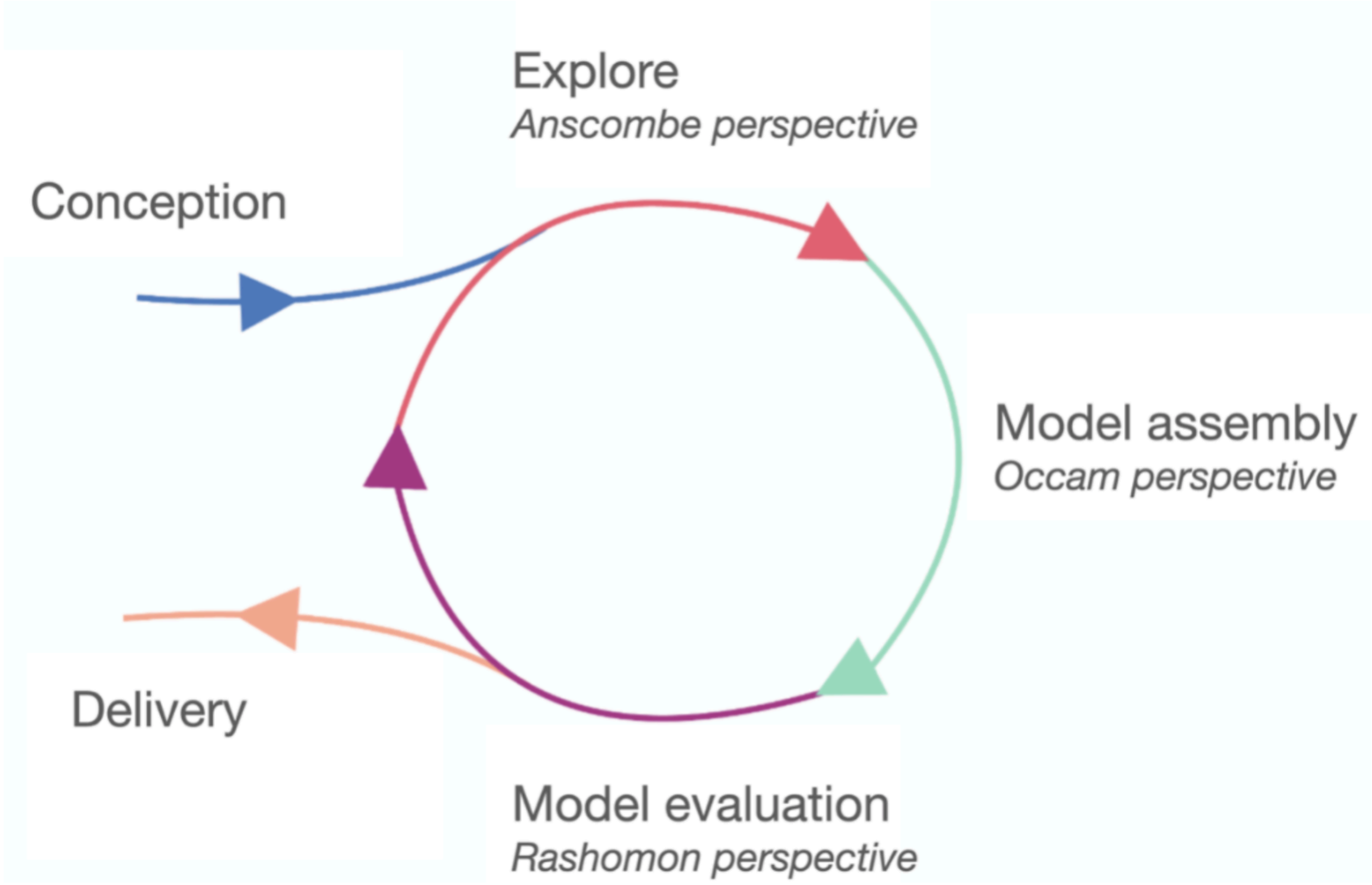
Introduction to Responsible Machine Learning

with mlr3 and DALEX

Przemyslaw Biecek

17-18.04.2021



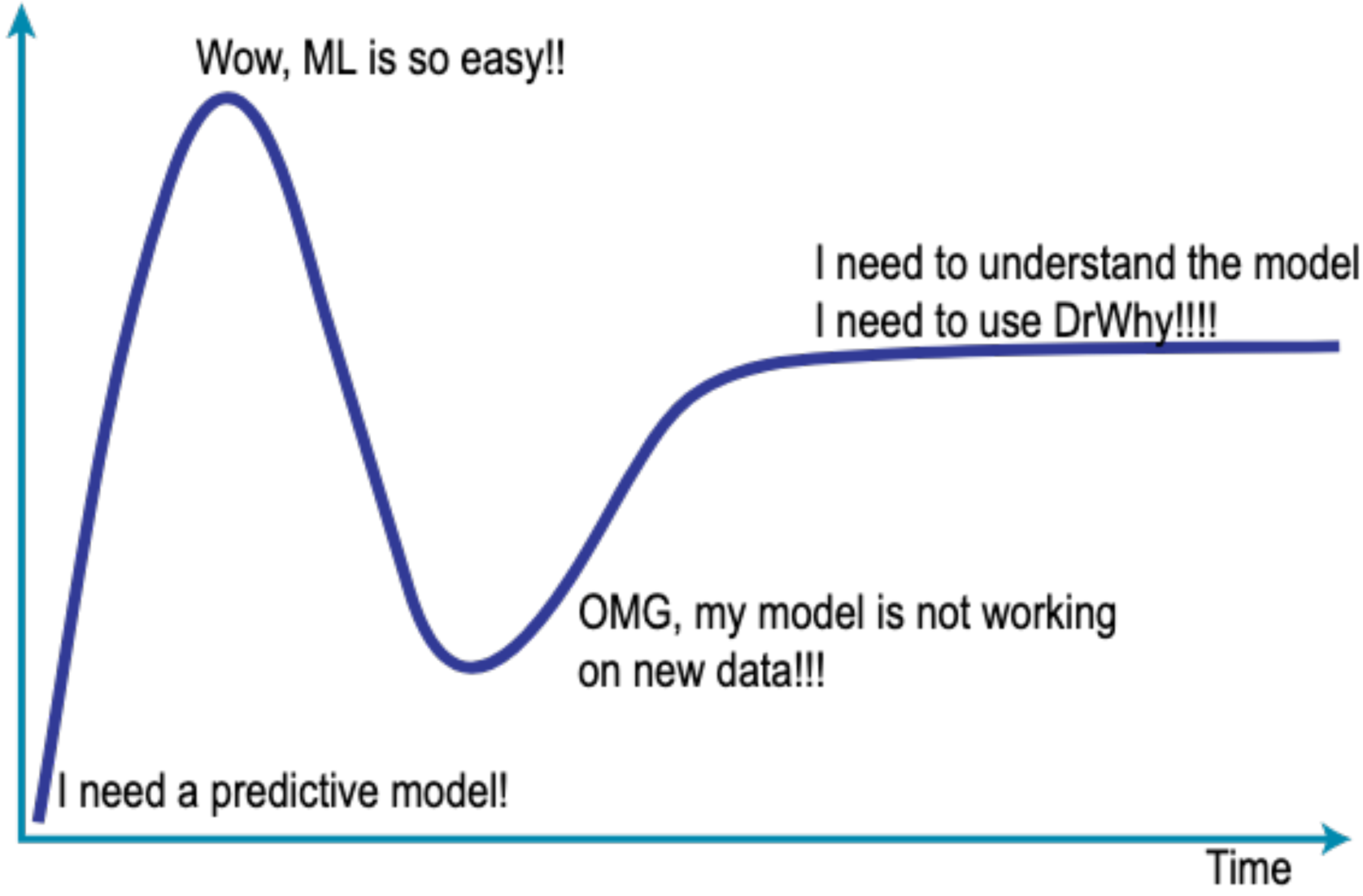
Day 1: Predictive modeling
Each block: 1.5h lecture + hands on exercises

09:00 - 10:30 Introduction to predictive modeling + EDA
10:30 - 10:45 Break
10:45 - 12:15 Hello model - first predictive model + How to measure performance
12:15 - 12:30 Break
12:30 - 14:00 Basics of random forest and boosting models
14:00 - 14:15 Break
14:15 - 15:45 Hyperparameter optimization + Wrap-up

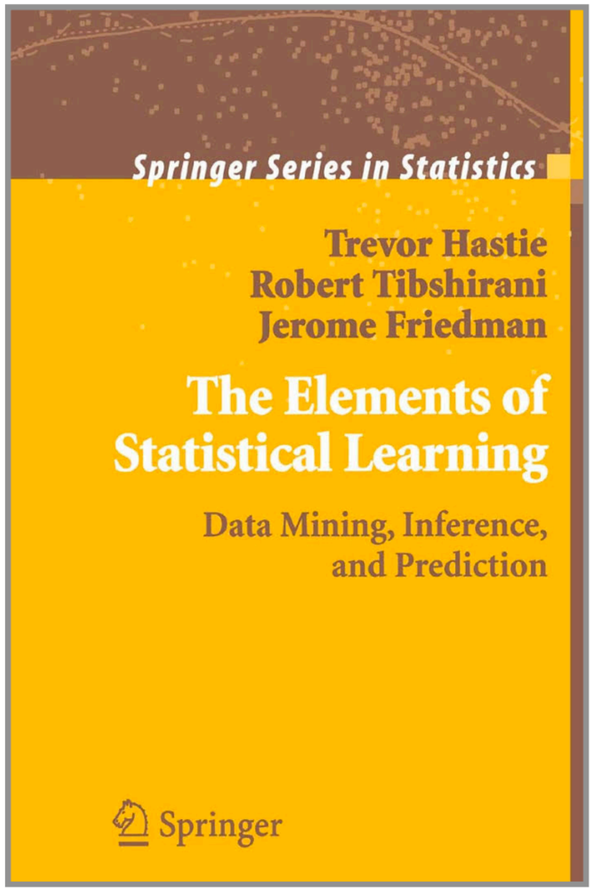
Day 2: Model exploration
Each block: 1.5h lecture + hands on exercises

09:00 - 10:30 Model level analysis - variable importance
10:30 - 10:45 Break
10:45 - 12:15 Model level analysis - variable profile
12:15 - 12:30 Break
12:30 - 14:00 Instance level analysis - variable attributions
14:00 - 14:15 Break
14:15 - 15:45 Instance level analysis - variable profile + Wrap-up

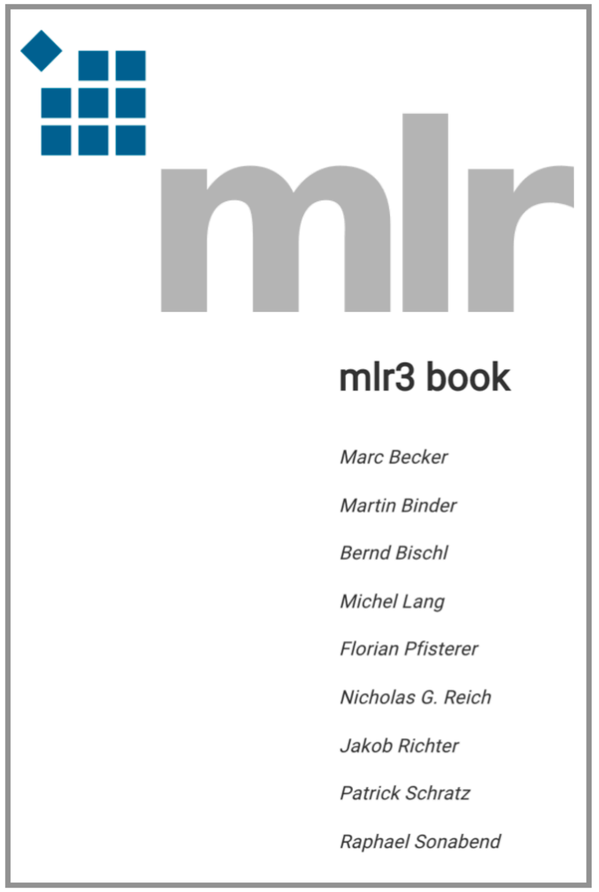
Hype Cycle for Predictive Models



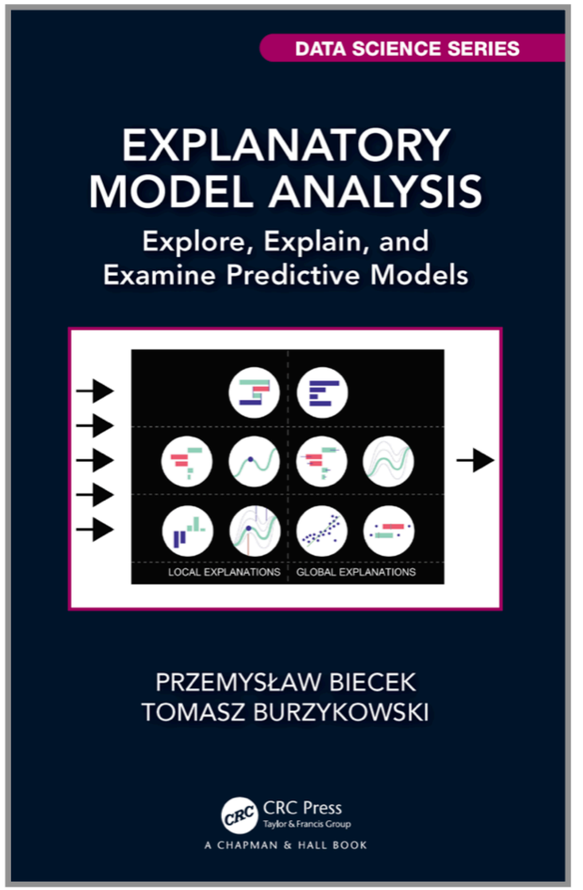
http://drwhy.ai



The Elements of Statistical Learning
https://www.statlearning.com/



The mlr3 book
https://mlr3book.mlr-org.com/



Explanatory Model Analysis
http://ema.drwhy.ai/

hbaniecki unify site & readme, closes #349 ✓ Latest commit ec055bf on 3 Dec 2020 History

5 contributors

107 lines (71 sloc) | 9 KB | Raw | Blame

moDel Agnostic Language for Exploration and eXplanation

Overview

Unverified black box model is the path to the failure. Opaqueness leads to distrust. Distrust leads to ignoration. Ignoration leads to rejection.

The `DALEX` package xrays any model and helps to explore and explain its behaviour, helps to understand how complex models are working. The main function `explain()` creates a wrapper around a predictive model. Wrapped models may then be explored and compared with a collection of local and global explainers. Recent developents from the area of Interpretable Machine Learning/eXplainable Artificial Intelligence.

The philosophy behind `DALEX` explanations is described in the Explanatory Model Analysis e-book. The `DALEX` package is a part of DrWhy.AI universe.

If you work with `scikit-learn`, `keras`, `H2O`, `tidymodels`, `xgboost`, `mlr` or `mlr3`, you may be interested in the DALEXtra package. It is an extension pack for `DALEX` with easy to use connectors to models created in these libraries.

https://github.com/ModelOriented/DALEX

Przemysław Biecek
(53.2% of academia, 21.3% of corpo, 10.6% of startups, 14.9% of NGOs)

Academy
6 years - Warsaw University of Technology
13 years - University of Warsaw
1 year - University of Wroclaw
1 year - Polish Academy of Sciences
4 years - Wroclaw University of Technology

Corpo
3 years - Samsung - Principal Data Scientist
1 year - OECD France - Research Fellow
2 years - iQor - Head of Data Vis
2 years - IBM Senior Data Scientist
2 years - Netezza Senior Data Scientist

