→ explainability
→ interpretability

Model Interpretation

Machine Learning for Data Science 1

explainable AI

interpretability:

① the extend to which cause and effect can be observed within a system

② the extend to which we are able to predict what is going to happen upon changes of the input

③ the degree to which us (human) can predict the model's result

exploitability:
extend to which the internal mechanics of the system can be exploited in human terms

background knowledge

example: chemical experiment

interpretability:
I know the protocol of the experiment

expl: understand the chemistry

Why?

- transparency, or the lack of
  - corporate self-interest
  - absence of governance and accountability

= ML gets involved in biggest businesses and politics
  ↳ Cambridge Analytica

= Criminal justice, medicine, healthcare, computer vision
  → accountability
  → transparency
  → trust

= understand the workings of the model

Mistakes

MIT Technology Review

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** January 21, 2019

AI might not seem to have a huge personal impact if your most frequent brush with machine-learning algorithms is through Facebook's news feed or Google's search rankings. But at the Data for Black Lives conference last weekend, technologists, legal experts, and community activists snapped things into perspective with a discussion of America's criminal justice system. There, an algorithm can determine the trajectory of your life.

The US imprisons more people than any other country in the world. At the end of 2016, nearly 2.2 million adults were being held in prisons or jails, and an additional 4.5 million were in other correctional facilities. Put another way, 1 in 38 adult Americans was under some form of correctional supervision. The nightmarishness of this situation is one of the few issues that unite politicians on both sides of the aisle.

FIRES

How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say

BY MICHAEL MCGOUGH

AUGUST 07, 2018 09:26 AM,

Smoke is affecting air quality all over California. Here's what it looks like at the Carr Fire, north of Redding, on July 31, 2018.

BY PAUL KITAGAKI JR.

INVESTING

Shares for another company called Zoom are flying, but some might be trading the wrong stock

PUBLISHED THU, APR 18 2019•11:19 AM EDT | UPDATED THU, APR 18 2019•4:27 PM EDT

Michael Sheetz
@THESHEETZTWEETZ

KEY POINTS

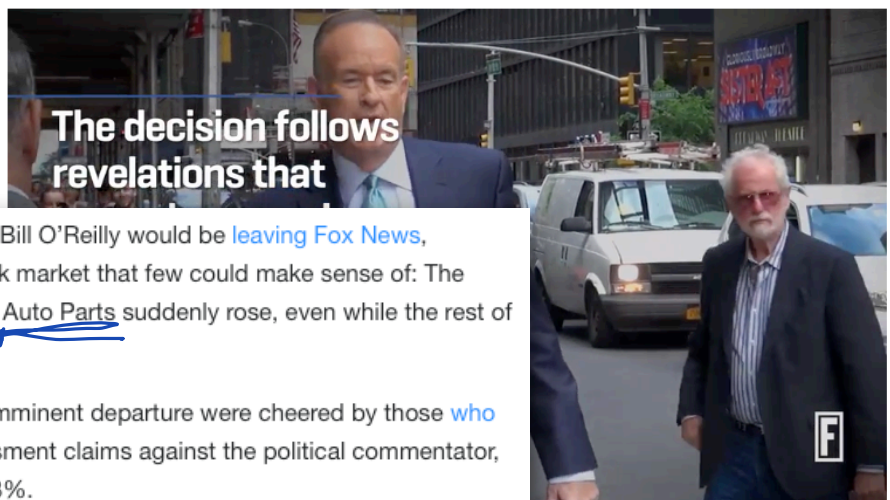
- Zoom Technologies (ticker ZOOM) is not the company Zoom Video Communications (ticker ZM) that began publicly trading Thursday on the Nasdaq.
- The former is a tiny Chinese wireless communications company that "does not have significant operations," according to its profile listing on Yahoo.
- Shares of Zoom Technologies hit a trading volume on Th
the amount of shares that change hands on the average

FINANCE • FOX NEWS

How Bill O'Reilly Leaving Fox Fired Up O'Reilly Auto Parts Stock

BY **JEN WIECZNER**

April 22, 2017 12:18 AM GMT+2



As news broke this week that TV anchor Bill O'Reilly would be leaving Fox News, investors observed a reaction in the stock market that few could make sense of: The stock of the company that owns O'Reilly Auto Parts suddenly rose, even while the rest of the market was falling.

On Wednesday, as reports of O'Reilly's imminent departure were cheered by those who pushed for his ouster amid sexual harassment claims against the political commentator, O'Reilly Automotive gained as much as 3%.

black-box models

1. a function too complicated for a human to comprehend
2. a proprietary function
   - ranking of the web pages
   - psychiatry ← prohibited
   - medicine

