What assumptions do we make when using black box predictive models?

Cynthia Rudin

Duke University

When we use a black box predictive model, we assume:

- … that the cost of the decision is low. Otherwise, we would build a model whose calculations we can easily double and triple check.
- … that information is correctly entered into the model.

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

232

Glenn Rodriguez was denied parole because of a miscalculated “COMPAS” score.

137 factors entered by hand for each survey

1% error rate → 75% chance of at least one typo on a survey

This is a serious disadvantage to complicated or proprietary models. "XAI" won't help.

In Florida….?

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

232

Name	COMPAS Violent Decile	# Arrests	# Charges	Selected Prior Charges	Selected Subseq. Charges
Shirley Darby	1	2	4	Aggravated Battery (F,1), Child Abuse (F,1), Resist Officer w/Violence (F,1)	
Joseph Salera	1	8	14	Battery on Law Enforc Officer (F,3), Aggravated Assault W/Dead Weap (F,1), Aggravated Battery (F,1), Resist/obstruct Officer W/viol (F,1)	
Bart Sandell	1	9	15	Attempted Murder 1st Degree (F,1), Resist/obstruct Officer W/viol (F,1), Agg Battery Grt/Bod/Harm (F,1), Carrying Concealed Firearm (F,1)	Armed Sex Batt/vict 12 Yrs + (F,2), Aggravated Assault W/dead Weap (F,3), Kidnapping (F,1)
Miguel Wilkins	1	11	22	Aggrav Battery w/Deadly Weapon (F,1), Driving Under The Influence (M,2), Carrying Concealed Firearm (F,1)	
Jonathan Gabbard	1	7	28	Robbery / Deadly Weapon (F,11), Poss Firearm Commission Felony (F,7)	
Brandon Jackel	1	22	40	Resist/obstruct Officer W/viol (F,3), Battery on Law Enforc Officer (F,2), Attempted Robbery Deadly Weapon (F,1), Robbery 1 / Deadly Weapon (F,1)	
Fernando Galarza	2	2	6	Murder in the First Degree (F,1), Aggrav Battery w/Deadly Weapon (F,1), Carrying Concealed Firearm (F,1)	

Continued on next page

Name	COMPAS Violent Decile	# Arrests	# Charges	Selected Prior Charges	Selected Subseq. Charges
Nathan Keller	2	8	17	Aggravated Assault (F,5), Aggravated Assault W/dead Weap (F,2), Shoot/throw Into Vehicle (F,2), Battery Upon Detainee (F,1)	
Zachary Campanelli	2	11	21	Armed Trafficking In Cocaine (F,1), Poss Weapon Commission Felony (F,1), Carrying Concealed Firearm (F,1)	
Aaron Coleburn	2	16	25	Attempt Murder in the First Degree (F,1), Carrying Concealed Firearm (F,1), Felon in Pos of Firearm or Amm (F,1)	
Bruce Poblano	2	22	39	Aggravated Battery (F,3), Robbery / Deadly Weapon (F,3), Kidnapping (F,1), Carrying Concealed Firearm (F,2)	Grand Theft in the 3rd Degree (F,3)
Phillip Sperry	3	11	16	Aggravated Assault W/dead Weap (F,1), Burglary Damage Property>$1000 (F,1), Burglary Unoccupied Dwelling (F,1)	
Dylan Azzi	3	11	17	Aggravated Assault W/dead Weap (F,2), Aggravated Assault w/Firearm (F,2), Discharge Firearm From Vehicle (F,1), Home Invasion Robbery (F,1)	Fail Register Vehicle (M,2)
Russell Michaels	3	9	23	Solicit to Commit Armed Robbery (F,1), Armed False Imprisonment (F,1), Home Invasion Robbery (F,1)	Driving While License Revoked (F,3)
Bradley Haddock	3	15	25	Attempt Sexual Batt / Vict 12+ (F,1), Resist/obstruct Officer W/viol (F,1), Poss Firearm W/alter/remov Id# (F,1)	
Randy Walkman	3	24	36	Murder in the First Degree (F,1), Poss Firearm Commission Felony (F,1), Solicit to Commit Armed Robbery (F,1)	Petit Theft $100-300$ (M,1)
Carol Hartman	4	5	16	Aggrav Battery w/Deadly Weapon (F,1), Felon in Pos of Firearm or Amm (F,4)	Resist/Obstruct W/O Violence (M,1), Possess Drug Paraphernalia (M,1)

Possibly typos in the COMPAS documentation from Northpointe?