Startups
1 year - Data Donuts
2 year - Data Hero
2 years - Positive Advisory

NGO
7 years - SmarterPoland / BetaBit

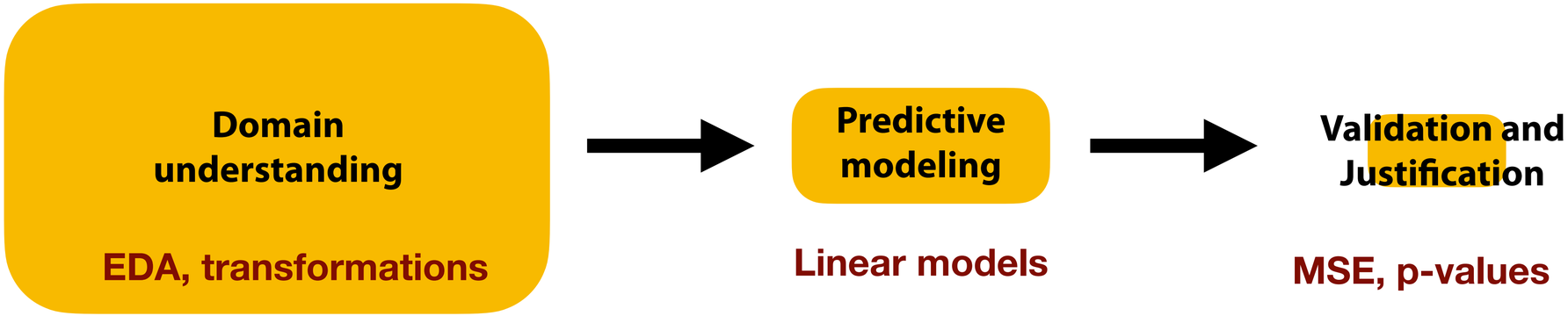
Day 1

Predictive modeling

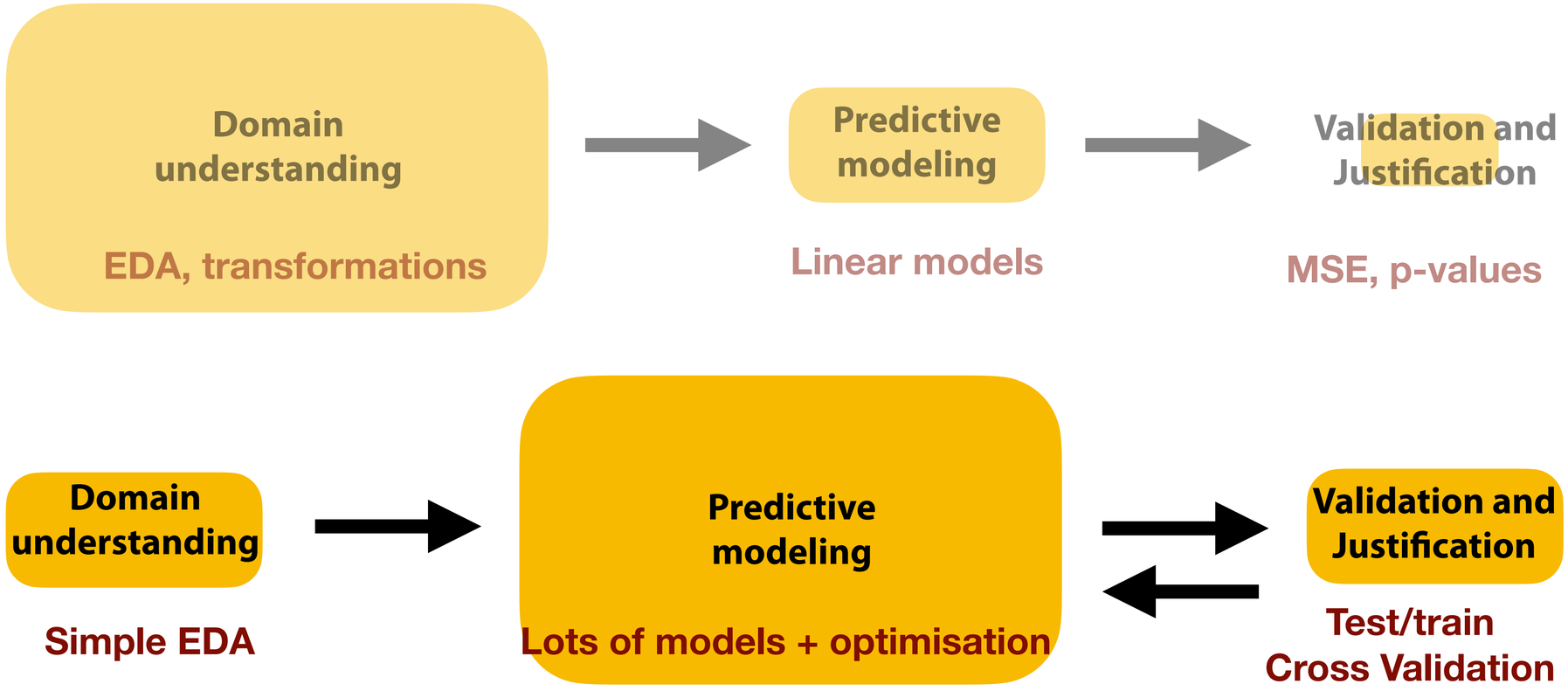
Part 1

Introduction to predictive modeling + EDA

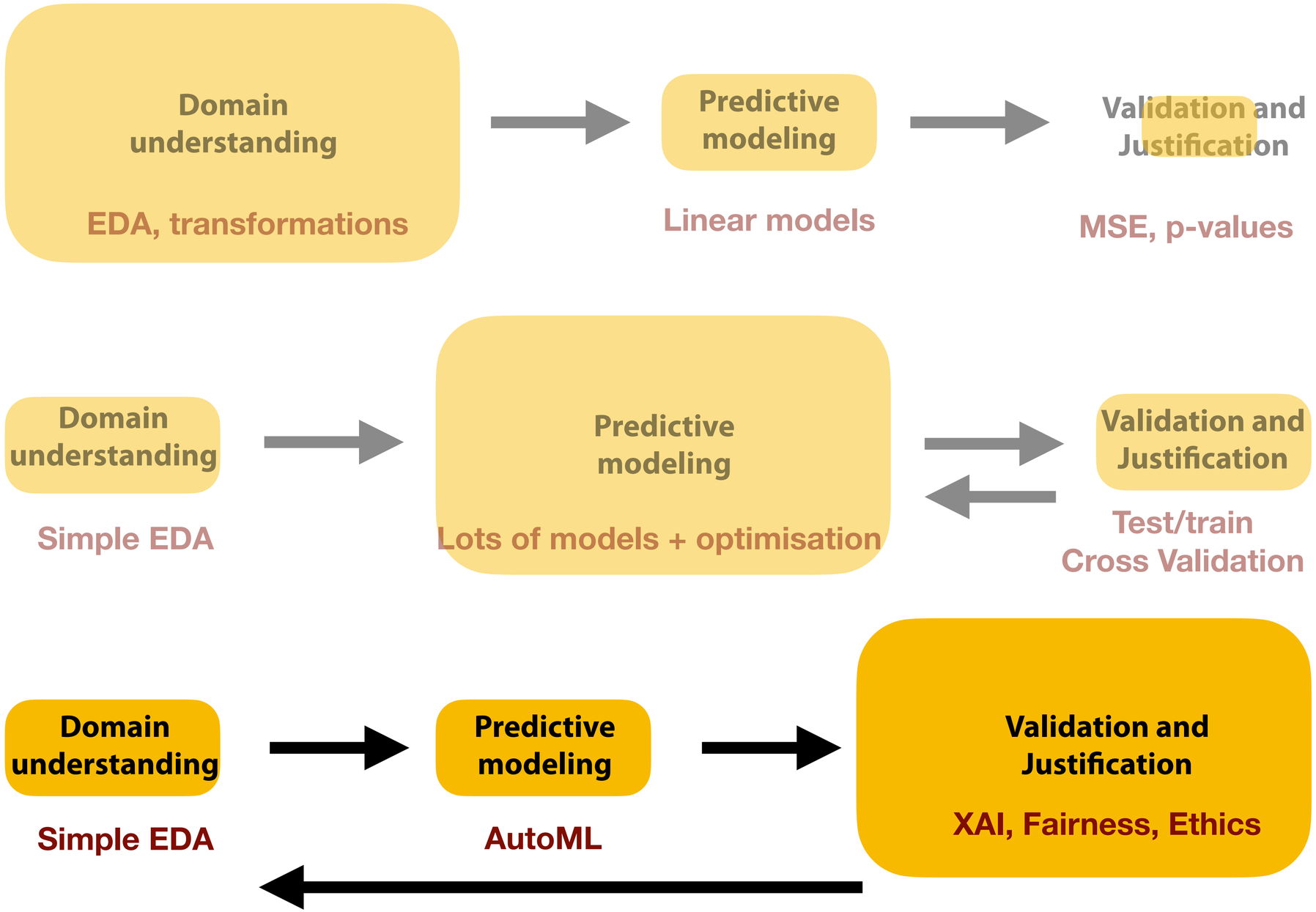
Shift in our focus: Statistics



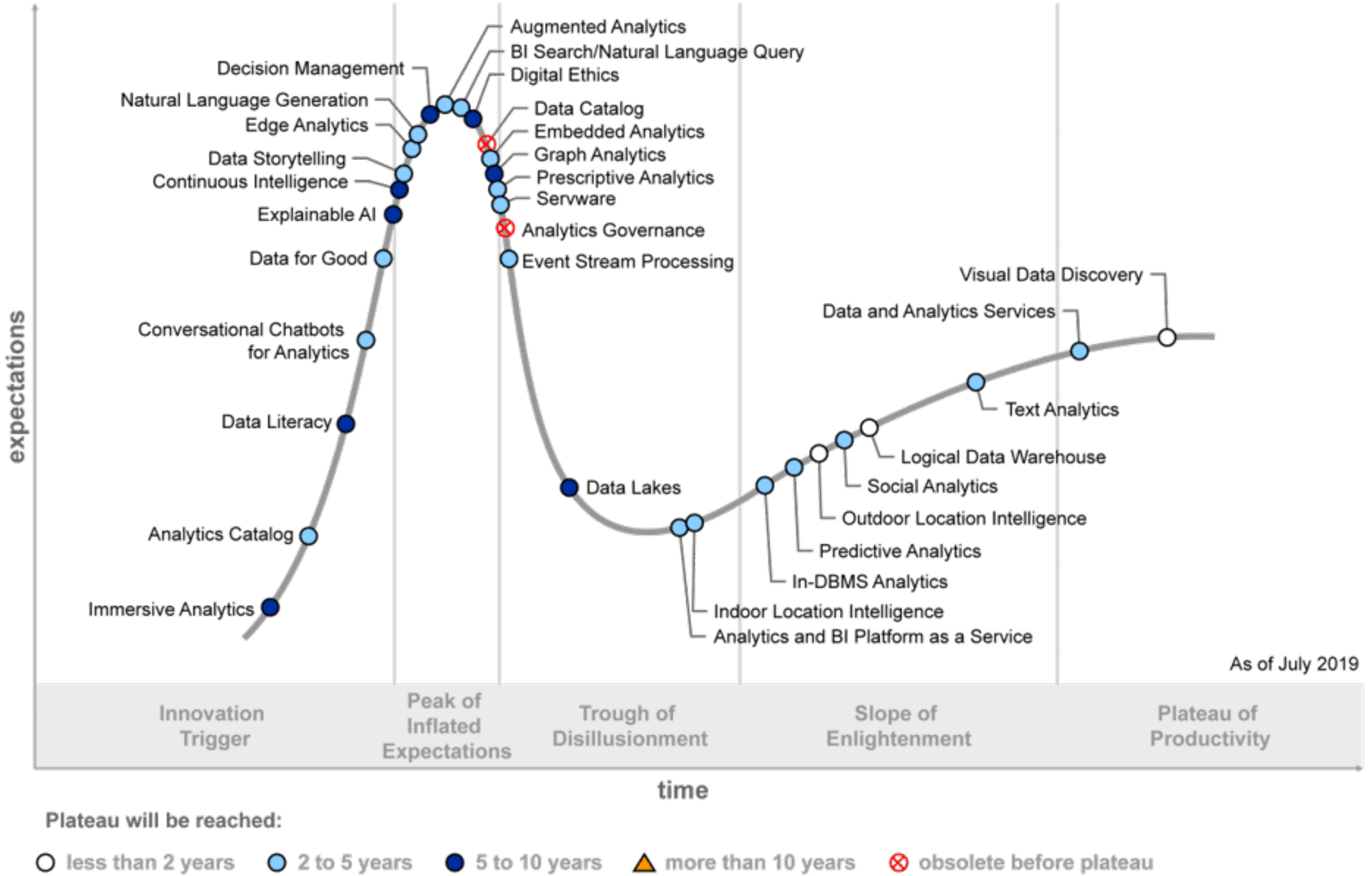
Shift in our focus: Machine Learning



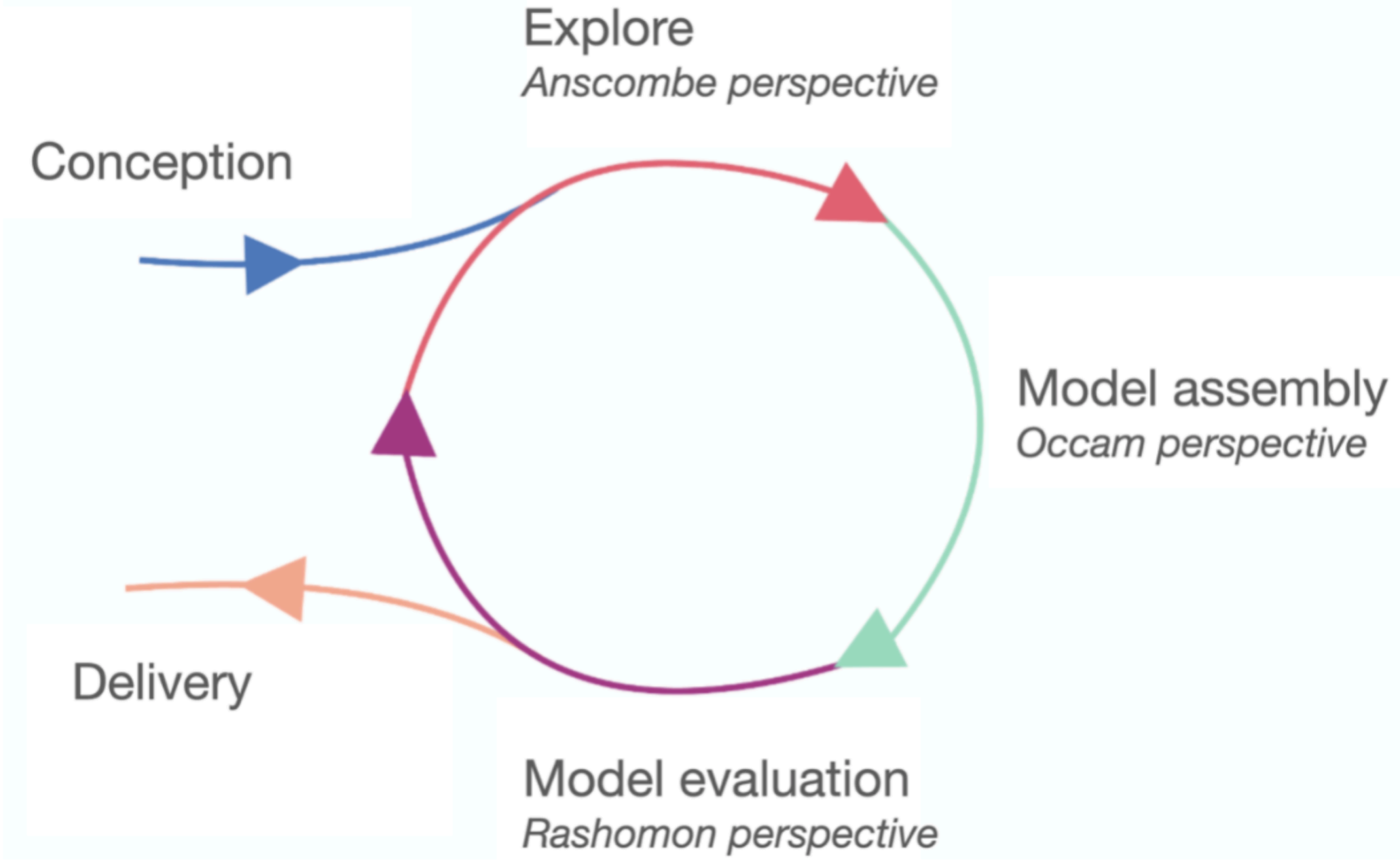
Shift in our focus: Human Oriented ML?



Hype Cycle for Analytics and Business Intelligence, 2019



Source: Gartner


Explore
Anscombe perspective
Conception
Model assembly
Occam perspective
Delivery
Model evaluation
Rashomon perspective


Problem formulation
Crisp Modeling
Fine tuning
Mantaining and decommissioning
Data preparation
Design of Experiment
Data acquisition
Data cleaning
Data understading
Data exploration
Feature selection
Feature engineering
Model assembly
Model selection
Parameters tuning
Model debbugging
Model audit
Residual analysis
Model benchmarking
Model diagnostic
Model delivery
Model deployment
Documentation
Communication
Iterations:
P1
P2
C1
C2
F1
F2
F3
M1
M2
M3
time

Day 1

Predictive modeling

Part 2

Hello model - first predictive model
How to measure performance

$$MSE(f) = \frac{1}{n}\sum_{i}^{n}(f(x_i) - y_i)^2$$

$$RMSE(f) = \sqrt{MSE(f, X, y)}$$

$$ACC(f) = (TP + TN)/n$$

$$Prec(f) = TP/(TP + FP)$$

$$Recall(f) = TP/(TP + FN)$$

$$F1(f) = 2\frac{Prec(f) * Recall(f)}{Prec(f) + Recall(f)}$$

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Pregnancy: Sensitivity and Specificity

Introduction: None written.

[Edit Diagnosis] [Merge dx] [Add prevalence]

Tags: None. *Tag this Diagnosis.*

The sensitivity and specificity of findings for Pregnancy are listed below. See the left navigation bar to change the display.

Sensitive and Specific Findings

Finding	*Sensitivity*	*Specificity*	*Comments, Study*
Serum beta-hCG	97%	100%	as compared to urine beta hCG, for all comers to an ED for whom pregnancy tests were ordered. In this study, urine beta hCG had sens = 100%, spec = 99.2% **Study:** Am J Emerg Med. 1993 Jul;11(4):434-6. PMID: 8216535

Specific Findings

Finding	*Sensitivity*	*Specificity*	*Comments, Study*
Chadwick Sign	51%	98%	Only 1 study from Dr Chadwick. Sign is "characteristic" deep violet-blue color of anterior vaginal wall. Rarely seen before 7 weeks. **Study:** JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281
Palpable Fundus	9%	97%	**Study:** JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281
Vaginal Exam Signs	18%	94%	not explicitly defined **Study:** JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281
Uterine Artery Pulsation	76%	93%	only 1 study **Study:** JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281
Morning Sickness	39%	85%	**Study:** JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281

Sensitive Findings

Finding	*Sensitivity*	*Specificity*	*Comments, Study*
Urine beta-hCG *Duplicate*	*Sensitivity = 97%*		7 days after missed period. Most urine b-hCG tests are sensitive to 20 mIU/ml. If sensitivity is at 0.5 mIU/ml then should be sensitive by 7 days after fertilization. **Study:** JAMA 2001 Oct 10;286(14):1759-61. PMID 11594902
Urine beta-hCG *Duplicate*	*Sensitivity = 90%*		on 1st day of missed period. Most urine b-hCG tests are sensitive to 20 mIU/ml. If sensitivity is at 0.5 mIU/ml then should be sensitive by 7 days after fertilization. **Study:** JAMA 2001 Oct 10;286(14):1759-61. PMID 11594902
Did Not Use Birth Control	88%	42%	**Study:** JAMA. 1997 Aug 20;278(7):586-91. PMID: 9268281

Morning sickness	Pregnant	Not Pregnant	
Has sickness	39	150	Positive predictive value: 20.63%
Has not	61	850	Negative predictive value: 93.30%
	Sensitivity: 39.00%	Specificity: 85.00%	Accuracy: 80.82%

miss rate or false negative rate (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{P} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

false discovery rate (FDR)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

false omission rate (FOR)

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

---

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Informedness or Bookmaker Informedness (BM)

$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1$$

Markedness (MK)

$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1$$

condition positive (P)

the number of real positive cases in the data

condition negative (N)

the number of real negative cases in the data

---

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

---

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$

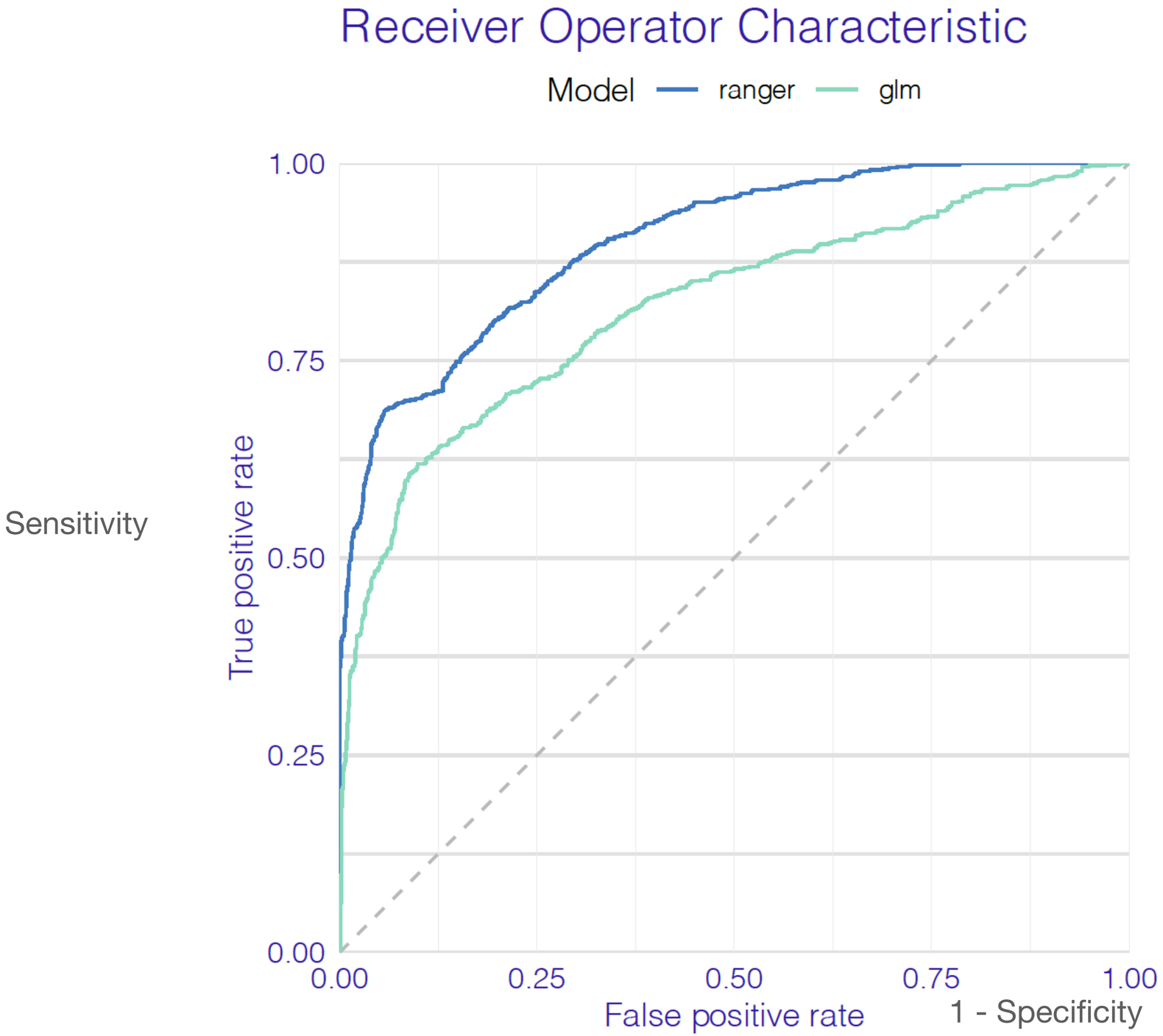
negative predictive value (NPV)