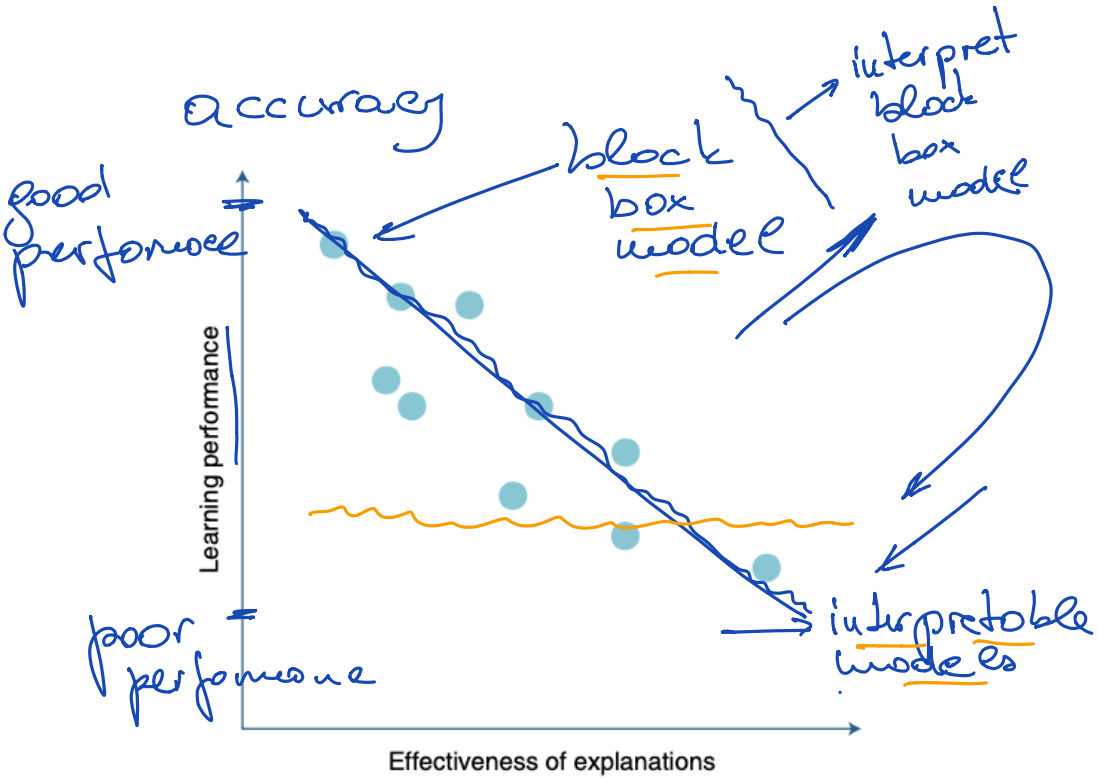


Fig. 1 | A fictional depiction of the accuracy–interpretability trade-off. Adapted from ref. [18], DARPA.

other benefits

- fairness, predictions not biased and do not discriminate (demographic, race...)
- privacy, exclude any private info
- ~~tautology~~
- reliability & robustness: small changes of input lead to large changes of output?
- causation
- trust

Some taxonomy

Model-specific ⟷ Model agnostic

Model-specific
↓
logistic reg. ⟸
ANN and ⟸

Model agnostic
↓
Interpretation capability does not depend on the type of the model
↓
black-boxes
(no internals used)

Local ⟷ Global

Local
Can you interpret prediction of $\underline{x}$
↓
$\hat{y}$? how?

Global
present the (whole) model in interpretable form

Evaluation, Doshi-Velez (2017)

- application level, did interpretation help
  requires a very good experimental setup
- human-level evaluation, layperson
- function level, optimize for simplicity

Properties of methods for interpretation

- expressive power
- translucency
- portability
- complexity

Properties of methods for individual explanations

- accuracy
- fidelity
- consistency
- stability
- comprehensibility
- certainty
- degree of importance
- novelty & background knowledge

LIME , Local Surrogate

black box
part of the feature space

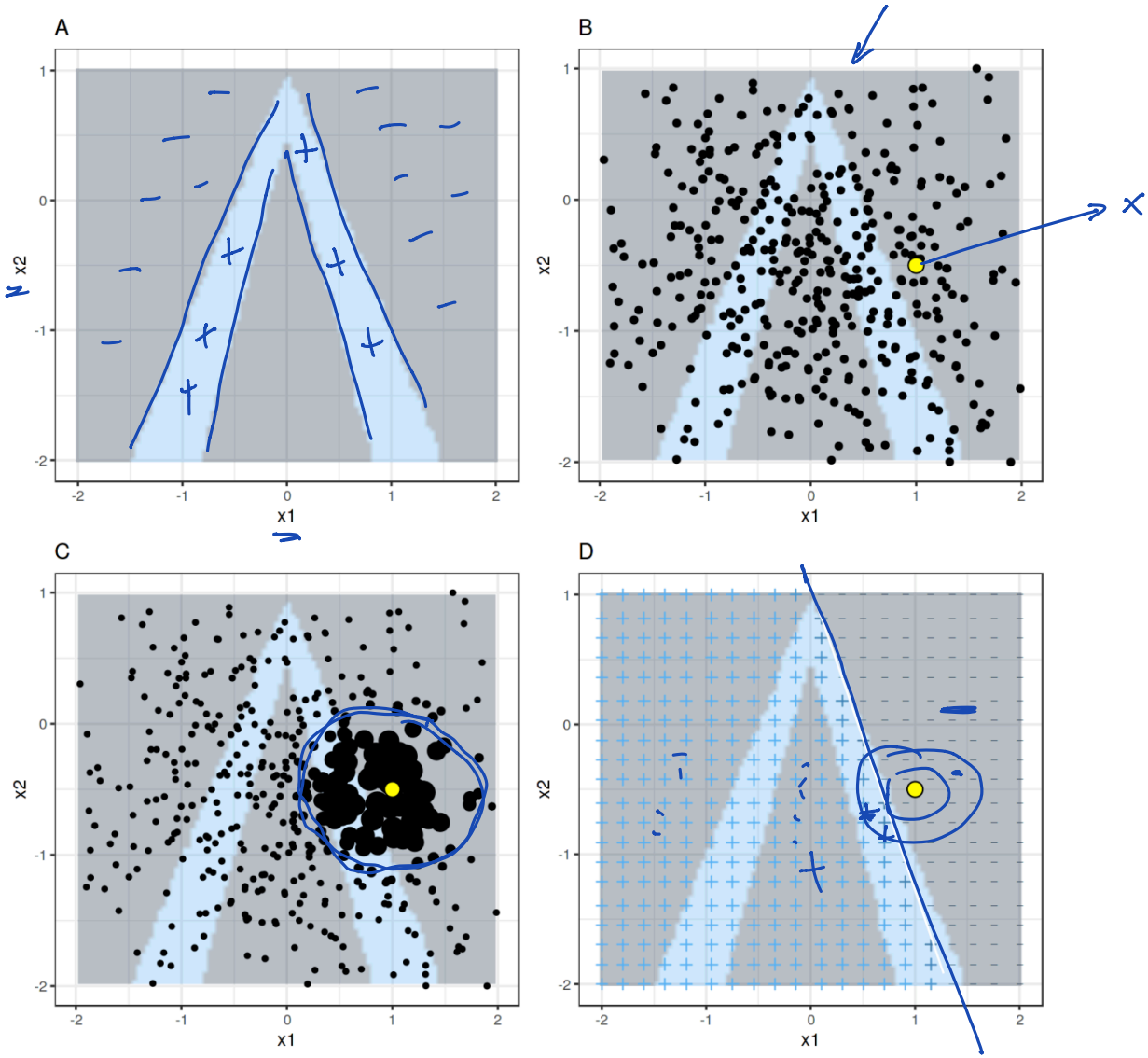
→ local interpretable model agnostic explanations