COMPAS Documentation

> Violent Recidivism Risk Score
>
> $$= (\text{age} * -\text{w}) + (\text{age-at-first-arrest} * -\text{w}) + (\text{history of violence} * \text{w})$$
> $$+ \quad (\text{vocation education} * \text{w}) + (\text{history of noncompliance} * \text{w})$$

Corrected version?

> Violent Recidivism Risk Score
>
> $$= (f(\text{age}) * -\text{w}) + (g(\text{age-at-first-arrest}) * -\text{w}) + (\text{history of violence} * \text{w})$$
> $$+ \quad (\text{vocation education} * \text{w}) + (\text{history of noncompliance} * \text{w}) \,,$$

where *f* and *g* are proprietary transformations of age, such as linear splines?

When we use a black box predictive model, we assume:

- … that the cost of the decision is low. Otherwise, we would build a model whose calculations we can easily double and triple check.
- … that information is correctly entered into the model.
- … the dataset is trustworthy. It is not.

Innovate Healthcare
HealthImaging
INSIGHTS IN IMAGING & INFORMATICS

Algorithm's 'unexpected' weakness raises larger concerns about AI's potential in broader populations

Matt O'Connor | *April 05, 2021* | *Artificial Intelligenc*

Deep learning detects intercranial hemorrhages

npj | digital medicine

Explore content About the journal Publish with us

nature > npj digital medicine > articles > article

Article | Open Access | Published: 30 April 2019

Deep learning predicts hip fracture using confounding patient and healthcare variables

Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder & Joel T. Dudley

npj Digital Medicine **2**, Article number: 31 (2019) | Cite this article

process data. If CAD algorithms are inexplicably leveraging patient and process variables in their predictions, it is unclear how radiologists should interpret their predictions in the context of other known patient data. Further research is needed to illuminate deep-learning decision processes so that computers and clinicians can effectively cooperate.

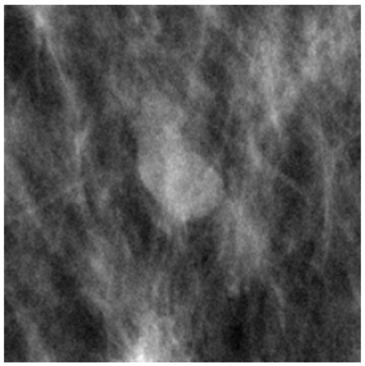
I propose something radically different: interpretable deep neural networks for radiology.

- Coming up:

(1) black box

(2) XAI-style ”explained” black box

(3) interpretable deep neural network

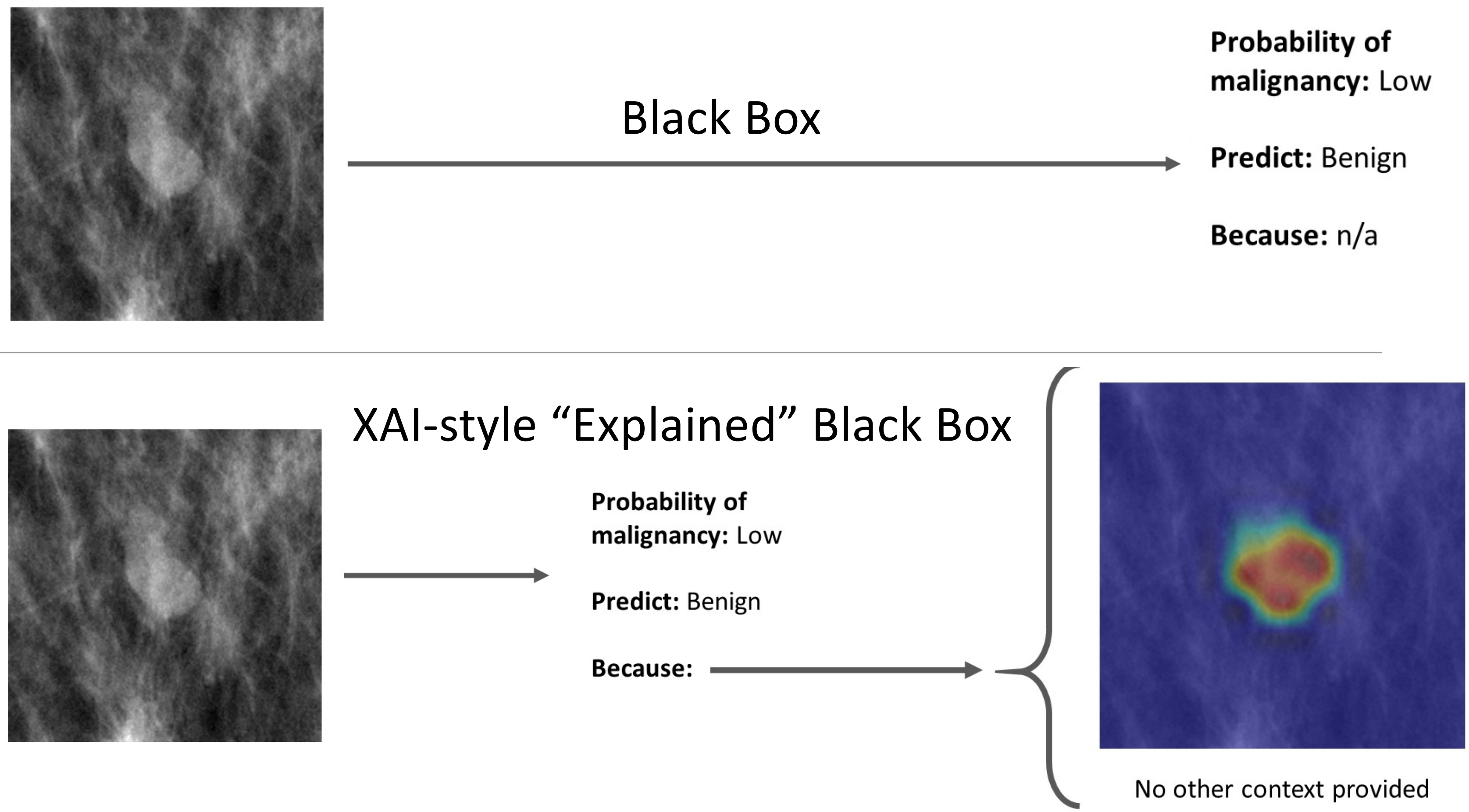


Black Box

Probability of malignancy: Low

Predict: Benign

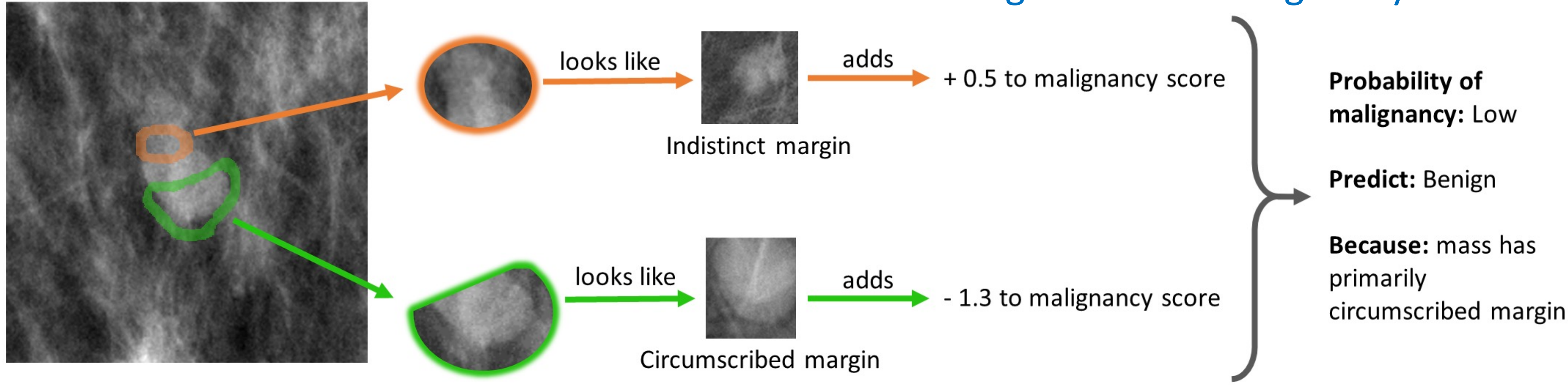
Because: n/a


Black Box
Probability of malignancy: Low
Predict: Benign
Because: n/a
XAI-style “Explained” Black Box
Probability of malignancy: Low
Predict: Benign
Because:
No other context provided

Interpretable AI algorithm for Breast Lesions (IAIA-BL)

Prototypes

Model decomposes to predict margins before malignancy



Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, Cynthia Rudin
IAIA-BL: A Case-based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography,
Nature Machine Intelligence, Accepted, 2021.