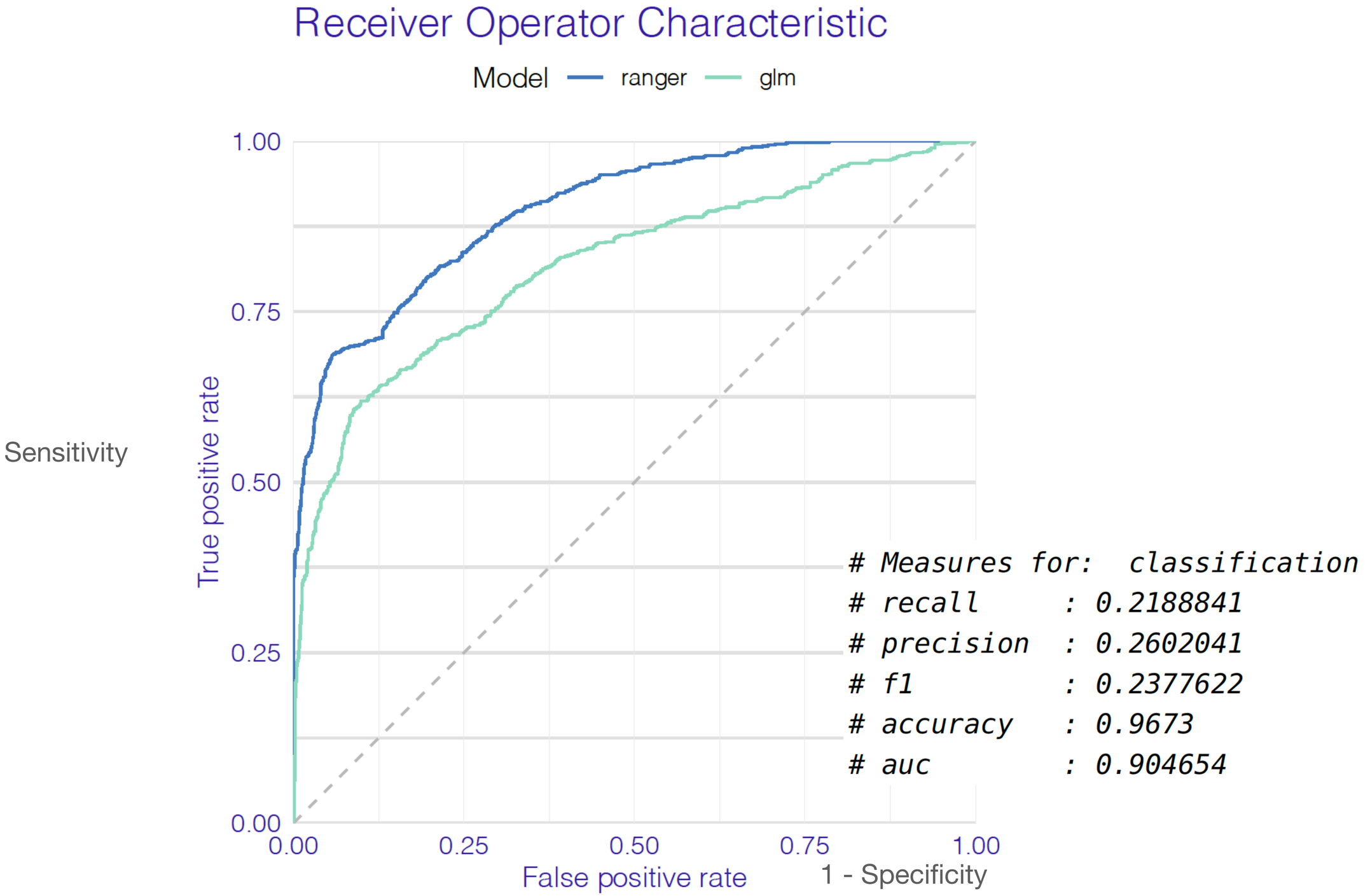
$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = 1 - \mathrm{FOR}$$

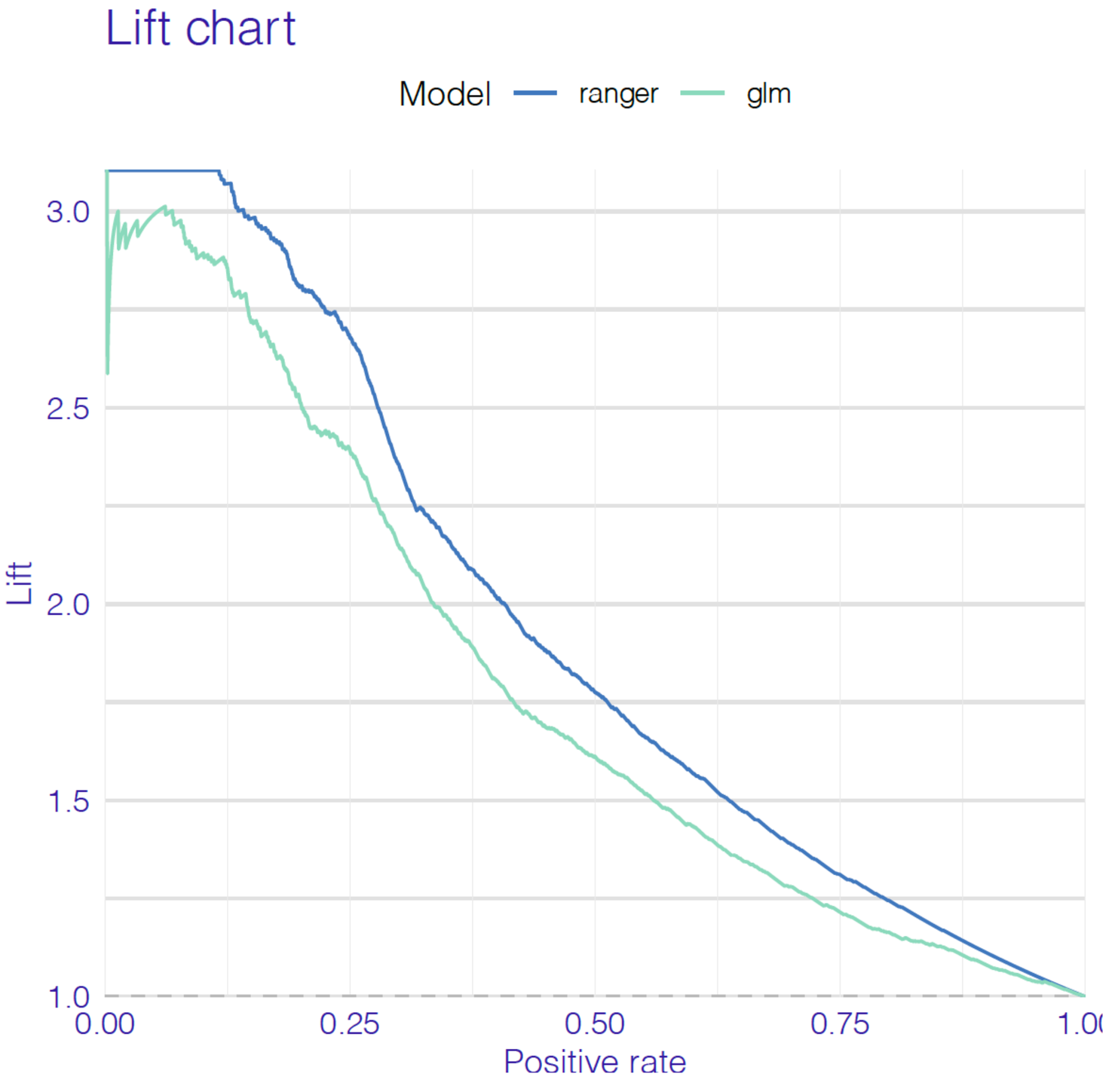
Country of development	US/China
Type of Serological Test	RDT
Authors/Company	Cellex Inc.
Description	RDT, lateral flow assay, which detects IgM and IgG to the nucelocapside protein of SARS-CoV-2. The sensitivity is 93.8% and specificity is 95.6%, when tested at 2 Chinese hospitals in a total of 128 COVID19 positive patients, and 250 COVID19 negative patients (as detected by RT-qPCR).
Sensitivity	93.8%
Specificity	95.6%
Phase of development	Approved by FDA for EUA on diagnostics, has CE approval
Proposed release	available for purchase by research labs/healthcare providers (product number 5513)
Date	April 1, 2020

	Infected	Not Infected	
Positive test (+)	18760	44000	Positive predictive value: 29.89%
Negative test (-)	1240	956000	Negative predictive value: 99.87%
	Sensitivity: 93.8%	Specificity: 95.6%	

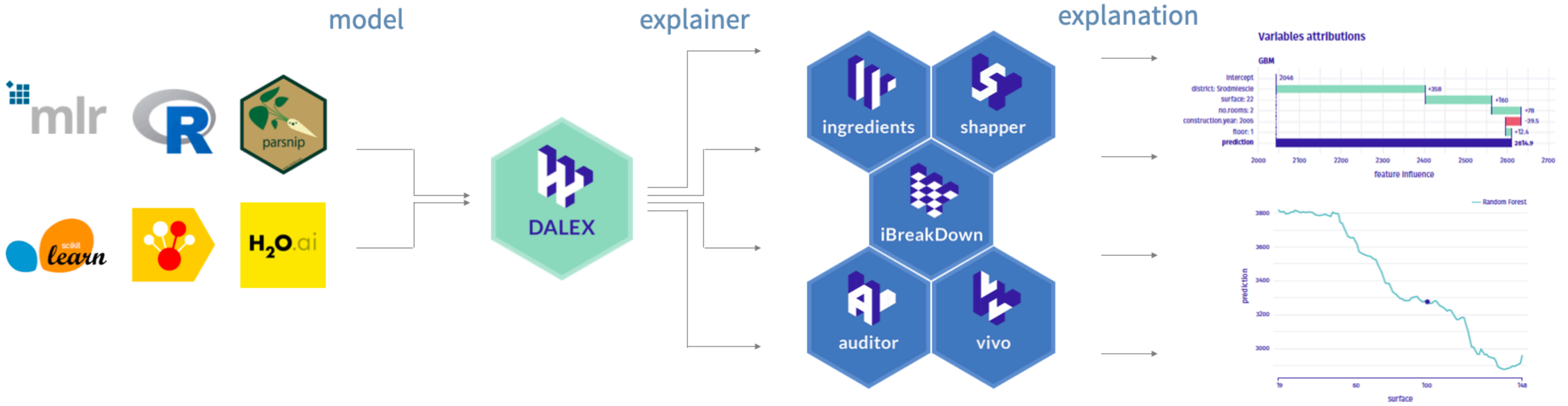
https://www.centerforhealthsecurity.org/resources/COVID-19/serology/Serology-based-tests-for-COVID-19.html

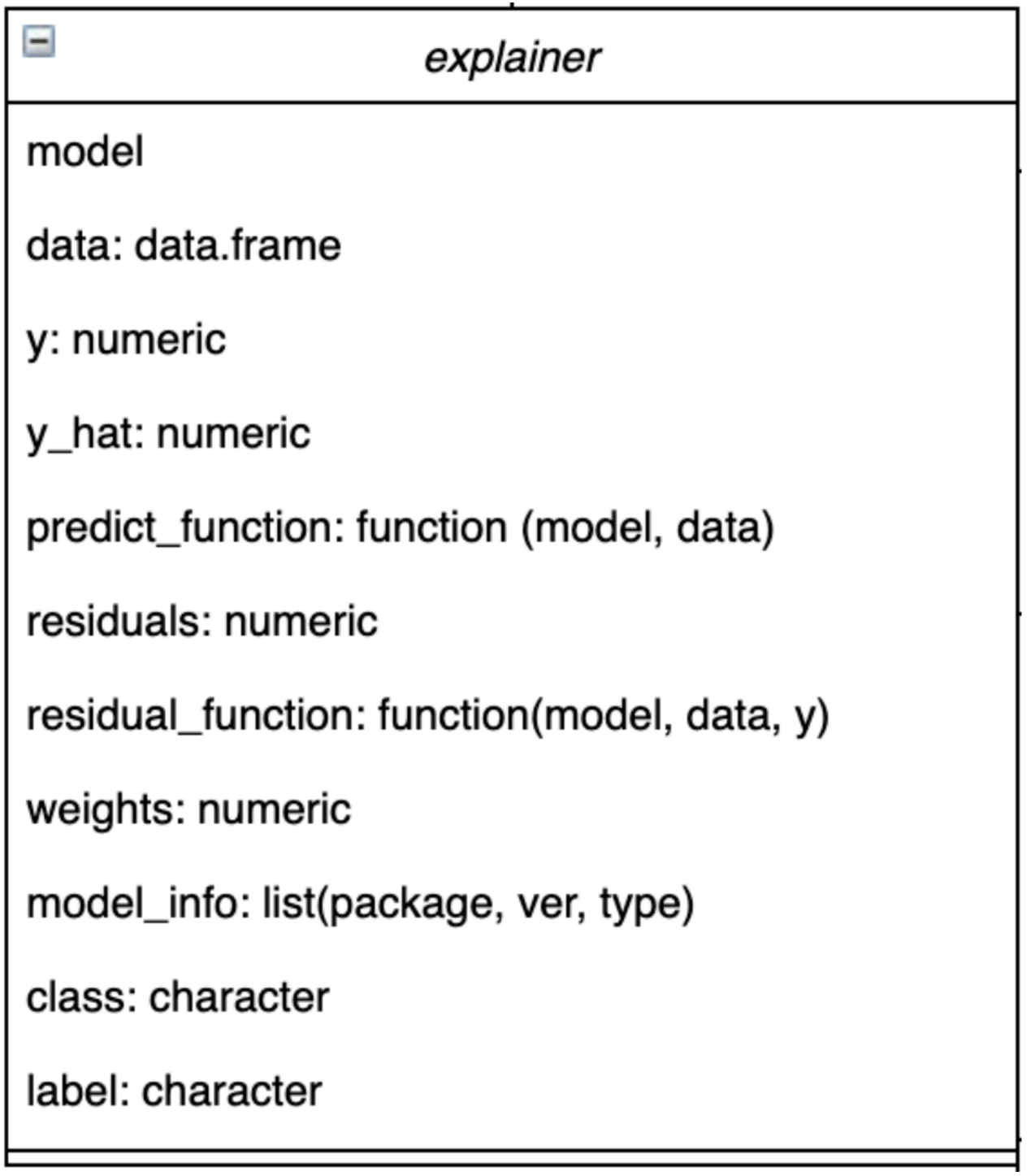

Receiver Operator Characteristic
Model
ranger
glm
1.00
0.75
0.50
0.25
0.00
True positive rate
Sensitivity
0.00
0.25
0.50
0.75
1.00
False positive rate
1 - Specificity


Receiver Operator Characteristic
Model
ranger
glm
1.00
0.75
0.50
0.25
0.00
True positive rate
Sensitivity
0.00
0.25
0.50
0.75
1.00
False positive rate
1 - Specificity
Measures for: classification
recall : 0.2188841
precision : 0.2602041
f1 : 0.2377622
accuracy : 0.9673
auc : 0.904654


Lift chart
Model
ranger
glm
3.0
2.5
2.0
1.5
1.0
Lift
0.00
0.25
0.50
0.75
1.0
Positive rate

```
ranger(y ~ ., data = df)  %>%  explain()  %>%  model_parts()  %>%  plot()
```



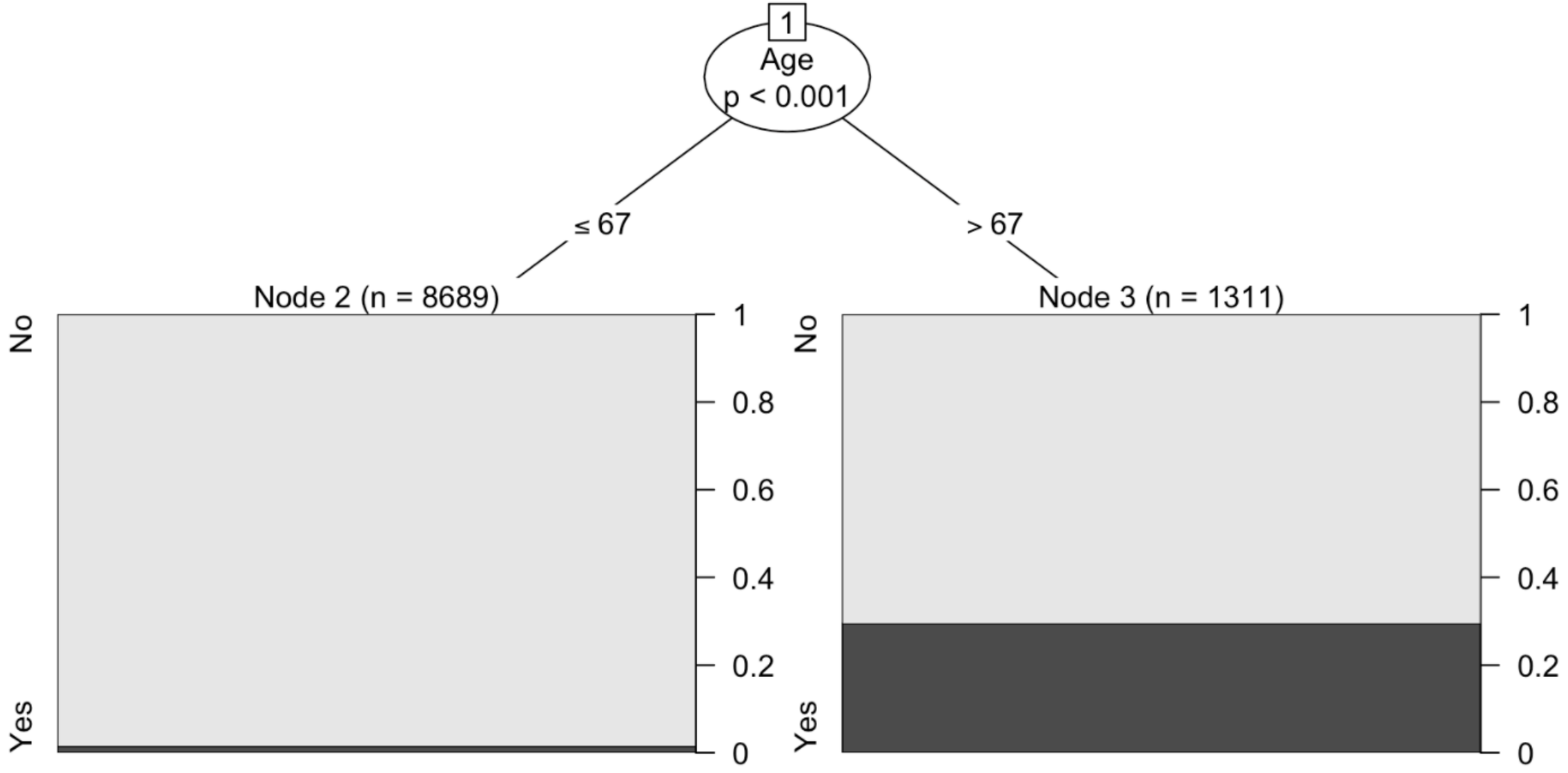

explainer
model
data: data.frame
y: numeric
y_hat: numeric
predict_function: function (model, data)
residuals: numeric
residual_function: function(model, data, y)
weights: numeric
model_info: list(package, ver, type)
class: character
label: character