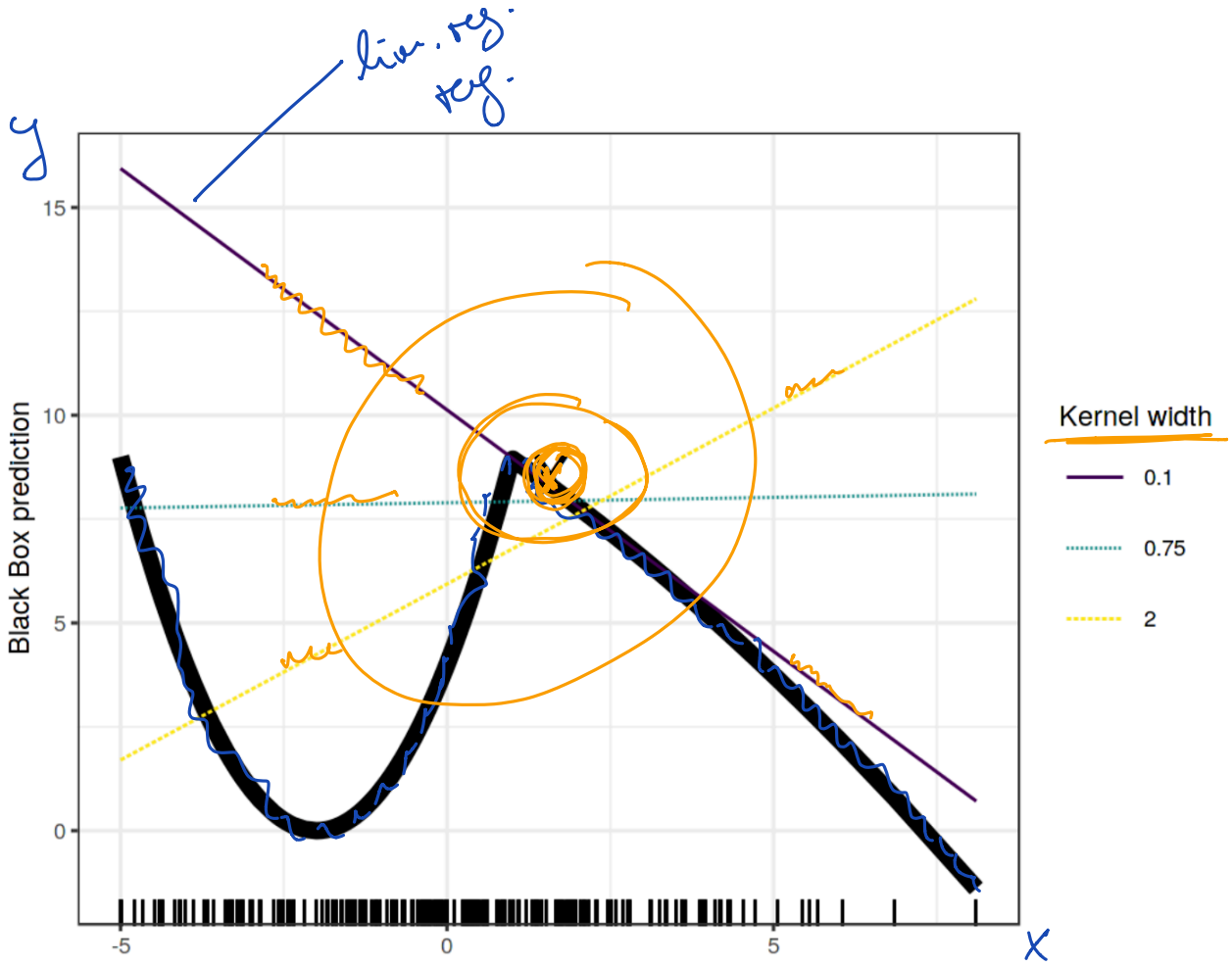
↓

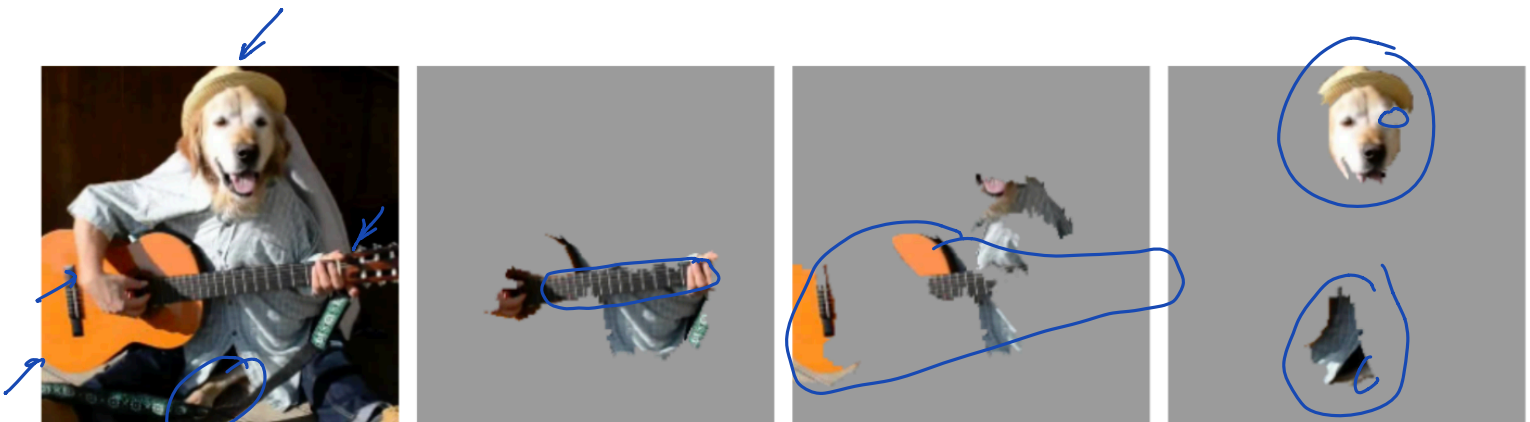
explain individual predictions

1. select an instance (row) of interest, x
2. perturb the data and get predictions
3. weight new samples according to similarity to x
4. weighted perturbations + outcomes of our BBM → interpretable model

tree, lin reg.
log reg. NBC


A
B
C
D
x1
x2
-2
-1
0
1
2


Black Box prediction
15
10
5
0
-5
0
5
X
Kernel width
0.1
0.75
2



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

LIME explanations for the top 3 classes for image classification made by Google's Inception neural network. The example is taken from the LIME paper (Ribeiro et. al., 2016).

LIME

Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

Shapley Values : effects of the feature values on the outcome of a model

each feature as a player in a game
distribute the payout among the feature

coalition game

↳ feature weighting

n players
v maps a subset of players to a real number
$v: 2^n \to \mathbb{R}$, $v(\emptyset) = 0$

Shapley value: distribute total gains to players (features) assuming they collaborate

S: coalition of players
$v(S)$: worth of coalition

The amount the player i gets in coalition game (v, n)

