models requires domain expertise, to even define what interpretability means in the specific domain. Interpretability constraints like sparsity, causality , additivity, etc make this a computationally hard problem in the worst case. However, it might actually be worthwhile to invest in this for a high stakes domain, so that much more expensive mistakes made later are avoided.

- Black Boxes seem to uncover hidden patterns

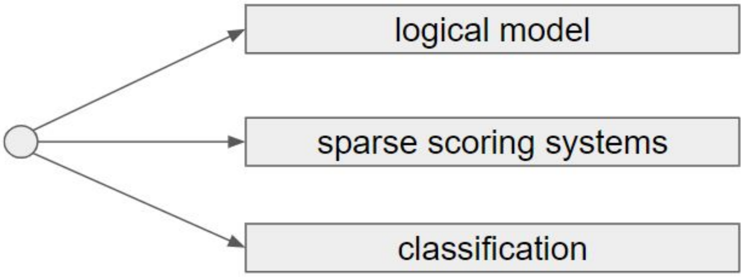
Interpretable models could also uncover similar patterns in data if these patterns were important enough to be leveraged for predictions. It majorly depends on the researcher's ability to create accurate-yet-interpretable models

Encouraging Responsible ML Governance: Two Proposals

Regulation does not need interpretability, the explanation is enough, So author has two proposals:
(1) For certain high-stakes decisions, no black box should be deployed when there exists an interpretable model with the same level of performance.(stressful)

(2) Let us consider the possibility that organizations that introduce black box models would be mandated to report the accuracy of interpretable modeling methods. (less stressful)

Algorithmic Challenges in Interpretable ML: Three cases



(1) logical models: consists of statements involving “or,” “and,” “if-then,” etc.

(such as Decision trees)

$$\min_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} 1_{[\text{training observation } i \text{ is misclassified by } f]} + \lambda \times \text{size}(f) \right)$$

The challenge is whether we can solve (or approximately solve) problems like this in practical ways by leveraging new theoretical techniques and advances in hardware.

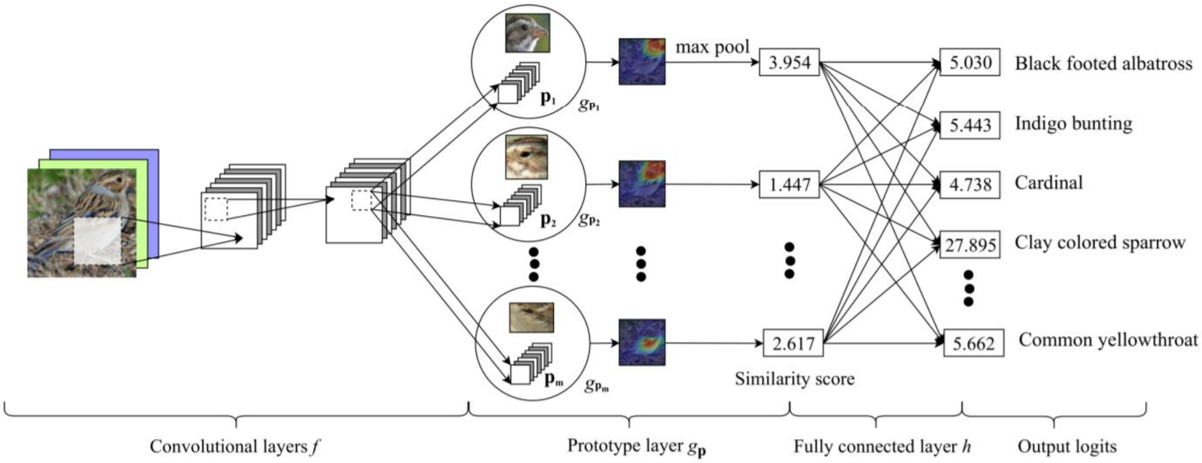
(2) sparse scoring systems: a sparse linear model with integer coefficients – the coefficients are the point scores.

$$\min_{b_1,b_2,..,b_p\in\{-10,-9,...,9,10\}} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-\sum_{j=1}^{p} b_j x_{i,j}\right)\right) + \lambda\sum_{j} 1_{[b_j\neq 0]},$$

the second challenge is to create algorithms for scoring systems that are computationally efficient

(3) Classification

The network must then make decisions by **reasoning about parts of the image** so that the explanations are real, and **not post hoc**.



Assumption of Interpretable Models Might Exist

Rashomon set is the set of reasonably accurate predictive models (say within a given accuracy from the best model accuracy) which has many close-to-optimal models that predict differently from each other, e.g. RF, NN, SVM. So with high probability it contains interpretable models, and interpretable accurate models.

Algorithm Stability

The instability might not be a disadvantage since it provides a large Rashomon set, (such as decision trees which are sensitive to data) so that we can select both accurate and interpretable models if we have great strategies.

Conclusion

The paper appeals that we should pay more attention and give more efforts to interpretability rather than explanation in both academic and industrial fields.