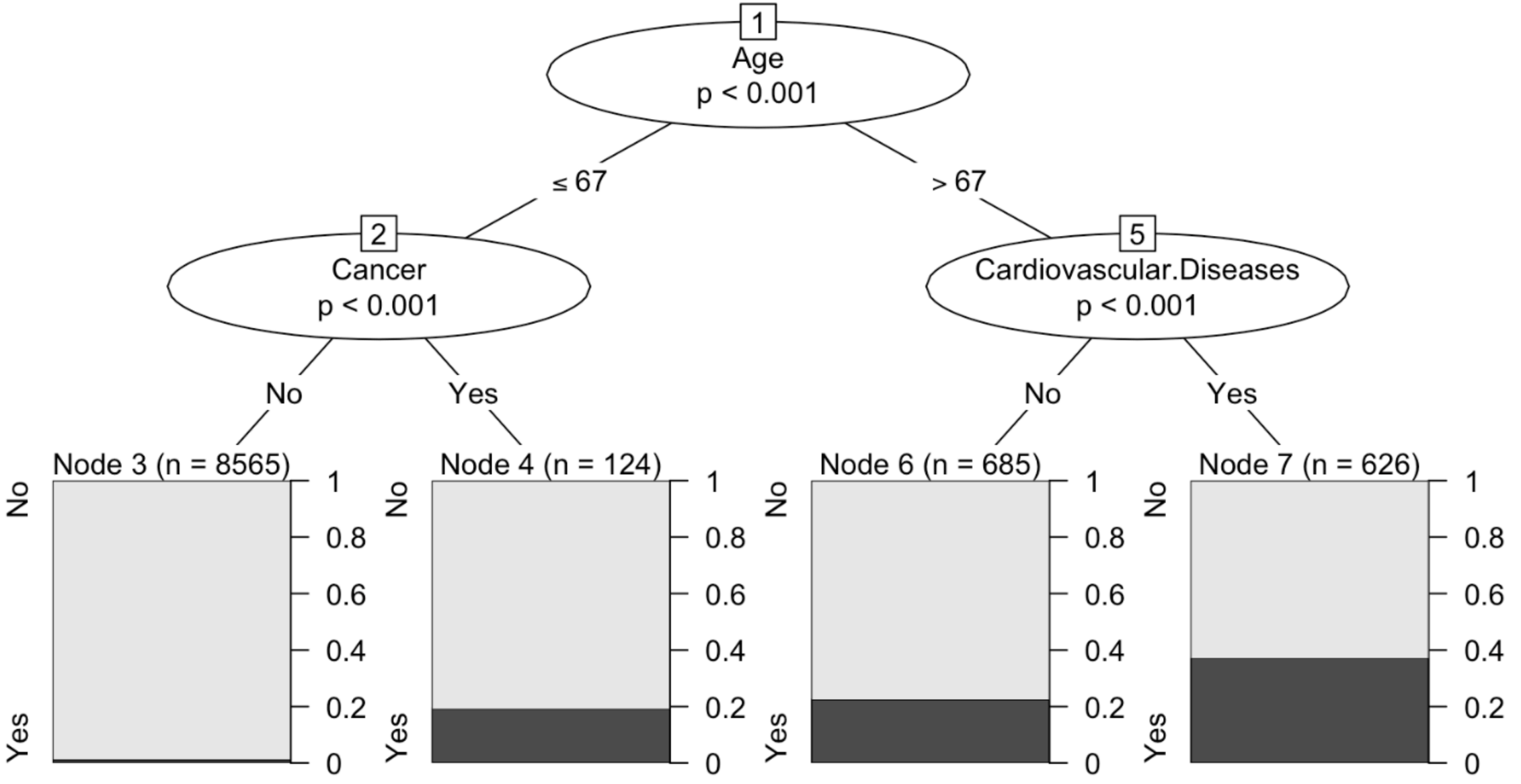
Day 1

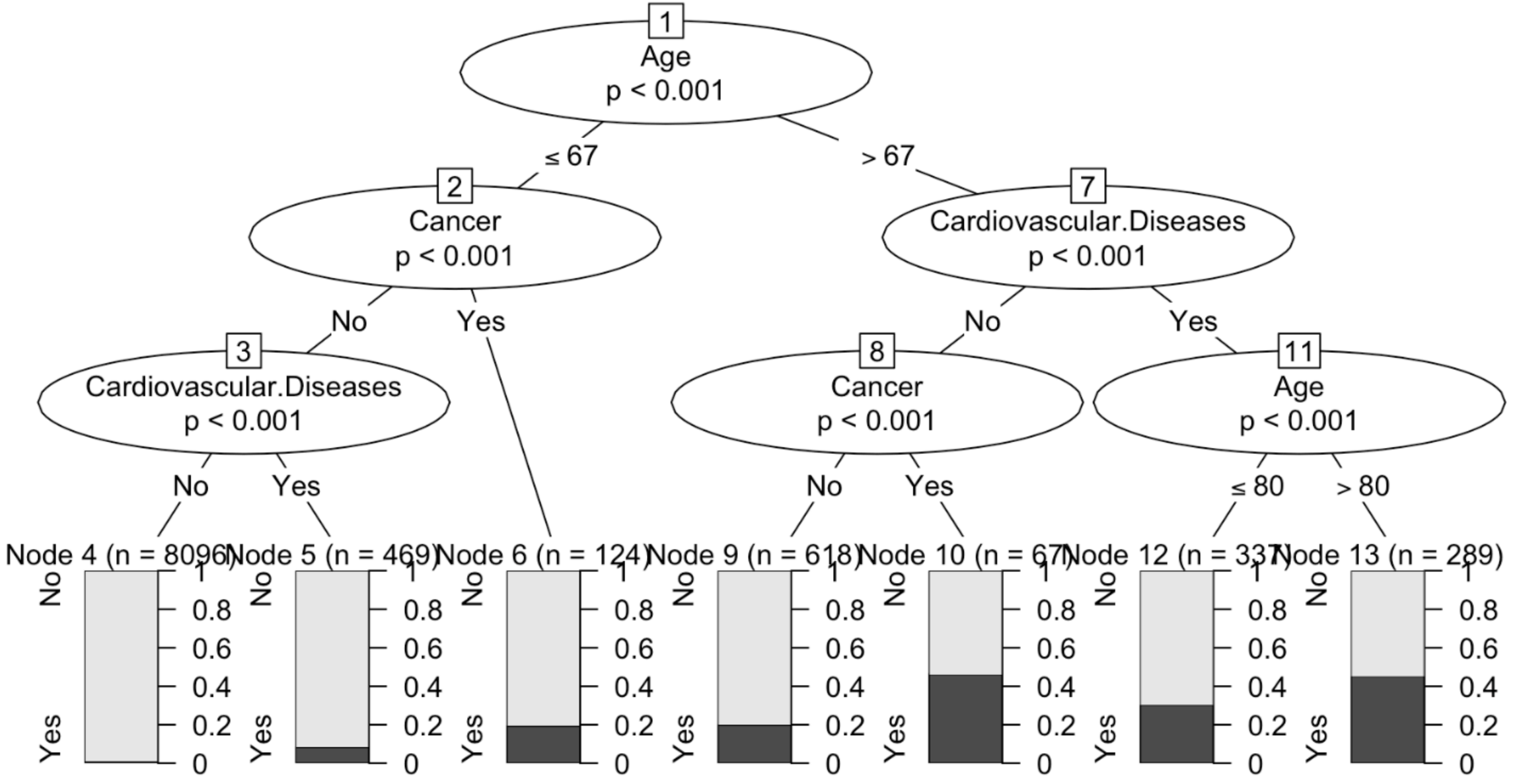
Predictive modeling

Part 3

Basics of random forest and boosting models


1
Age
p < 0.001
≤ 67
> 67
Node 2 (n = 8689)
Node 3 (n = 1311)
No
Yes
1
0.8
0.6
0.4
0.2
0


1
Age
p < 0.001
≤ 67
> 67
2
Cancer
p < 0.001
5
Cardiovascular.Diseases
p < 0.001
No
Yes
No
Yes
Node 3 (n = 8565)
Node 4 (n = 124)
Node 6 (n = 685)
Node 7 (n = 626)
No
Yes
1
0.8
0.6
0.4
0.2
0

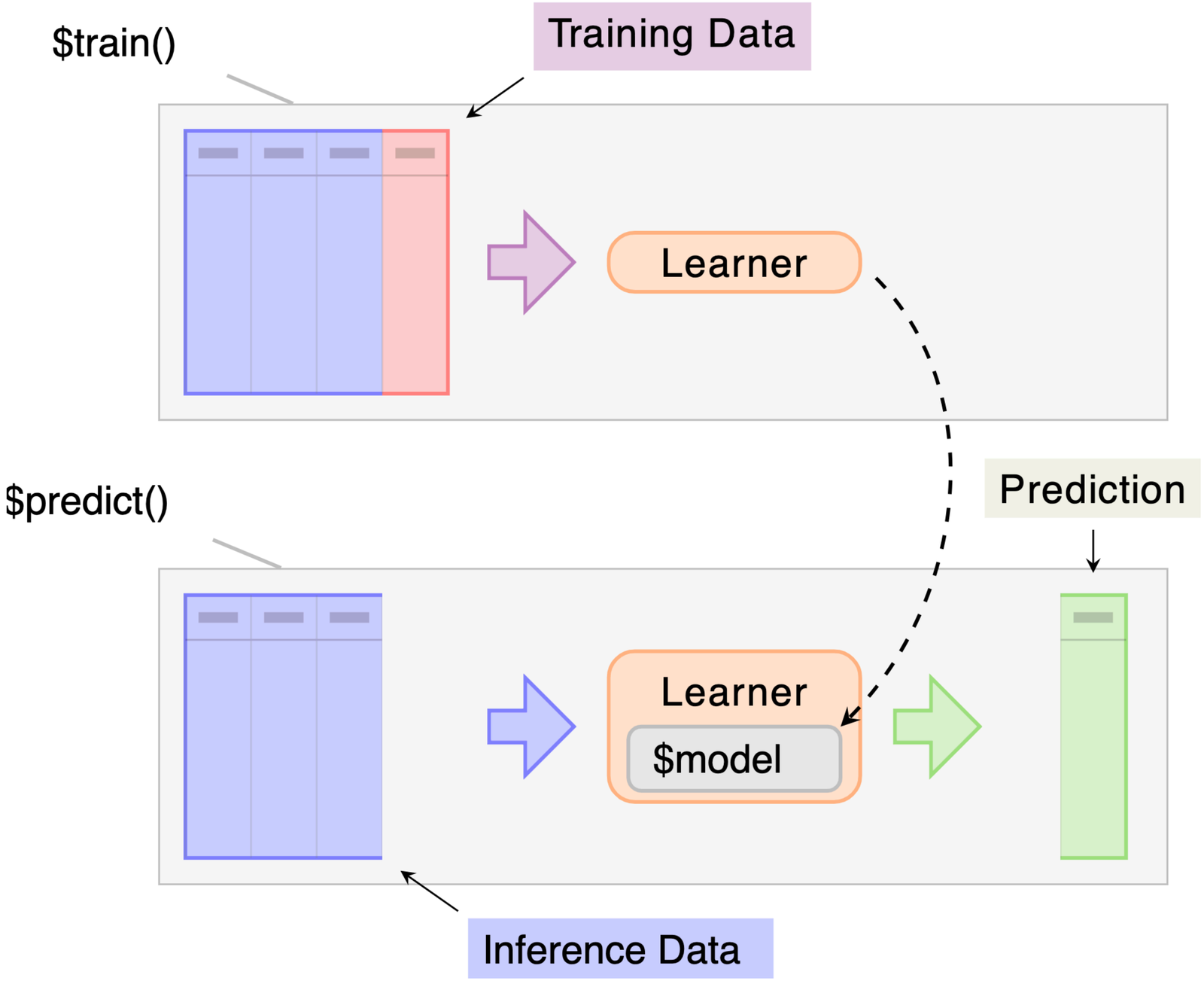

1
Age
p < 0.001
≤ 67
> 67
2
Cancer
p < 0.001
7
Cardiovascular.Diseases
p < 0.001
No
Yes
No
Yes
3
Cardiovascular.Diseases
p < 0.001
8
Cancer
p < 0.001
11
Age
p < 0.001
No
Yes
No
Yes
≤ 80
> 80
Node 4 (n = 8096)
Node 5 (n = 469)
Node 6 (n = 124)
Node 9 (n = 618)
Node 10 (n = 67)
Node 12 (n = 337)
Node 13 (n = 289)
No
Yes
1
0.8
0.6
0.4
0.2
0

Random Forests grows many classification/regression trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but *with replacement*, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

Source: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm



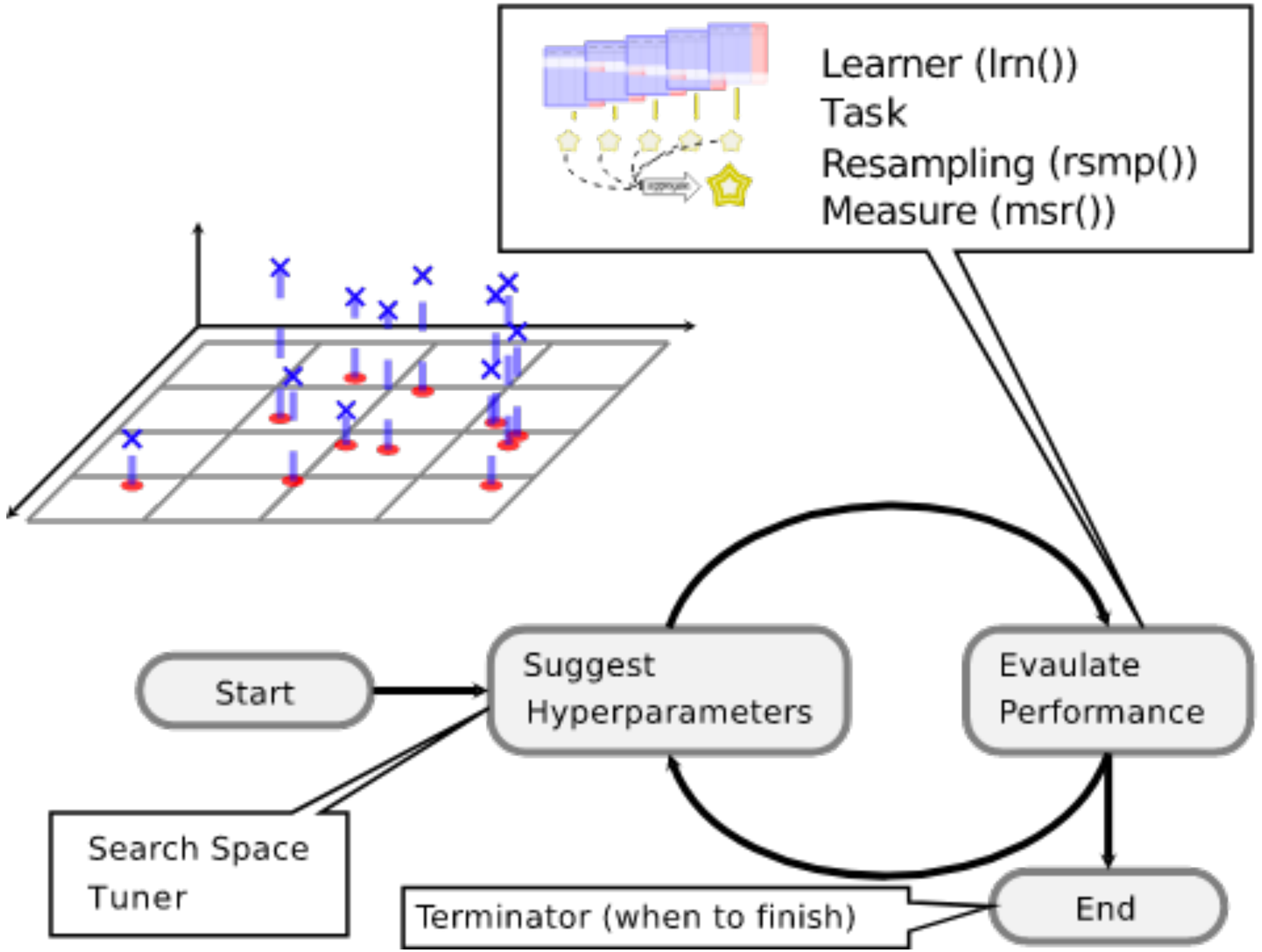
Source: https://mlr3book.mlr-org.com/learners.htm