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n-|S|-1)!}{n!} \Big(v(S \cup \{i\}) - v(S) \Big)$$

all possible coalitions

1000

approximation (Štrumbelj, Kononenko 2011)

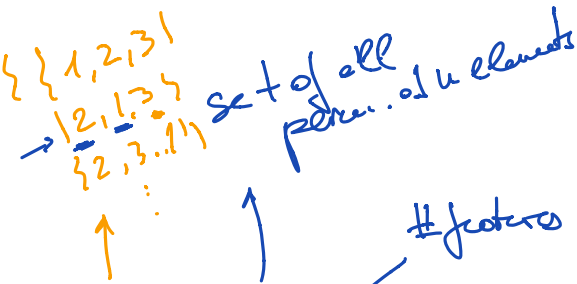
$$\phi_i(v, x) = \frac{1}{M} \sum_{m=1}^{M} \Big(\hat{f}(x_{+i}^{m}) - f(x_{-i}^{m}) \Big)$$

Algorithm, approximate $\phi_i(x)$, the importance of the i-th feature's value for instance x and for model f

$\phi_i(x) \leftarrow 0$

for $k = 1$ to K

select a random permutation of $S \in \Pi(n)$ and some random instance $y \in X$



$x_{+i} \leftarrow$ [feat. preceding i in S | i | features after i in S]

come from x | come from y

$x_{-i} \leftarrow$ [| |]

come from x | come from y

Local importance

$$\phi_i(x) = \phi_i(x) + f(x_{+i}) - f(x_{-i})$$

$$\phi_i(x) = \phi_i(x) / K$$

Algorithm: approximate $\phi_{i,j}$, global importance of the i-th feature value j for a model f

$\phi_{i,j} \leftarrow 0$

for $k = 1$ to K

select random instance $y \in X$

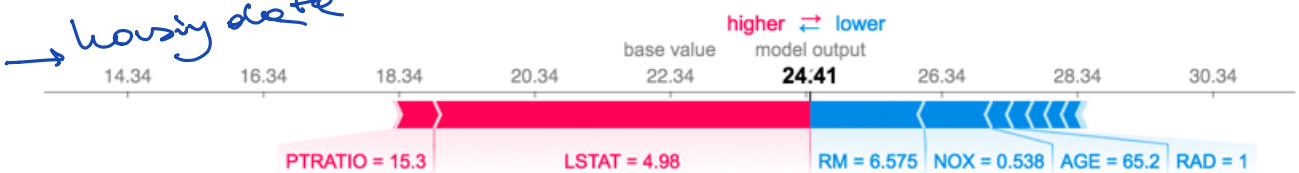
$x_1 \leftarrow$ set i-th value to j, take all other values from y

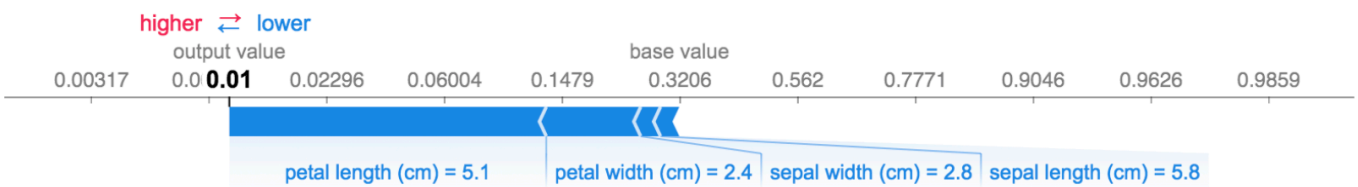
$\phi_{i,j} \leftarrow \phi_{i,j} + f(x_1) - f(y)$