When we use a black box predictive model, we assume:

- … that the cost of the decision is low. Otherwise, we would build a model whose calculations we can easily double and triple check.
- … that information is correctly entered into the model.
- … the dataset is trustworthy. It is not.
- … that reported accuracy scores represent the population of interest (e.g., IVF).

Multicenter Study > Hum Reprod. 2019 Jun 4;34(6):1011-1018. doi: 10.1093/humrep/dez064.

Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer

D Tran [1], S Cooke [2], P J Illingworth [2], D K Gardner [3]

Affiliations + expand

PMID: 31111884 PMCID: PMC6554189 DOI: 10.1093/humrep/dez064

Free PMC article

Abstract

Study question: Can a deep learning model predict the probability of pregnancy with fetal heart (FH) from time-lapse videos?

Main results and the role of chance: The deep learning model was able to predict FH pregnancy from time-lapse videos with an AUC of 0.93 [95% CI 0.92-0.94] in 5-fold stratified cross-validation. A hold-out validation test across eight laboratories showed that the AUC was reproducible, ranging from 0.95 to 0.90 across different laboratories with different culture and laboratory processes.

Ethical Implementation of Artificial Intelligence to Select Embryos in *In Vitro Fertilization*

Michael Anis Mihdi Afnan
Department of Medicine
Imperial College London
London, UK
michaelafnan@icloud.com

Cynthia Rudin
Departments of Computer Science, Electrical Engineering and Statistical Science
Duke University
Durham, North Carolina, USA
cynthia@cs.duke.edu

Vincent Conitzer
Departments of Computer Science, Economics and Philosophy & Institute for Ethics in AI and Departments of Computer Science and Philosophy
Duke University & Oxford University
Durham, North Carolina, USA
conitzer@cs.duke.edu

Julian Savulescu
Uehiro Centre for Practical Ethics & Wellcome Centre for Ethics and Humanities & Murdoch Children's Research Institute
Oxford University & Oxford University & Royal Children's Hospital
Oxford, UK
julian.savulescu@philosophy.ox.ac.uk

Abhishek Mishra
Uehiro Centre for Practical Ethics
Oxford University
Oxford, UK
abhishek.mishra@philosophy.ox.ac.uk

Yanhe Liu
Monash IVF Group & School of Human Sciences & School of Medical and Health Sciences
Monash IVF Group & University of Western Australia & Edith Cowan University
Southport, Australia
gift0409@yahoo.com.au

Masoud Afnan
Department of Obstetrics and Gynaecology
Qingdao United Family Hospital
Qingdao, China
masoudafnan@me.com

Comment > Hum Reprod. 2020 Jun 1;35(6):1473. doi: 10.1093/humrep/deaa083.

Can deep learning automatically predict fetal heart pregnancy with almost perfect accuracy?

Yoav Kan-Tor [1], Assaf Ben-Meir [2], Amnon Buxboim [1] [3] [4]

Affiliations + expand

PMID: 32458001 DOI: 10.1093/humrep/deaa083

Adding "obvious" cases artificially inflates the performance!

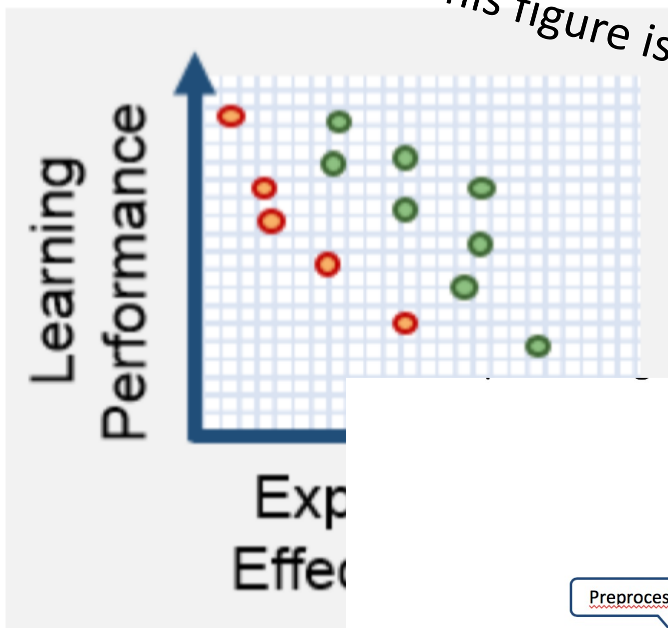
When we use a black box predictive model, we assume:

- … that the cost of the decision is low. Otherwise, we would build a model whose calculations we can easily double and triple check.
- … that information is correctly entered into the model.
- … the dataset is trustworthy. It is not.
- … that reported accuracy scores represent the population of interest (e.g., IVF).
- … that AI is incapable of explaining itself while maintaining accuracy.

The Accuracy/Interpretability Tradeoff is a Myth

This figure is phony baloney

The tradeoff doesn't happen like this

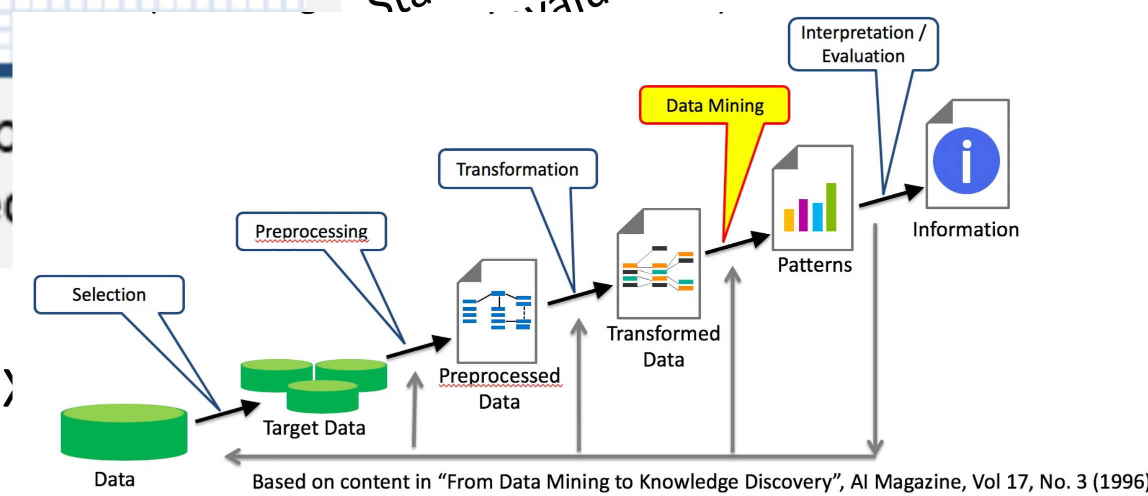


Static dataset?

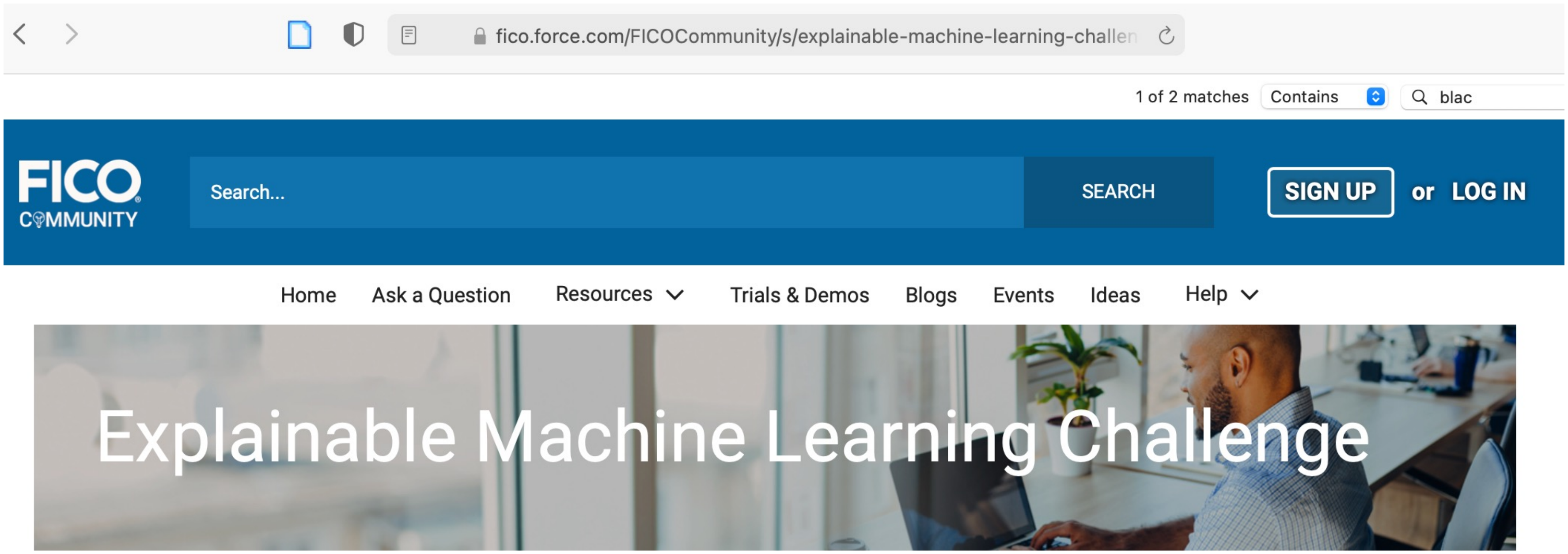
valuation metric?