Day 1

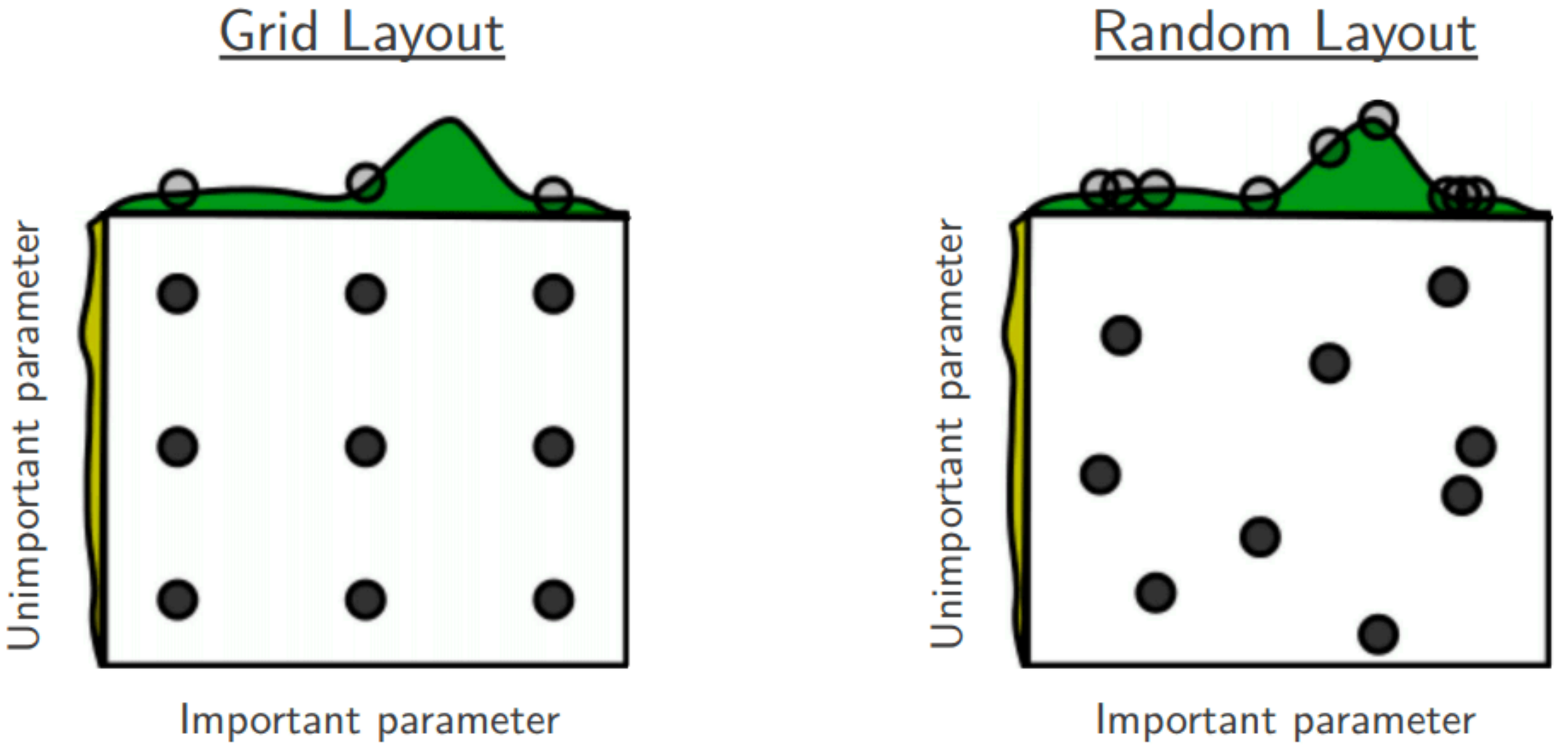
Predictive modeling

Part 4

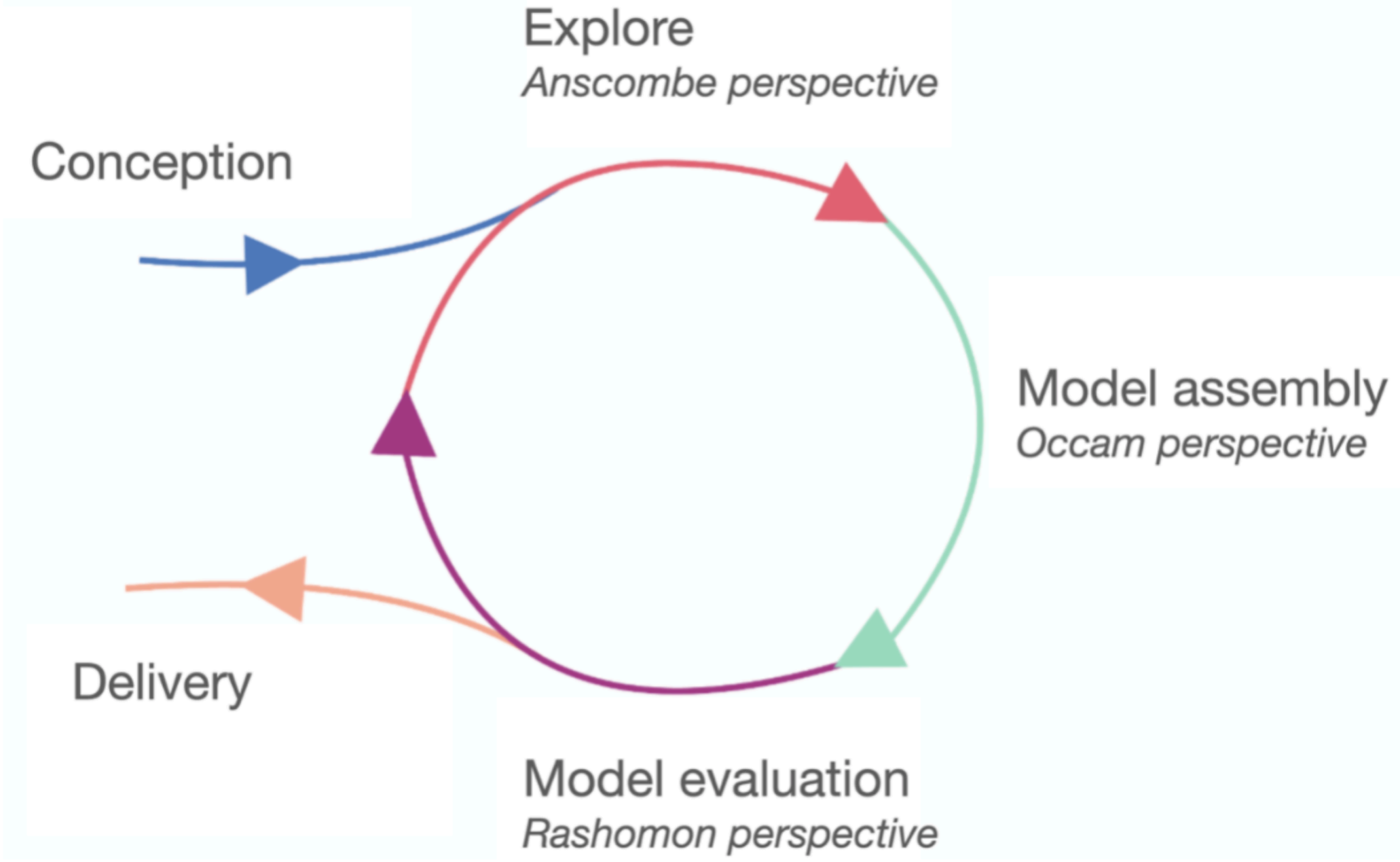
Hyperparameter optimization + Wrap-up



Source: https://mlr3book.mlr-org.com/tuning.html



Source: Random Search for Hyper-Parameter Optimization James Bergstra, Yoshua Bengio


Explore
Anscombe perspective
Conception
Model assembly
Occam perspective
Delivery
Model evaluation
Rashomon perspective

Day 2

Model exploration

Machine Learning is Creating a Crisis in Science

The adoption of machine-learning techniques is contributing to a worrying number of research findings that cannot be repeated by other researchers.

Kevin McCaney
Wed, 02/27/2019 - 11:28

Photo credit: metamorworks/iStock

Google Flu Trends

From Wikipedia, the free encyclopedia

Google Flu Trends was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google Search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.[1]

Google Flu Trends is now no longer publishing current estimates. Historical estimates are still available for download, and current data are offerec

Forbes Billionaires Innovation Leadership Money Con

61,215 views | Mar 23, 2014, 09:00am

Why Google Flu Is A Failu

Steven Salzberg Contributor
Pharma & Healthcare

It seemed like such a good idea at the time.

Support The Guardian
Contribute Subscribe
Search jobs Sign in Search International edition
The Guardian

News Opinion Sport Culture Lifestyle More

World UK Science Cities Global development Football Tech Business Environment Obituaries

Google

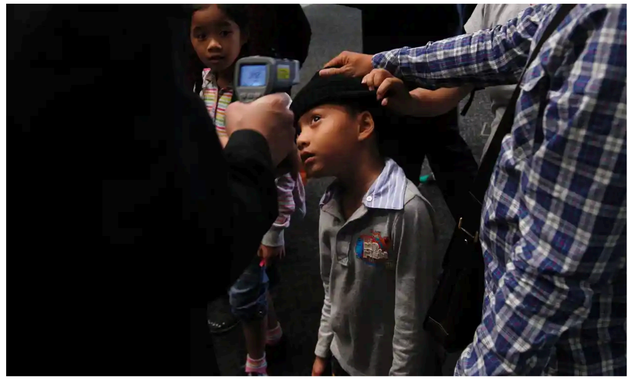
Google Flu Trends is no longer good at predicting flu, scientists find

Researchers warn of 'big data hubris' and the importance of updating analytical models, claiming Google has made inaccurate forecasts for 100 of 108 weeks

Charles Arthur
@charlesarthur
Thu 27 Mar 2014 10.27 GMT

200 7

This article is over 4 years old



Airport security personnel take a body temperature reading of a boy as he arrives at Hong Kong International Airport April 9,2013, following concerns over a deadly strain of bird flu. Photograph: Tyrone Siu/Reuters

most viewed


Dozens of Indian paramilitaries killed in Kashmir car bombing

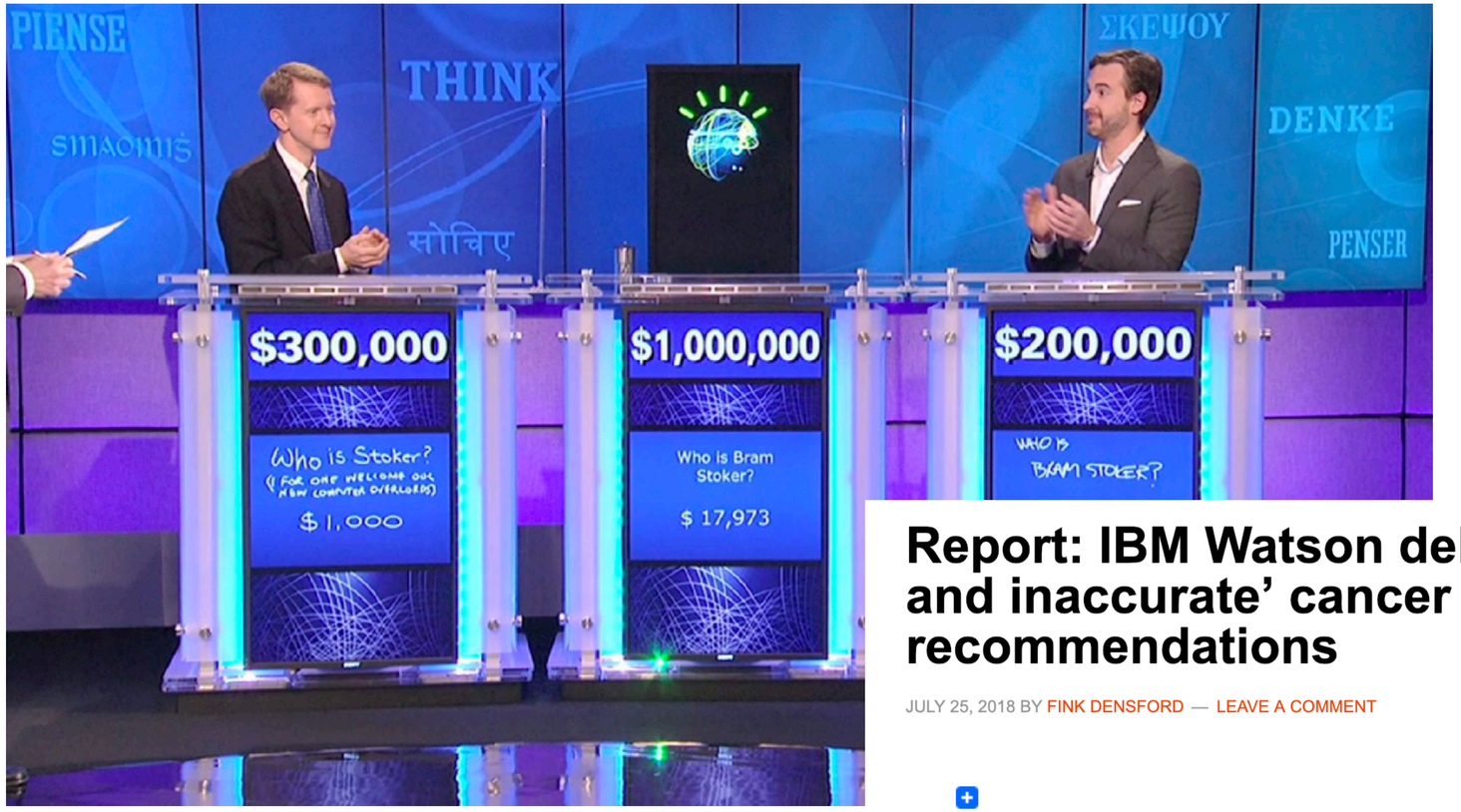

Live Brexit: blow to May's authority as MPs reject her motion by 303 votes to 258 - as it happened


Theresa May defeated on Brexit again as ERG Tories abstain


Live Trump to sign government funding bill and declare national emergency - live


Andrew McCabe says officials discussed removing Trump after Comey firing

https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/



Report: IBM Watson delivered ‘unsafe and inaccurate’ cancer recommendations

JULY 25, 2018 BY FINK DENSFORD — LEAVE A COMMENT

Internal documents from IBM Watson Health (NYSE:IBM) indicate that the company’s Watson for Oncology product often returns “multiple examples of unsafe and incorrect treatment recommendations,” according to a new report from *STAT News*.

The documents come from slides presented last year by IBM Watson Health’s deputy chief health officer, according to the report, and include feedback from customers that indicated the product is “often inaccurate” and that its recommendations bring to light “serious questions about the process for building content and the underlying technology.”

The issues were blamed on training the Watson product received by IBM engineers and physicians at the Memorial Sloan Kettering Cancer Center, which included “synthetic,” or hypothetical patients and cases, instead of real patient data, *STAT* reports.

https://www.massdevice.com/report-ibm-watson-delivered-unsafe-and-inaccurate-cancer-recommendations/

REUTERS World Business Markets Politics TV

Midterm Elections | Imprisoned In Myanmar | Sectors Up Close | Breakingviews | Investing | Future of Money | Charged: The Future of Aut

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

Amazon scraps secret AI showed bias against wor

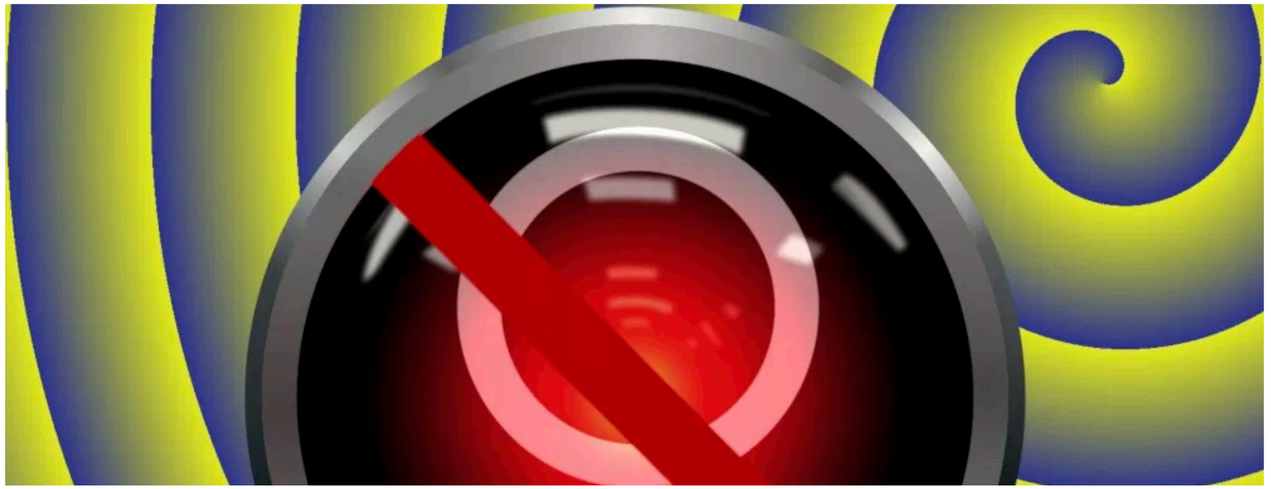
Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's specialists uncovered a big problem: their

The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.

Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist



THE VERGE TECH SCIENCE C

TECH AMAZON ARTIFICIAL INTELLIGENCE

21

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

COMPAS (software)

From Wikipedia, the free encyclopedia

COMPAS, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a case management and decision support tool developed by Northpointe (now equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[1][2]

Contents [hide]

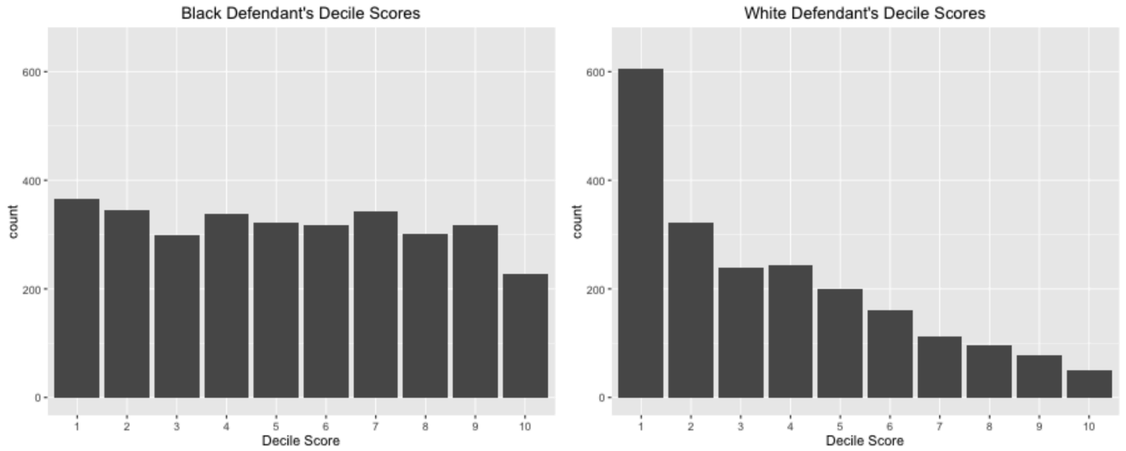
Risk Assessment [edit]

The COMPAS software uses an algorithm to assess poten
scales for general and violent recidivism, and for pretrial r
Practitioner's Guide, the scales were designed using beha
high relevance to recidivism and criminal careers."[3]

Analysis

We analyzed the COMPAS scores for "Risk of Recidivism" and "Risk of Violent Recidivism." We did not analyze the COMPAS score for "Risk of Failure to Appear."

We began by looking at the risk of recidivism score. Our initial analysis looked at the simple distribution of the COMPAS decile scores among whites and blacks. We plotted the distribution of these scores for 6,172 defendants who had not been arrested for a new offense or who had recidivated within two years.



These histograms show that scores for white defendants were skewed toward lower-risk categories, while black defendants were evenly distributed across scores. In our two-year sample, there were 3,175 black defendants and 2,103 white defendants, with 1,175 female defendants and 4,997 male defendants. There were 2,809 defendants who recidivated within two years in this sample.