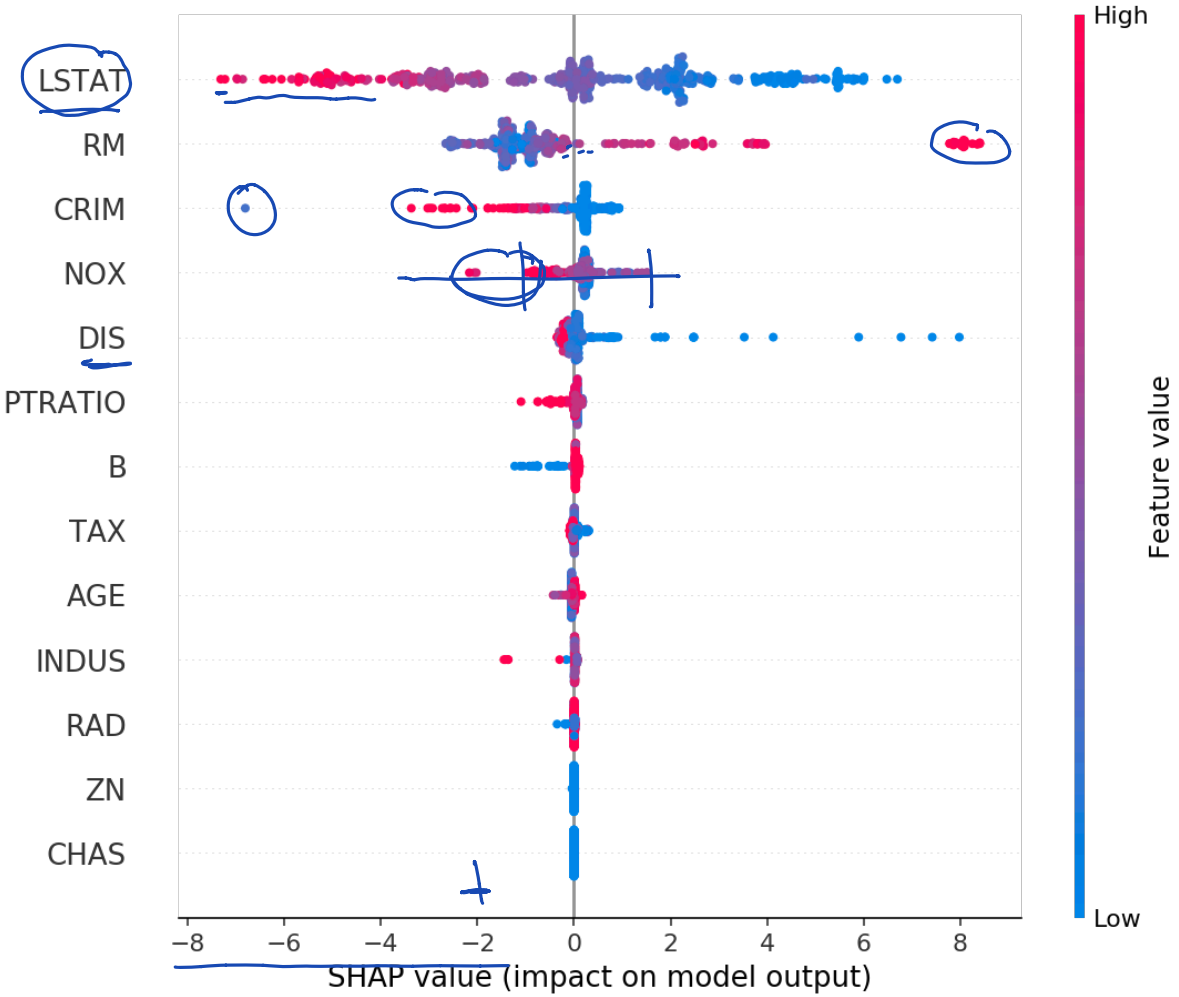
$\phi_{i,j} \leftarrow \phi_{i,j} / K$

SHAP

→ housing data






LSTAT
RM
CRIM
NOX
DIS
PTRATIO
B
TAX
AGE
INDUS
RAD
ZN
CHAS
−8
−6
−4
−2
0
2
4
6
8
SHAP value (impact on model output)
High
Feature value
Low

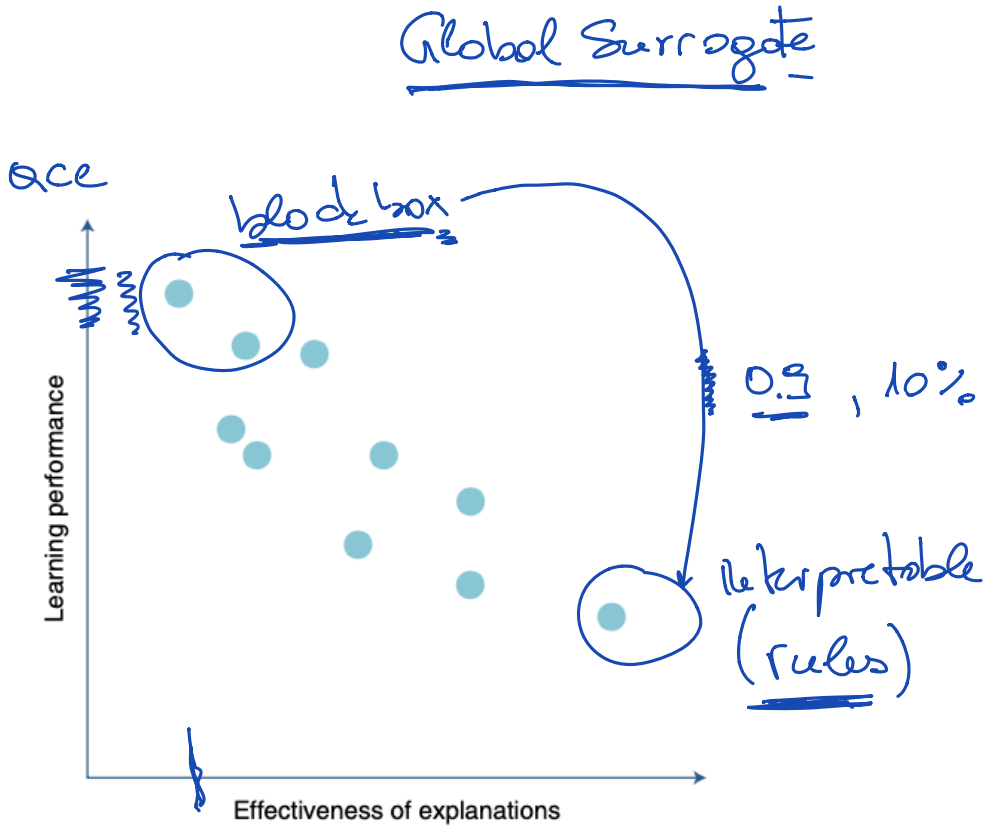


Fig. 1 | A fictional depiction of the accuracy–interpretability trade-off. Adapted from ref. [18], DARPA.