Are they talking about explaining black boxes?

From the DARPA X



Based on content in "From Data Mining to Knowledge Discovery", AI Magazine, Vol 17, No. 3 (1996)
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230



Home Equity Line of Credit (HELOC) Dataset

This competition focuses on an anonymized dataset of Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price). The customers in this dataset have requested a credit line in the range of $5,000 - $150,000. The fundamental task is to use the information about the applicant in their credit report to predict whether they will repay their HELOC account within 2 years. This prediction is then used to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.

About the data

- ~10K loan applicants
- Factors:
  - External Risk Estimate
  - Months Since Oldest Trade Open
  - Months Since Most Recent Trade Open
  - Average Months In File
  - Number of Satisfactory Trades
  - Number Trades 60+ Ever
  - Number Trades 90+ Ever
  - Number of Total Trades
  - Number Trades Open In Last 12 Months
  - Percent Trades Never Delinquent
  - Months Since Most Recent Delinquency
  - Max Delinquency / Public Records Last 12 Months
  - Max Delinquency Ever
  - Percent Installment Trades
  - Net Fraction of Installment Burden
  - Number of Installment Trades with Balance
  - Months Since Most Recent Inquiry excluding 7 days
  - Number of Inquiries in Last 6 Months
  - Number of Inquiries in Last 6 Months excluding 7 days.
  - Net Fraction Revolving Burden. (Revolving balance divided by credit limit.)
  - Number Revolving Trades with Balance
  - Number Bank/Natl Trades with high utilization ratio
  - Percent of Trades with Balance

Best black box accuracy
(boosted decision trees) 73%

Best black box AUC
(2-layer neural network) .80

Best black box accuracy
(boosted decision trees) 73%

Best black box AUC
(2-layer neural network) .80

IBM model (First Prize): 6 questions
Accuracy = 71.8%
AUC = .62

Our entry (won FICO Recognition Prize):
Two-layer additive risk model
10 subscales + one final scoring model

Accuracy = 73.8%
AUC = .806

Go to http://dukedatasciencefico.cs.duke.edu

When we use a black box predictive model, we assume:

- … that the cost of the decision is low. Otherwise, we would build a model whose calculations we can easily double and triple check.
- … that information is correctly entered into the model.
- … the dataset is trustworthy. It is not.
- … that reported accuracy scores represent the population of interest (e.g., IVF).
- … that AI is incapable of explaining itself while maintaining accuracy.
- … that we can explain the black box.

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

There is no scientific evidence for a general tradeoff between accuracy and interpretability

models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are inter-

Even for deep learning in computer vision, interpretable models can be built at the same accuracy as a black box deep neural network

For tabular data, most machine learning methods are equally accurate, including sparse models.

diction applications that deeply impact human lives. Many of

models that are lightly constrained in model form (such as models

Explaining a black box gives it unnecessary authority.

When we use a black box predictive model, we assume:

- … that the cost of the decision is low. Otherwise, we would build a model whose calculations we can easily double and triple check.
- … that information is correctly entered into the model.
- … the dataset is trustworthy. It is not.
- … that reported accuracy scores represent the population of interest (e.g., IVF).
- … that AI is incapable of explaining itself while maintaining accuracy.
- … that we can explain the black box.

Luckily…

We don’t need a black box.