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

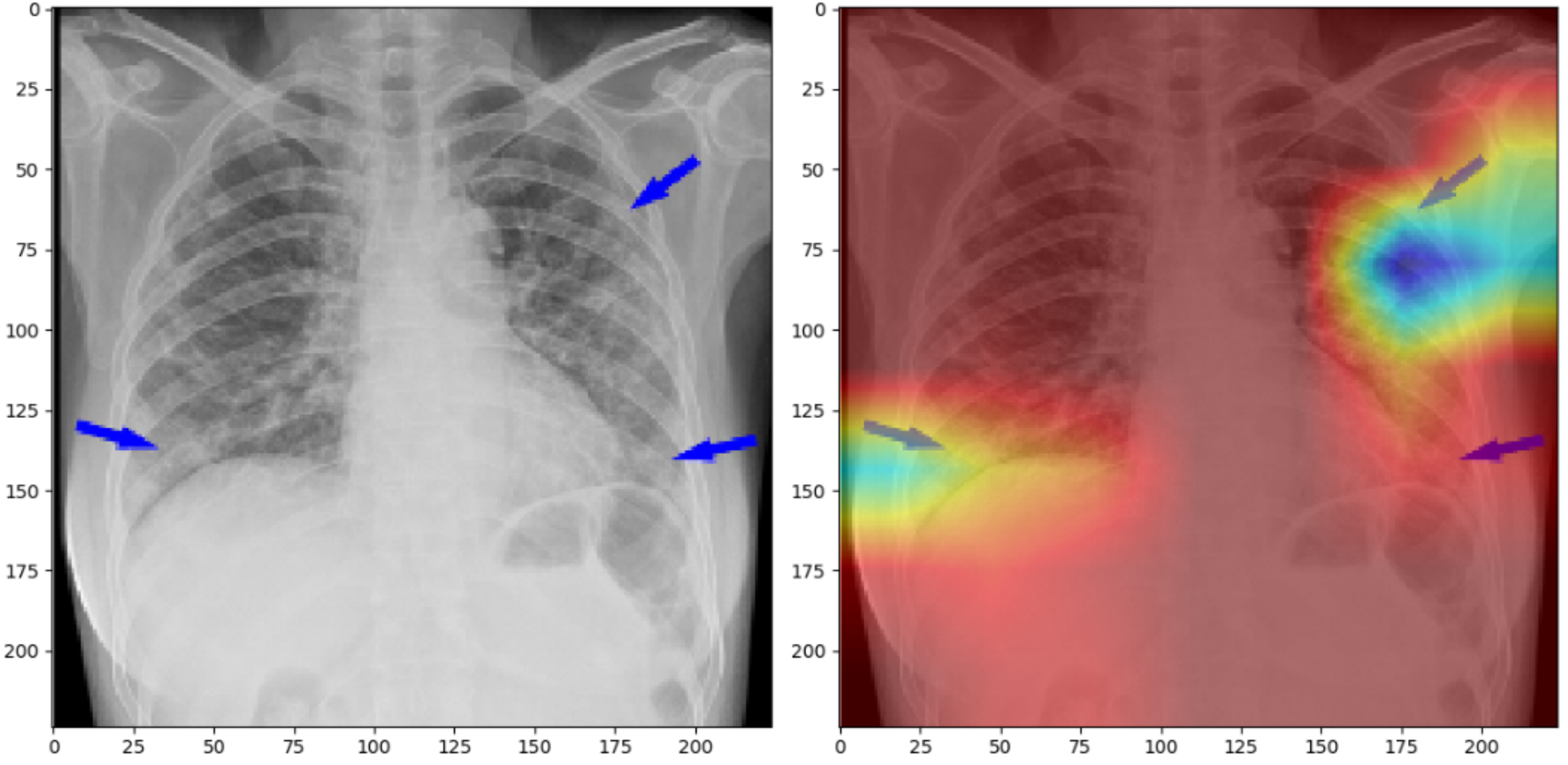
Model debugging

(a) Husky classified as wolf

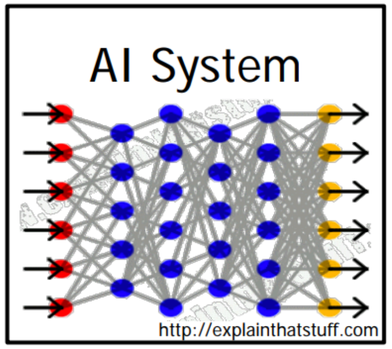
(b) Explanation

"Why Should I Trust You?" Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). https://arxiv.org/pdf/1602.04938.pdf

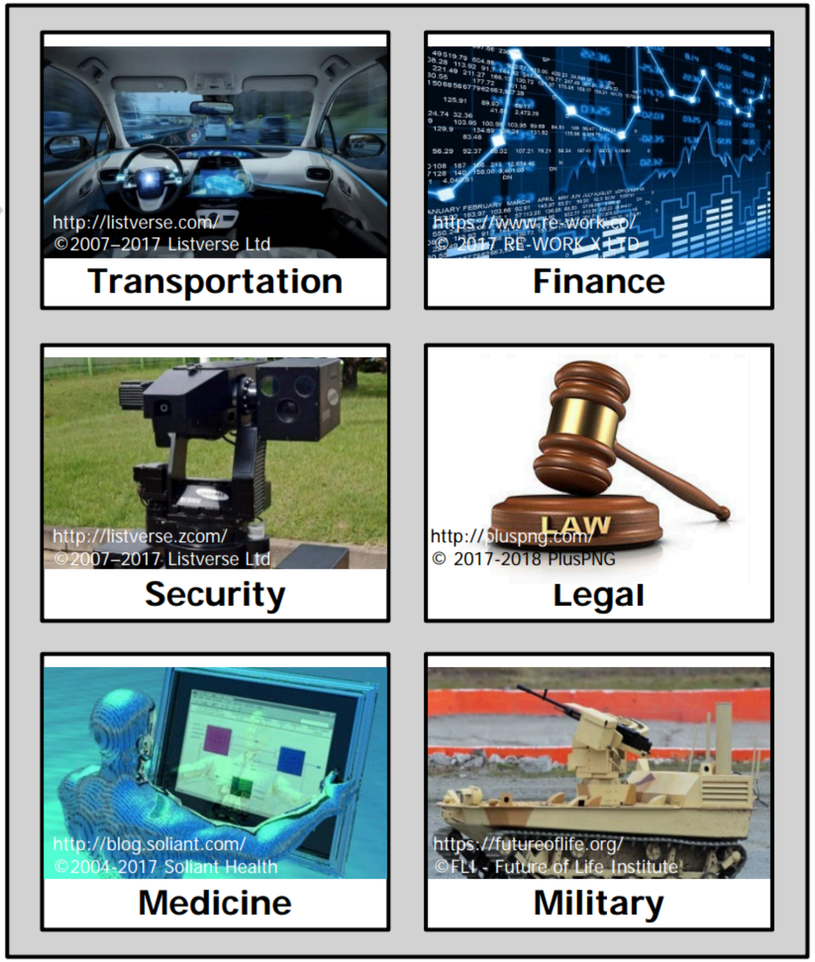
Ground Truth Class: 1 (COVID-19)
Predicted Class: 1 (COVID-19)
Prediction probabilities: ['0.00', '1.00']



The Need for Explainable AI



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- The current generation of AI systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to users
- Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners

Right to explanation

From Wikipedia, the free encyclopedia

In the regulation of algorithms, particularly artificial intelligence and its subfield of machine learning, a **right to explanation** (or **right to *an* explanation**) is a right to be given an explanation for an output of the algorithm. Such rights primarily refer to individual rights to be given an explanation for decisions that significantly affect an individual, particularly legally or financially. For example, a person who applies for a loan and is denied may ask for an explanation, which could be "Credit bureau X reports that you declared bankruptcy last year; this is the main factor in considering you too likely to default, and thus we will not give you the loan you applied for."

Some such legal rights already exist, while the scope of a general "right to explanation" is a matter of ongoing debate.

Contents [hide]

Prawo do wyjaśnienia decyzji kredytowej dla każdego! Sukces Panoptykonu!

13.02.2019

Po dzisiejszym posiedzeniu połączonych komisji sejmowych mamy dobre wiadomości: udało się przekonać rząd i posłów do wprowadzenia zmian w prawie bankowym, które zmienią proces przyznawania kredytów. W miejsce „czarnej skrzynki", która wypluwa niezrozumiałe decyzje bez uzasadnienia, ma się pojawić przejrzysta procedura, wzmacniająca prawa osób ubiegających się o kredyt. To duży przełom w relacjach bank–klient i odpowiedź ustawodawcy na postulaty, które zgłaszamy od dawna.

Krótka historia długiej walki

Od momentu, w którym pojawił się rządowy projekt tzw. ustawy sektorowej (wdrażającej RODO w 160 ustawach) i propozycje zmian w prawie bankowym, walczyliśmy o to, żeby każdy konsument starający się o kredyt miał prawo zażądać od banku wyjaśnienia oceny zdolności kredytowej i decyzji, którą w jego sprawie podjął bank. W pierwszej wersji projektu było przewidziane jedynie szczątkowe i nieprecyzyjnie sformułowane prawo do wyjaśnienia, dotyczące wyłącznie decyzji podejmowanych w sposób automatyczny, czyli bez udziału człowieka. Uznaliśmy, że to za mało, i walczyliśmy o więcej. Tym bardziej, że od kilku lat prawo do wyjaśnienia oceny zdolności kredytowej (w każdej sytuacji) mają przedsiębiorcy.

Na pierwszym posiedzeniu specjalnej podkomisji powołanej do prac nad tym pakietem zaapelowaliśmy do rządu i posłów o wprowadzanie dalej idących gwarancji dla konsumentów. Ten pomysł poparli posłowie wszystkich partii, a przewodniczący podkomisji Edward Siarka wezwał Ministerstwo Cyfryzacji do wypracowania kompromisu ze „stroną społeczną" i przedstawicielami banków. Negocjacje i praca nad brzmieniem przepisów zajęły trzy tygodnie, ale warto było poczekać.

Koniec z „czarną skrzynką" przy udzielaniu kredytów

Na wniosek ubiegającego się o kredyt konsumenta bank przedstawi mu czynniki, w tym dane osobowe,

Na wniosek ubiegającego się o kredyt konsumenta bank przedstawi mu czynniki, w tym dane osobowe, które miały wpływ na ocenę zdolności kredytowej. To kwintesencja zgłoszonej dzisiaj przez rząd i popartej przez posłów ze wszystkich ugrupowań zmiany w prawie bankowym.

Cathy O'Neil:
The era of blind faith in ~~big data~~ black boxes must end

- “You don’t see a lot of skepticism,” she says. “The algorithms are like shiny new toys that we can’t resist using. We trust them so much that we project meaning on to them.”
- Ultimately algorithms, according to O’Neil, reinforce discrimination and widen inequality, “using people’s fear and trust of mathematics to prevent them from asking questions”.

Day 2

Model exploration

Part 1

Model level analysis - variable importance

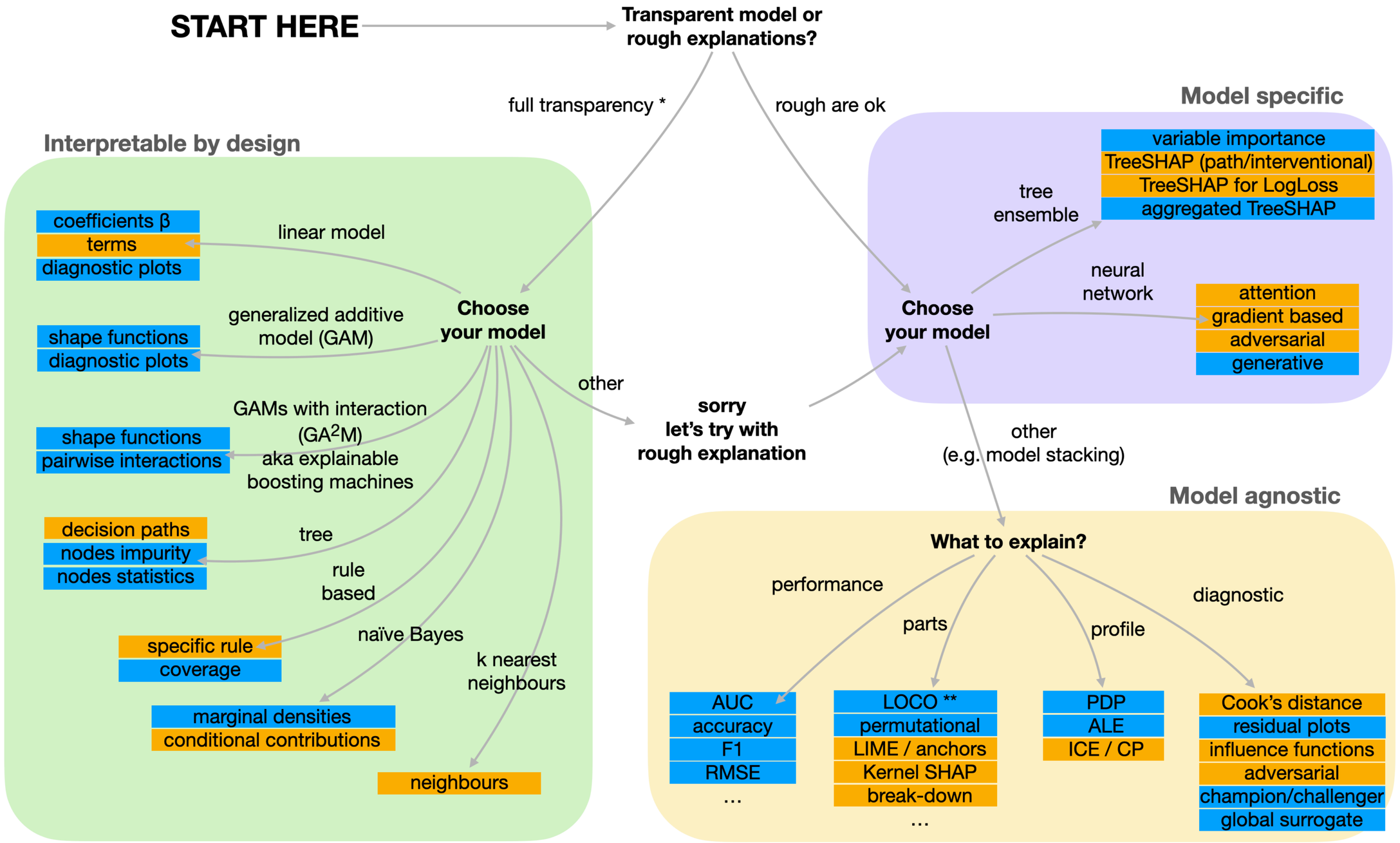
The central assumption of the machine learning models:

The future will be similar to the past

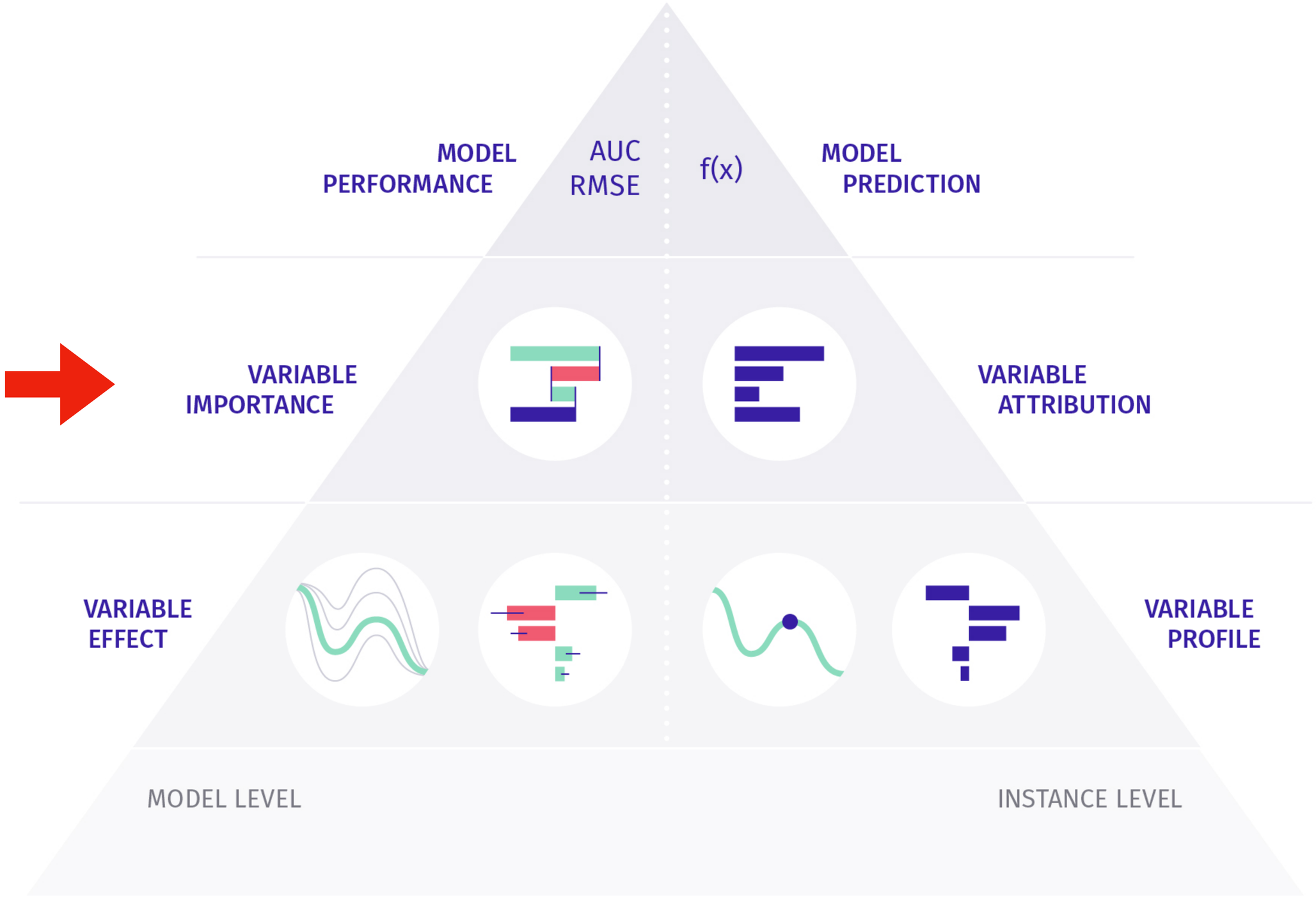
The central assumption of the machine learning models:

COVID19

The future will be similar to the past



https://arxiv.org/abs/2009.13248


MODEL PERFORMANCE
AUC
RMSE
f(x)
MODEL PREDICTION
VARIABLE IMPORTANCE
VARIABLE ATTRIBUTION
VARIABLE EFFECT
VARIABLE PROFILE
MODEL LEVEL
INSTANCE LEVEL

Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective

Statistics > Methodology

All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously

Aaron Fisher, Cynthia Rudin, Francesca Dominici

(Submitted on 4 Jan 2018 (v1), last revised 23 Dec 2019 (this version, v5))

Variable importance (VI) tools describe how much covariates contribute to a prediction model's accuracy. However, important variables for one well-performing model (for example, a linear model $f(\mathbf{x}) = \mathbf{x}^T\beta$ with a fixed coefficient vector β) may be unimportant for another model. In this paper, we propose model class reliance (MCR) as the range of VI values across all well-performing model in a prespecified class. Thus, MCR gives a more comprehensive description of importance by accounting for the fact that many prediction models, possibly of different parametric forms, may fit the data well. In the process of deriving MCR, we show several informative results for permutation-based VI estimates, based on the VI measures used in Random Forests. Specifically, we derive connections between permutation importance estimates for a single prediction model, U-statistics, conditional variable importance, conditional causal effects, and linear model coefficients. We then give probabilistic bounds for MCR, using a novel, generalizable technique. We apply MCR to a public data set of Broward County criminal records to study the reliance of recidivism prediction models on sex and race. In this application, MCR can be used to help inform VI for unknown, proprietary models.