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

$$\min_{f \in \mathcal{F}} \left(\frac{1}{N} \sum 1[\text{training case is misclassified}] + \lambda \, \text{size}(f) \right)$$

error — Complexity

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

RiskSLIM

Table 2 | Comparison of COMPAS and CORELS models

COMPAS	CORELS
Black box; 130+ factors; might include socio-economic info; expensive (software licence); within software used in US justice system	Full model is in Table 1; only age, priors, gender (optional); no other information; free, transparent

Table 1 | Machine learning model from the CORELS algorithm

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 offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

This model from ref. [39] is the minimizer of a special case of equation (1) discussed later in the challenges section. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data from Florida needed to produce this model.

Table 3 | Scoring system for risk of recidivism

1.	Prior arrests ≥ 2	1 point	
2.	Prior arrests ≥ 5	1 point	
3.	Prior arrests for local ordinance	1 point	
4.	Age at release between 18 to 24	1 point	
5.	Age at release ≥ 40	−1 point	
		Score	

Score	−1	0	1	2	3	4
Risk (%)	11.9	26.9	50.0	73.1	88.1	95.3

This system is from ref. [21], which was developed from refs. [29,46]. The model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

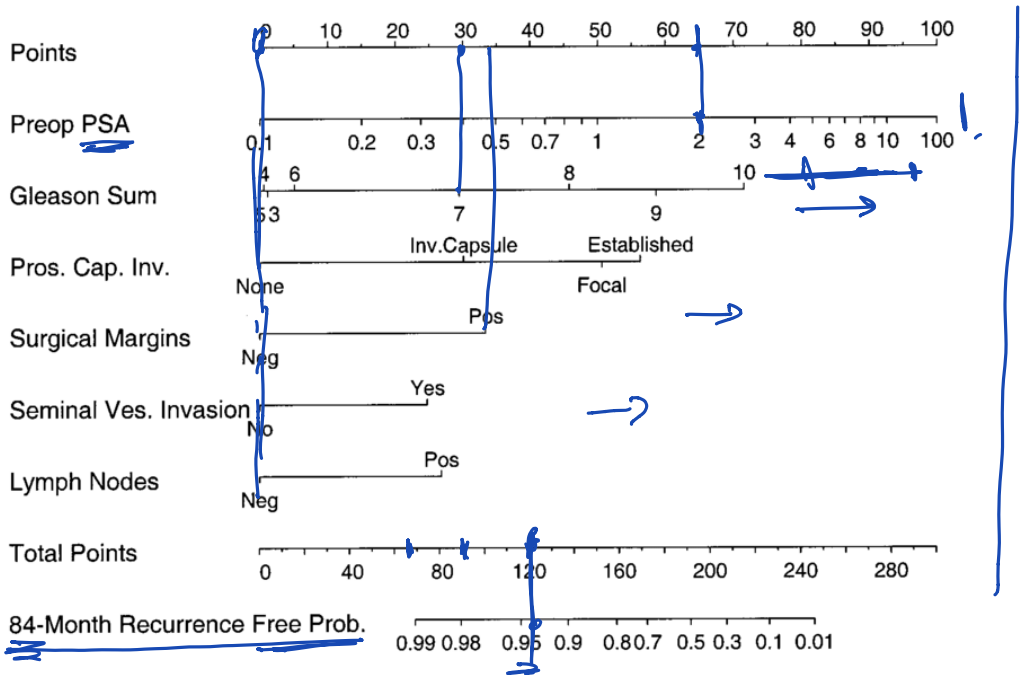
Logistic regression $\hat{y} = p = \frac{1}{1 + e^{-x\omega^T}}$

$$\log \frac{p}{1-p} = \omega_0 + \omega_1 x_1 + \dots + \omega_n x_n$$

$$\log \frac{p}{1-p} - \omega_0 = \omega_1 x_1 + \dots + \omega_n x_n = S$$

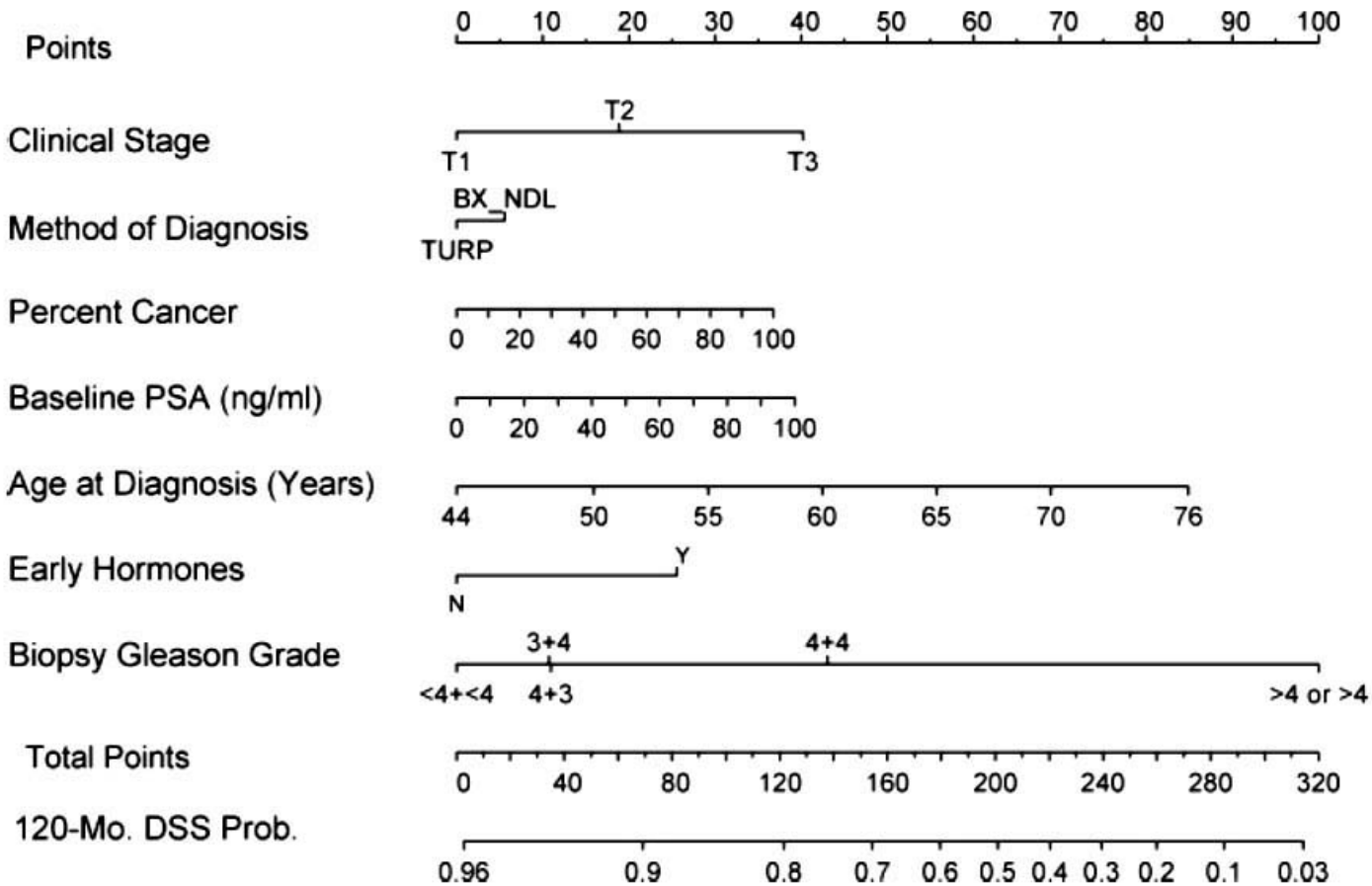
looking for p given S — $p(S)$ ...

Kattan & Scardino 1997



Instructions for Physician: Locate the patient's PSA on the **PSA** axis. Draw a line straight upwards to the **Points** axis to determine how many points towards recurrence the patient receives for his PSA. Repeat this process for the other axes, each time drawing straight upward to the **Points** axis. Sum the points achieved for each predictor and locate this sum on the **Total Points** axis. Draw a line straight down to find the patient's probability of remaining recurrence free for 84 months assuming he does not die of another cause first.

Instruction to Patient: "Mr. X, if we had 100 men exactly like you, we would expect between <predicted percentage from nomogram – 10%> and <predicted percentage + 10%> to remain free of their disease at 7 years following radical prostatectomy, and recurrence after 7 years is very rare."



American Cancer Society

Platinum Priority – Kidney Cancer
Editorial by Christian Bolenz and Yair Lotan on pp. 296–298 of this issue
EUROPEAN UROLOGY 55 (2009) 287–295

A Preoperative Prognostic Model for Patients Treated with Nephrectomy for Renal Cell Carcinoma

Pierre I. Karakiewicz [a,*], Nazareno Suardi [a,b], Umberto Capitanio [a,b], Claudio Jeldres [a],

- Accuracy: 84-88%

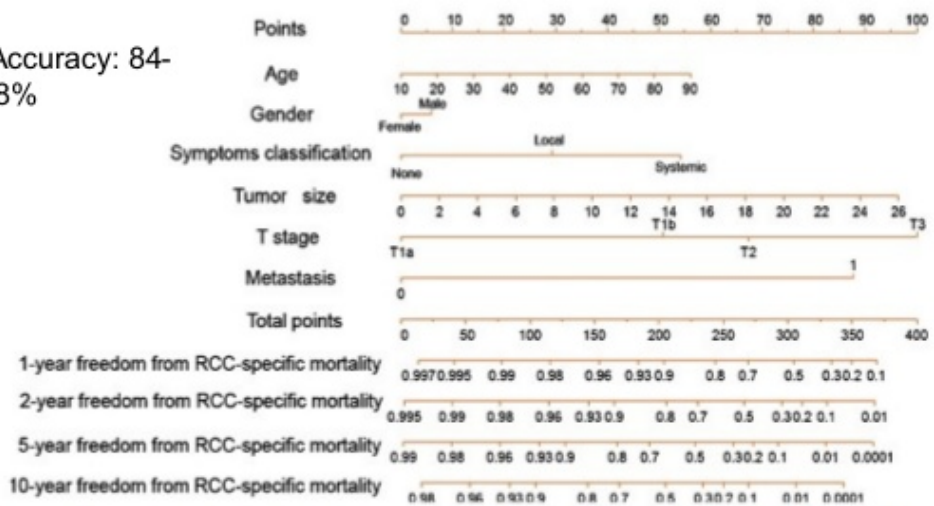


Fig. 2 – Preoperative nomogram predicting renal cell carcinoma (RCC)–specific survival at 1 yr, 2 yr, 5 yr, and 10 yr. Abbreviations: S classification, symptoms classification; M, metastases (0 = absent; 1 = present).