Comments:	The final, published article is now available at this http URL . This version contains minor changes to the introductory sections, and some typo corrections. The title of this article changed twice during the revision process (see v1 and v3 on arxiv)
Subjects:	**Methodology (stat.ME)**
Journal reference:	Journal of Machine Learning Research 20 (177), 1-81, 2019
Cite as:	arXiv:1801.01489 [stat.ME]
	(or arXiv:1801.01489v5 [stat.ME] for this version)

https://arxiv.org/abs/1801.01489

Cynthia Rudin

OBSERWUJ

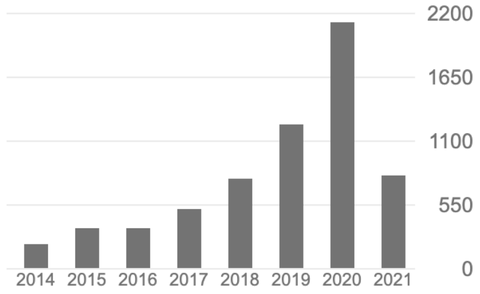
Professor of Computer Science, ECE, and Statistics, Duke University
Zweryfikowany adres z cs.duke.edu - Strona główna

machine learning interpretability data science

TYTUŁ	CYTOWANE PRZEZ	ROK
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead C Rudin Nature Machine Intelligence 1 (5), 206-215	976 *	2019
Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model B Letham, C Rudin, TH McCormick, D Madigan Annals of Applied Statistics 9 (3), 1350-1371	544 *	2015
All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. A Fisher, C Rudin, F Dominici Journal of Machine Learning Research 20 (177), 1-81	341 *	2019
The P-Norm Push: A simple convex ranking algorithm that concentrates at the top of the list C Rudin The Journal of Machine Learning Research 10, 2233-2271	235 *	2009
Supersparse linear integer models for optimized medical scoring systems B Ustun, C Rudin Machine Learning 102 (3), 349-391	230 *	2016
Machine learning for the New York City power grid C Rudin, D Waltz, RN Anderson, A Boulanger, A Salleb-Aouissi, M Chow, ... IEEE transactions on pattern analysis and machine intelligence 34 (2), 328-345	229	2011
Margin-based ranking and an equivalence between AdaBoost and RankBoost C Rudin, RE Schapire The Journal of Machine Learning Research 10, 2193-2232	214 *	2009
The bayesian case model: A generative approach for case-based reasoning and prototype classification B Kim. C Rudin. JA Shah	199	2014

Cytowane przez WYŚWIETL WSZYSTKO

	Wszystkie	Od 2016
Cytowania	7304	5854
h-indeks	41	35
i10-indeks	89	73



Dostęp publiczny WYŚWIETL WSZYSTKO

0 artykułów 23 artykuły

niedostępne dostępne

Objęte finansowaniem

Współautorzy WYŚWIETL WSZYSTKICH


Berk Ustun
Google AI / UCSD


Tyler H. McCormick
University of Washington


David Madigan
Professor of Statistics, Northeast…

Variable importance for random forest

Variable importance

In every tree grown in the forest, put down the oob cases and count the number of votes cast for the correct class. Now randomly permute the values of variable m in the oob cases and put these cases down the tree. Subtract the number of votes for the correct class in the variable-m-permuted oob data from the number of votes for the correct class in the untouched oob data. The average of this number over all trees in the forest is the raw importance score for variable m.

If the values of this score from tree to tree are independent, then the standard error can be computed by a standard computation. The correlations of these scores between trees have been computed for a number of data sets and proved to be quite low, therefore we compute standard errors in the classical way, divide the raw score by its standard error to get a z-score, ands assign a significance level to the z-score assuming normality.

If the number of variables is very large, forests can be run once with all the variables, then run again using only the most important variables from the first run.

For each case, consider all the trees for which it is oob. Subtract the percentage of votes for the correct class in the variable-m-permuted oob data from the percentage of votes for the correct class in the untouched oob data. This is the local importance score for variable m for this case, and is used in the graphics program RAFT.

https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#varimp

Do we need a new measure for Variable Importance?

Several common approaches for variable selection, or for describing relationships between variables, do not necessarily capture a variable's importance. Null hypothesis testing methods may identify a relationship, but do not describe the relationship's strength. Similarly, checking whether a variable is included by a sparse model-fitting algorithm, such as the Lasso (Hastie et al., 2009), does not describe the extent to which the variable is relied on. Partial dependence plots (Breiman et al., 2001; Hastie et al., 2009) can be difficult to interpret if multiple variables are of interest, or if the prediction model contains interaction effects.

Another common VI procedure is to run a model-fitting algorithm twice, first on all of the data, and then again after removing X_1 from the data set. The losses for the two resulting models are then compared to determine the importance, or "necessity," of X_1 (Gevrey et al., 2003). Because this measure is a function of two prediction models rather than one, it does not measure how much either individual model relies on X_1. We refer

http://www.jmlr.org/papers/volume20/18-760/18-760.pdf

Variable importance as the drop in the loss after variable permutation

$$MR(f) := \frac{\text{Expected loss of } f \text{ under noise}}{\text{Expected loss of } f \text{ without noise}}. \tag{2.1}$$

The added noise must satisfy certain properties, namely, it must render X_1 completely uninformative of the outcome Y, without altering the marginal distribution of X_1 (for details, see Section 3, as well as Breiman, 2001; Breiman et al., 2001).

$$\widehat{MR}(f) := \frac{\text{In-sample loss of } f \text{ under noise}}{\text{In-sample loss of } f \text{ without noise}}, \tag{2.3}$$

http://www.jmlr.org/papers/volume20/18-760/18-760.pdf

One permutation is not enough. Check all of them

As a reference point, we compare $e_{\text{switch}}(f)$ against the standard expected loss when none of the variables are switched, $e_{\text{orig}}(f) := \mathbb{E}L(f, (Y, X_1, X_2))$. From these two quantities, we formally define *model reliance* (MR) as the ratio,

$$MR(f) := \frac{e_{\text{switch}}(f)}{e_{\text{orig}}(f)}, \tag{3.1}$$

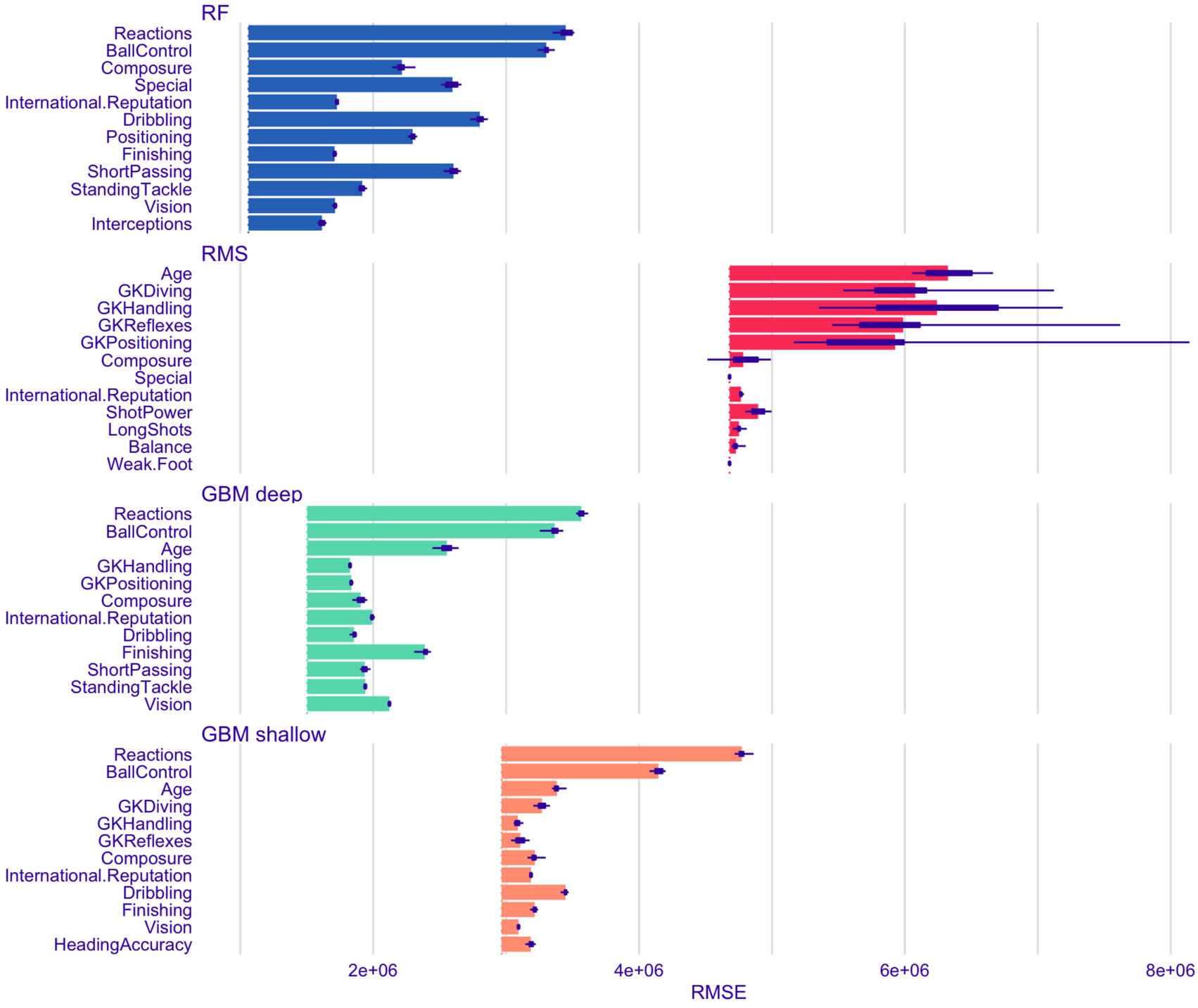
$$e_{\text{switch}}(f) := \mathbb{E}L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\}$$

Given a model f and data set $\mathbf{Z} = \begin{bmatrix} \mathbf{y} & \mathbf{X} \end{bmatrix}$, we estimate $MR(f)$ by separately estimating the numerator and denominator of Eq 3.1. We estimate $e_{\text{orig}}(f)$ with the standard empirical loss,

$$\hat{e}_{\text{orig}}(f) := \frac{1}{n} \sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}. \tag{3.2}$$

We estimate $e_{\text{switch}}(f)$ by performing a "switch" operation across all observed pairs, as in

$$\hat{e}_{\text{switch}}(f) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L\{f, (\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\}. \tag{3.3}$$

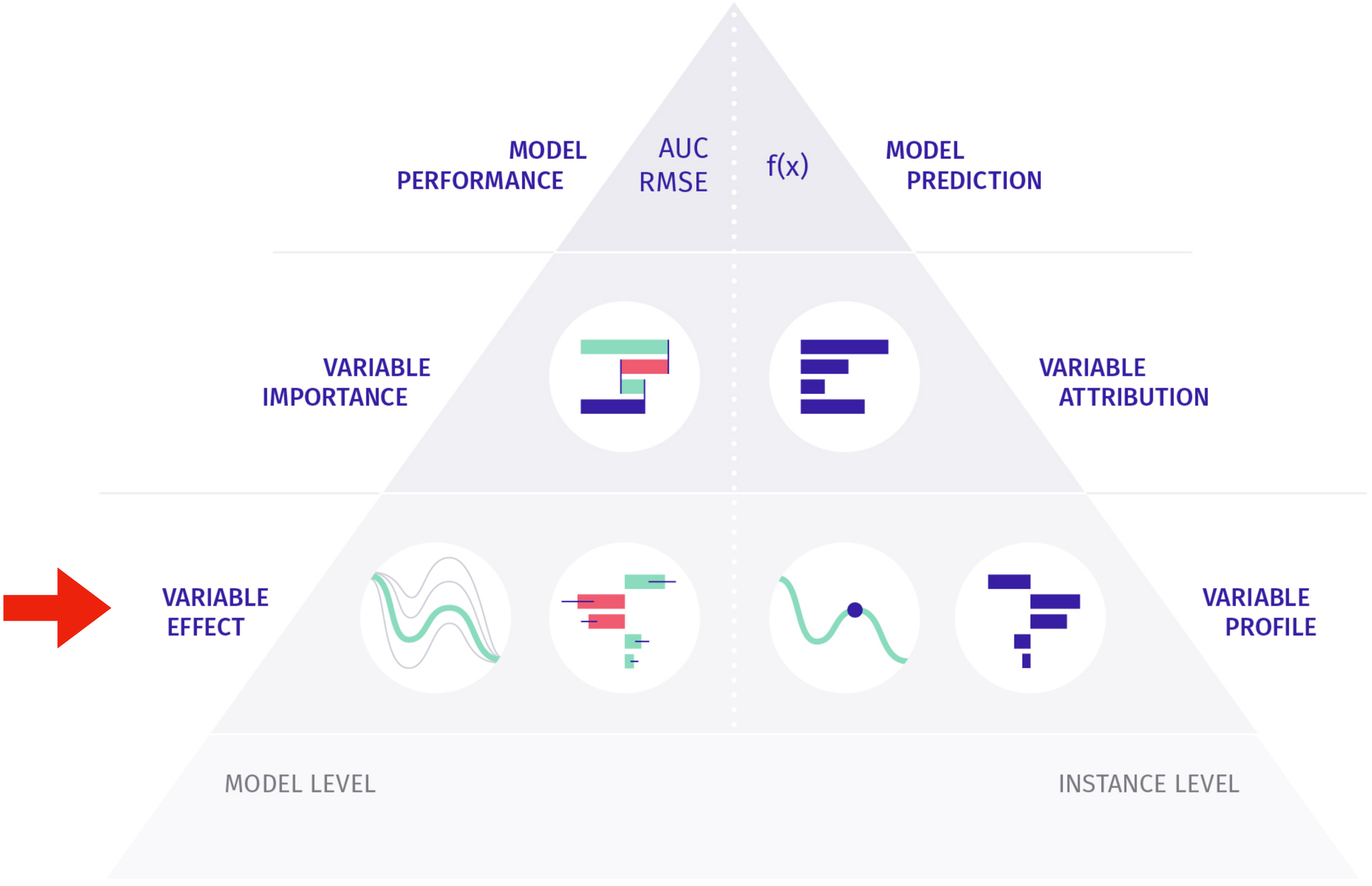

RF
Reactions
BallControl
Composure
Special
International.Reputation
Dribbling
Positioning
Finishing
ShortPassing
StandingTackle
Vision
Interceptions
RMS
Age
GKDiving
GKHandling
GKReflexes
GKPositioning
Composure
Special
International.Reputation
ShotPower
LongShots
Balance
Weak.Foot
GBM deep
Reactions
BallControl
Age
GKHandling
GKPositioning
Composure
International.Reputation
Dribbling
Finishing
ShortPassing
StandingTackle
Vision
GBM shallow
Reactions
BallControl
Age
GKDiving
GKHandling
GKReflexes
Composure
International.Reputation
Dribbling
Finishing
Vision
HeadingAccuracy
2e+06
4e+06
6e+06
8e+06
RMSE

Day 2

Model exploration

Part 2

Model level analysis - variable profile


MODEL PERFORMANCE
AUC
RMSE
f(x)
MODEL PREDICTION
VARIABLE IMPORTANCE
VARIABLE ATTRIBUTION
VARIABLE EFFECT
VARIABLE PROFILE
MODEL LEVEL
INSTANCE LEVEL

Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models

Statistics > Methodology

Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models

Daniel W. Apley, Jingyu Zhu

(Submitted on 27 Dec 2016 (v1), last revised 19 Aug 2019 (this version, v2))

When fitting black box supervised learning models (e.g., complex trees, neural networks, boosted trees, random forests, nearest neighbors, local kernel-weighted methods, etc.), visualizing the main effects of the individual predictor variables and their low-order interaction effects is often important, and partial dependence (PD) plots are the most popular approach for accomplishing this. However, PD plots involve a serious pitfall if the predictor variables are far from independent, which is quite common with large observational data sets. Namely, PD plots require extrapolation of the response at predictor values that are far outside the multivariate envelope of the training data, which can render the PD plots unreliable. Although marginal plots (M plots) do not require such extrapolation, they produce substantially biased and misleading results when the predictors are dependent, analogous to the omitted variable bias in regression. We present a new visualization approach that we term accumulated local effects (ALE) plots, which inherits the desirable characteristics of PD and M plots, without inheriting their preceding shortcomings. Like M plots, ALE plots do not require extrapolation; and like PD plots, they are not biased by the omitted variable phenomenon. Moreover, ALE plots are far less computationally expensive than PD plots.