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 38 (2005) 389–394
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Nomographic representation of logistic regression models: A case study using patient self-assessment data

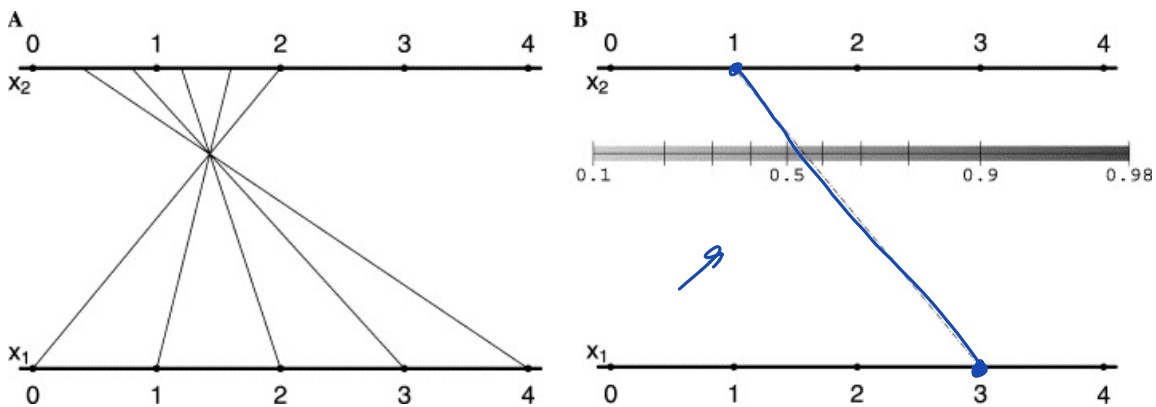
Stephan Dreiseitl [a,*], Alexandra Harbauer [b], Michael Binder [b], Harald Kittler [b]

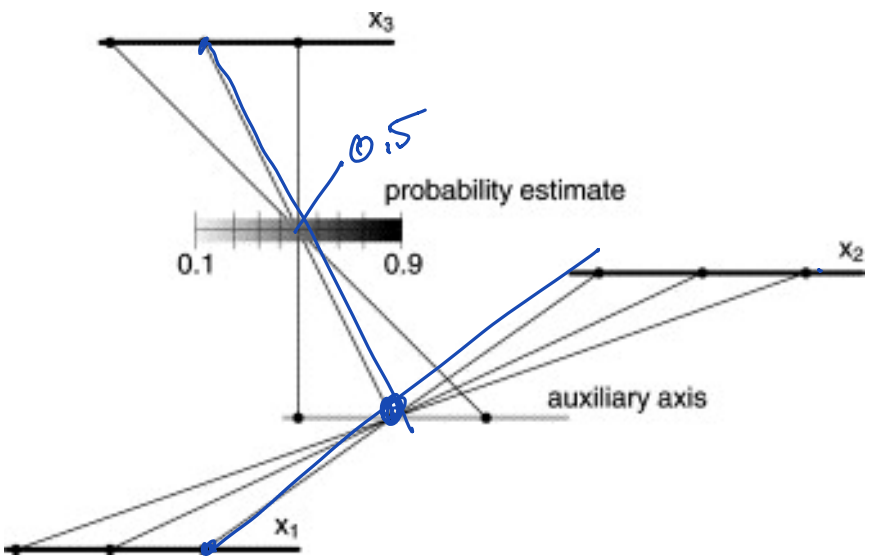
[a] Department of Software Engineering, University of Applied Sciences Upper Austria at Hagenberg, Austria
[b] Department of Dermatology, Medical University of Vienna, Austria

Received 24 December 2004
Available online 17 March 2005

log. → graphical device

$$p = \frac{1}{1+e^{-(-2+0.4x_1+x_2)}}$$




x3
0.5
probability estimate
0.1
0.9
x2
auxiliary axis
x1

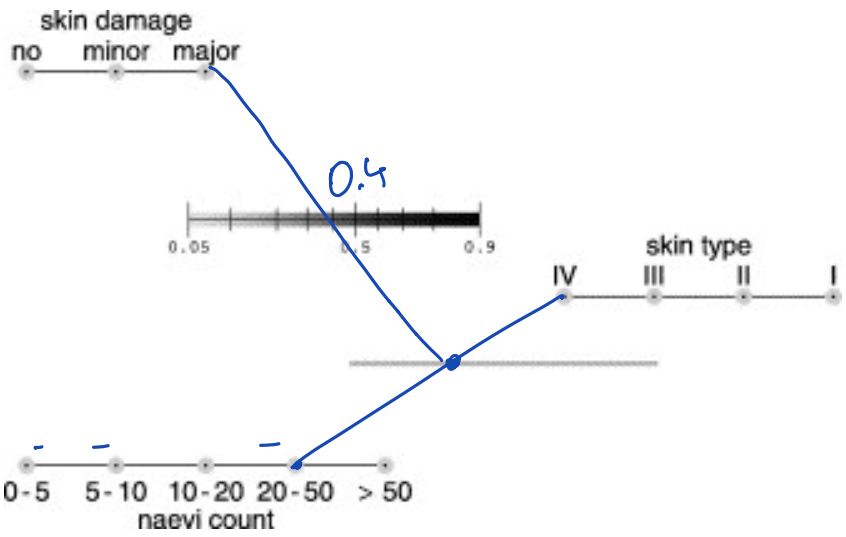

skin damage
no
minor
major
0.4
0.05
0.5
0.9
skin type
IV
III
II
I
0-5
5-10
10-20
20-50
> 50
naevi count

Fig. 3 | Image from the authors of ref. [48], indicating that parts of the test image on the left are similar to prototypical parts of training examples. The test image to be classified is on the left, the most similar prototypes are in the middle column, and the heatmaps that show which part of the test image is similar to the prototype are on the right. We included copies of the test image on the right so that it is easier to see to what part of the bird the heatmaps are referring. The similarities of the prototypes to the test image are what determine the predicted class label of the image. Here, the image is predicted to be a clay-coloured sparrow. The top prototype seems to be comparing the bird's head to a prototypical head of a clay-coloured sparrow, the second prototype considers the throat of the bird, the third looks at feathers, and the last seems to consider the abdomen and leg. Credit: Image constructed by Alina Barnett, Duke University

Dustin concludes

- block box models are not necessary needed for accurate predictions
- focus of feature engineering
- more effort
  Safety & trust ML models