Comments: The R package ALEPlot is available on CRAN. The new version contains refined definitions of ALE effects, a new illustrative example, theorems and proofs of asymptotic properties of ALE effects and estimators, and extra implementation details

Subjects: **Methodology (stat.ME)**

Cite as: arXiv:1612.08468 [stat.ME]

(or arXiv:1612.08468v2 [stat.ME] for this version)

Bibliographic data

Submission history

https://arxiv.org/abs/1612.08468

Daniel Apley

OBSERWUJ

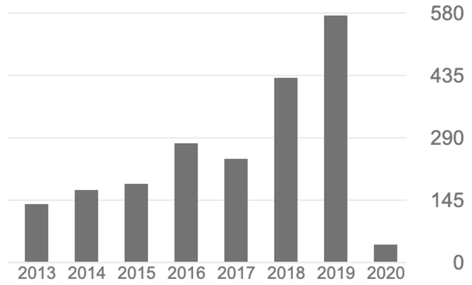
Northwestern University
Zweryfikowany adres z u.northwestern.edu

statistics machine learning manufacturing quality control

TYTUŁ	CYTOWANE PRZEZ	ROK
Quantification of model uncertainty: Calibration, model discrepancy, and identifiability PD Arendt, DW Apley, W Chen Journal of Mechanical Design 134 (10), 100908	198	2012
Understanding the effects of model uncertainty in robust design with computer experiments DW Apley, J Liu, W Chen Journal of Mechanical Design 128 (4), 945-958	183	2006
Diagnosis of multiple fixture faults in panel assembly DW Apley, J Shi	172	1998
Local Gaussian process approximation for large computer experiments RB Gramacy, DW Apley Journal of Computational and Graphical Statistics 24 (2), 561-578	158	2015
The GLRT for statistical process control of autocorrelated processes DW Apley, J Shi IIE transactions 31 (12), 1123-1134	155	1999
A non-stationary covariance-based Kriging method for metamodelling in engineering design Y Xiong, W Chen, D Apley, X Ding International Journal for Numerical Methods in Engineering 71 (6), 733-756	148	2007
A better understanding of model updating strategies in validating engineering models Y Xiong, W Chen, KL Tsui, DW Apley Computer methods in applied mechanics and engineering 198 (15-16), 1327-1337	139	2009
A factor-analysis method for diagnosing variability in mulitvariate manufacturing processes DW Apley, J Shi Technometrics 43 (1), 84-95	126	2001
A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality MA Bessa, R Bostanabad, Z Liu, A Hu, DW Apley, C Brinson, W Chen, ... Computer Methods in Applied Mechanics and Engineering 320, 633-667	125	2017
The Autoregressive *T*2 Chart for Monitoring Univariate Autocorrelated Processes DW Apley, F Tsung Journal of Quality Technology 34 (1), 80-96	116	2002

Cytowane przez WYŚWIETL WSZYSTKO

	Wszystkie	Od 2015
Cytowania	3202	1754
h-indeks	26	22
i10-indeks	61	40



Dataset level

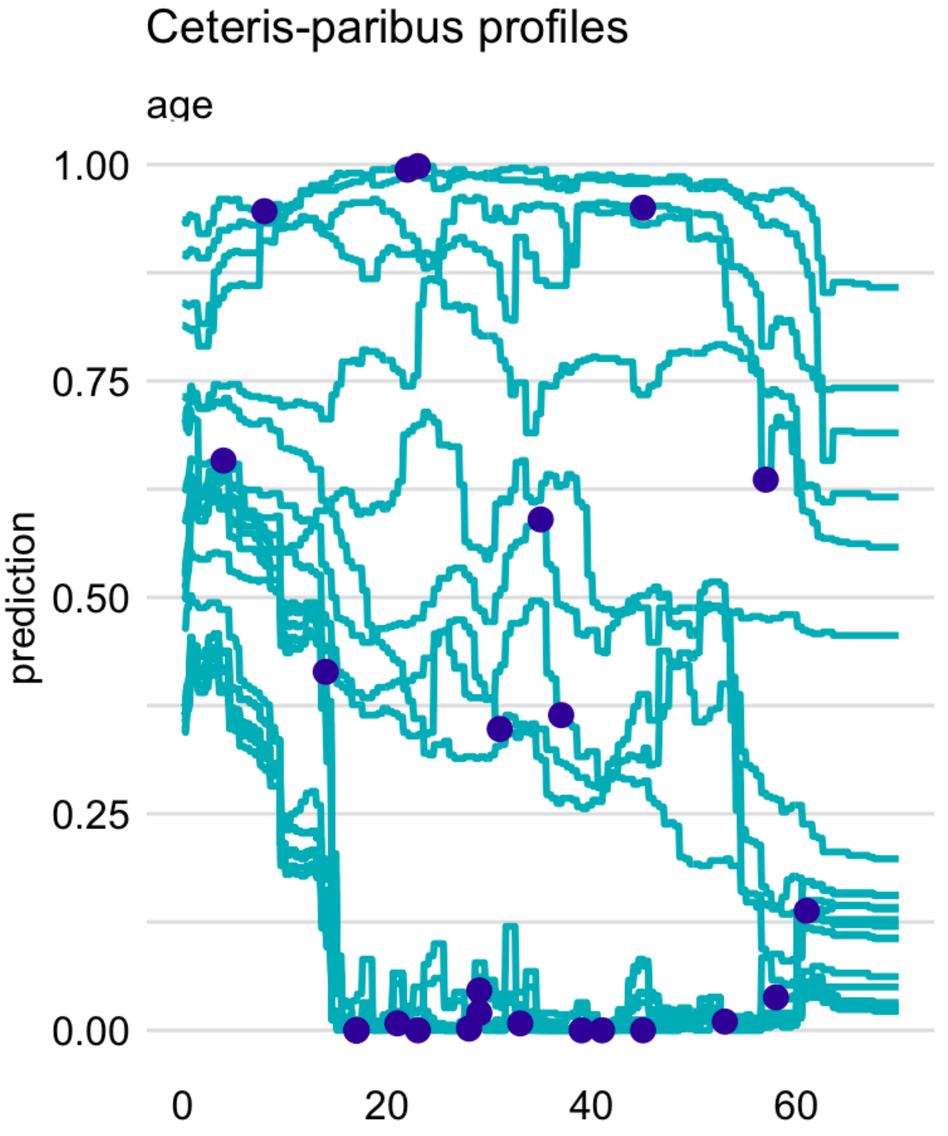
The value of a PD profile for model $f()$ and explanatory variable X^j at z is defined as follows:

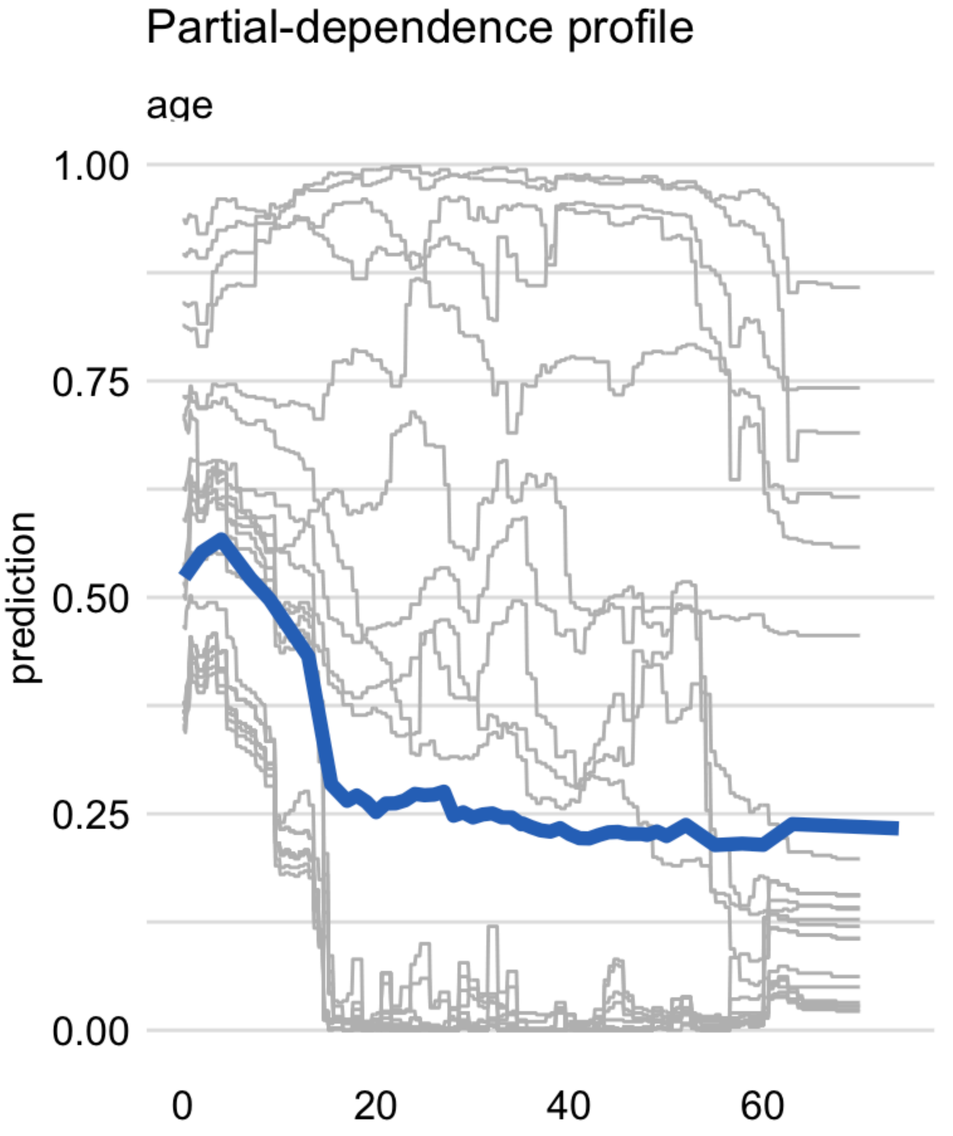
$$g_{PD}^{f,j}(z) = E_{X^{-j}}[f(X^{j|=z})]. \tag{18.1}$$

Usually, we do not know the true distribution of X^{-j}. We can estimate it, however, by the empirical distribution of N, say, observations available in a training dataset. This leads to the use of the average of CP profiles for X^j as an estimator of the PD profile:

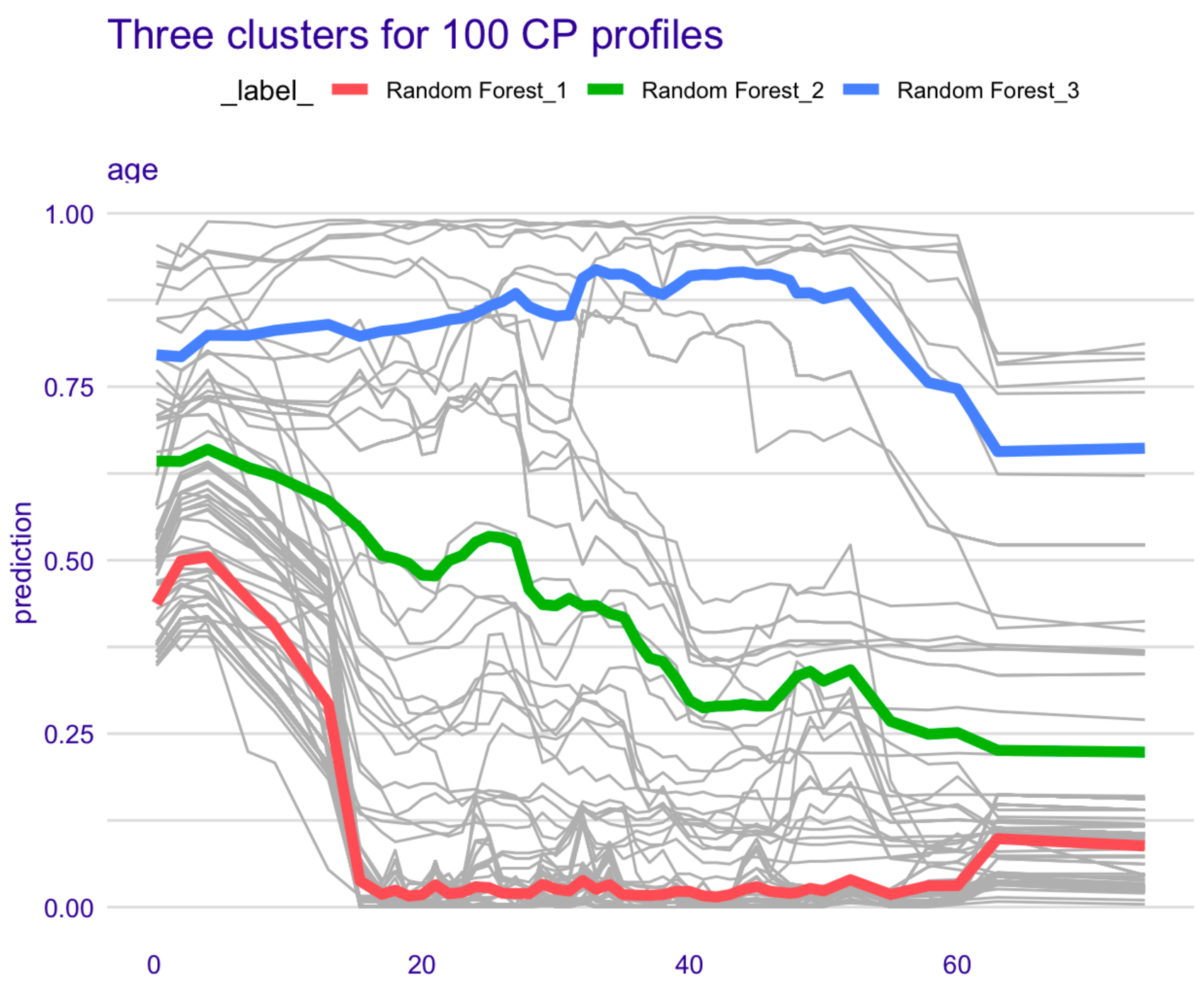
$$\hat{g}_{PD}^{f,j}(z) = \frac{1}{N}\sum_{i=1}^{N} f(x_i^{j|=z}). \tag{18.2}$$

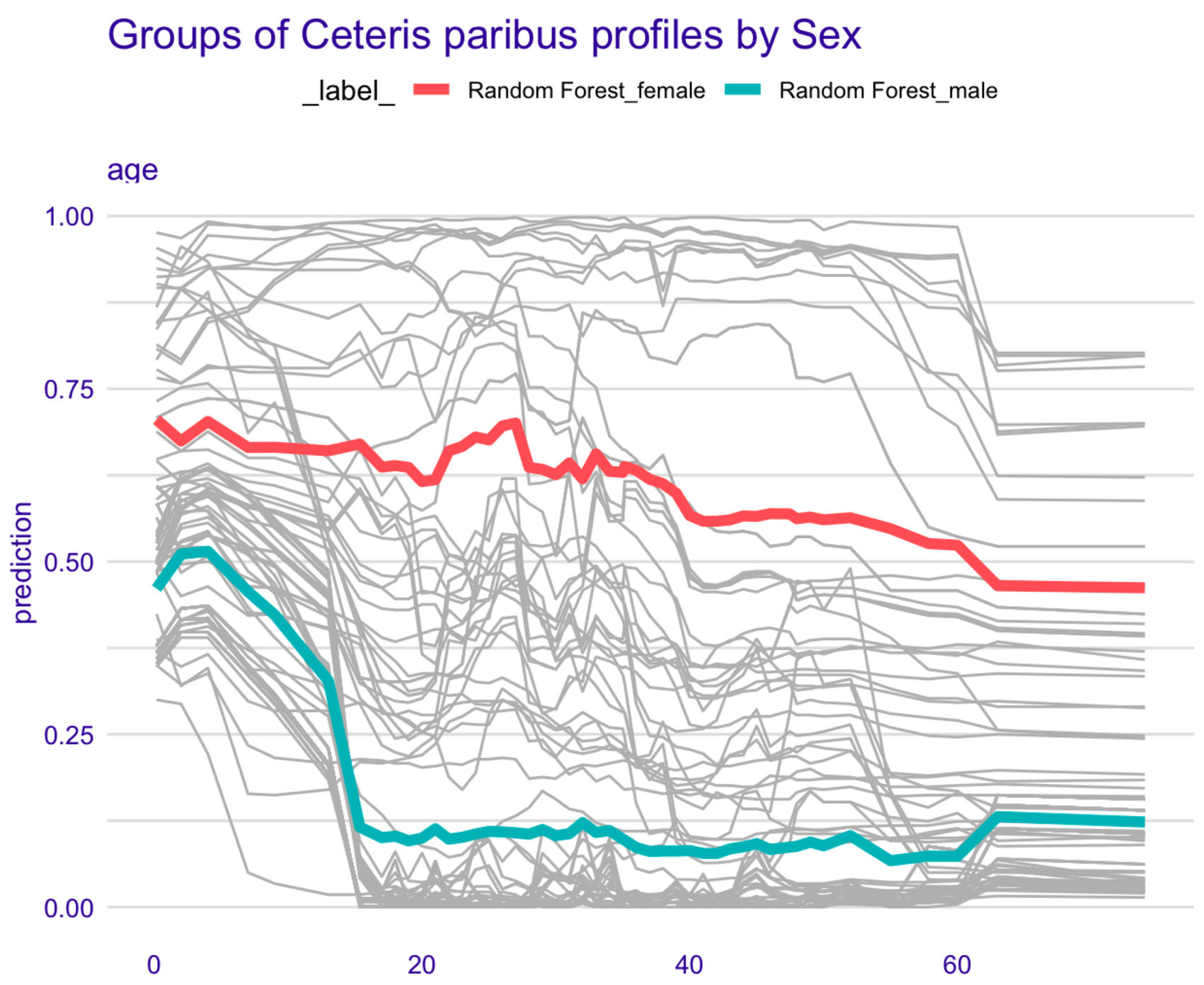
https://ema.drwhy.ai/accumulatedLocalProfiles.html

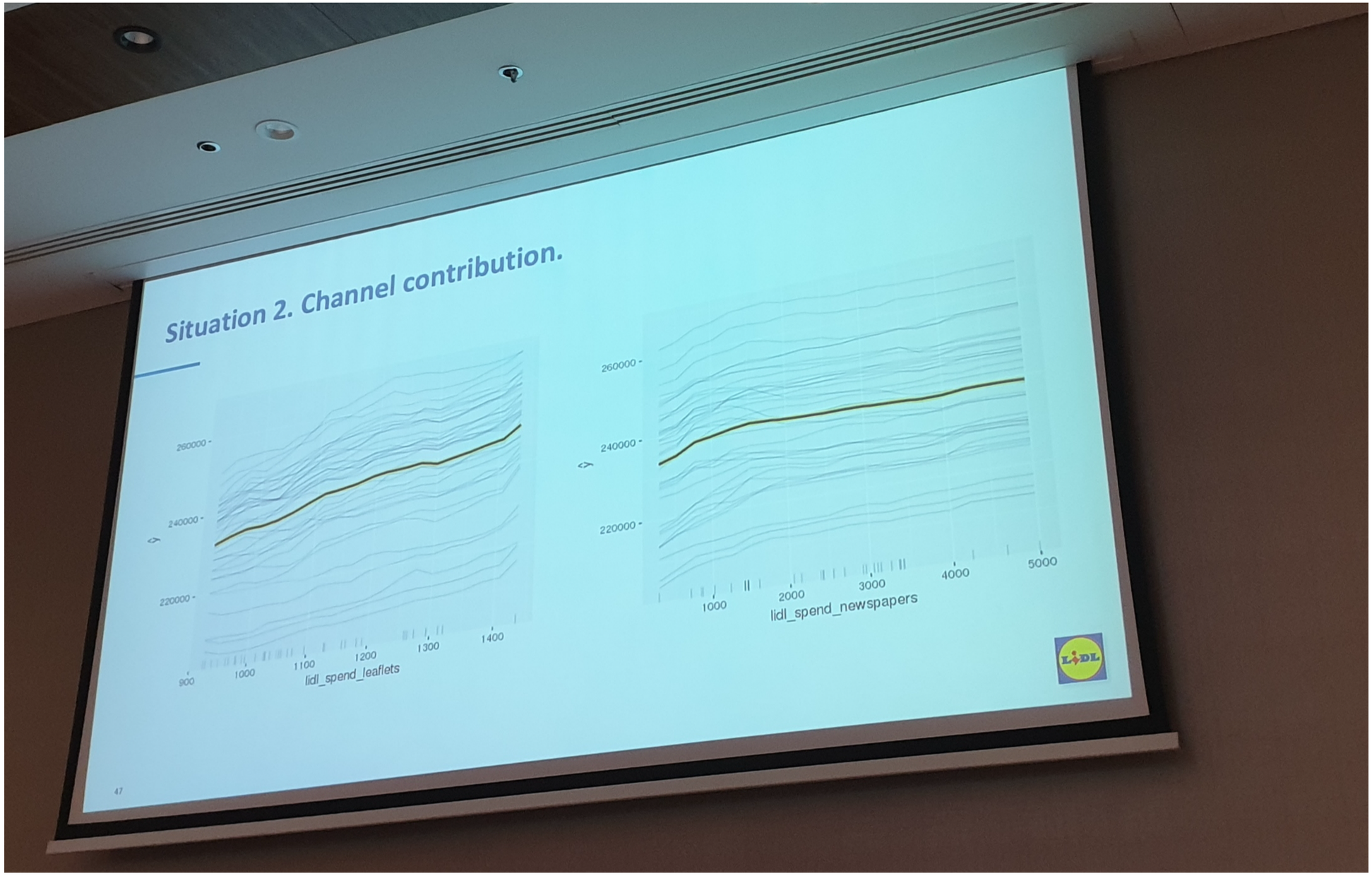




http://ema.drwhy.ai/partialDependenceProfiles.html


Three clusters for 100 CP profiles
label
Random Forest_1
Random Forest_2
Random Forest_3
age
prediction
1.00
0.75
0.50
0.25
0.00
0
20
40
60


Groups of Ceteris paribus profiles by Sex
label
Random Forest_female
Random Forest_male
age
1.00
0.75
0.50
0.25
0.00
prediction
0
20
40
60

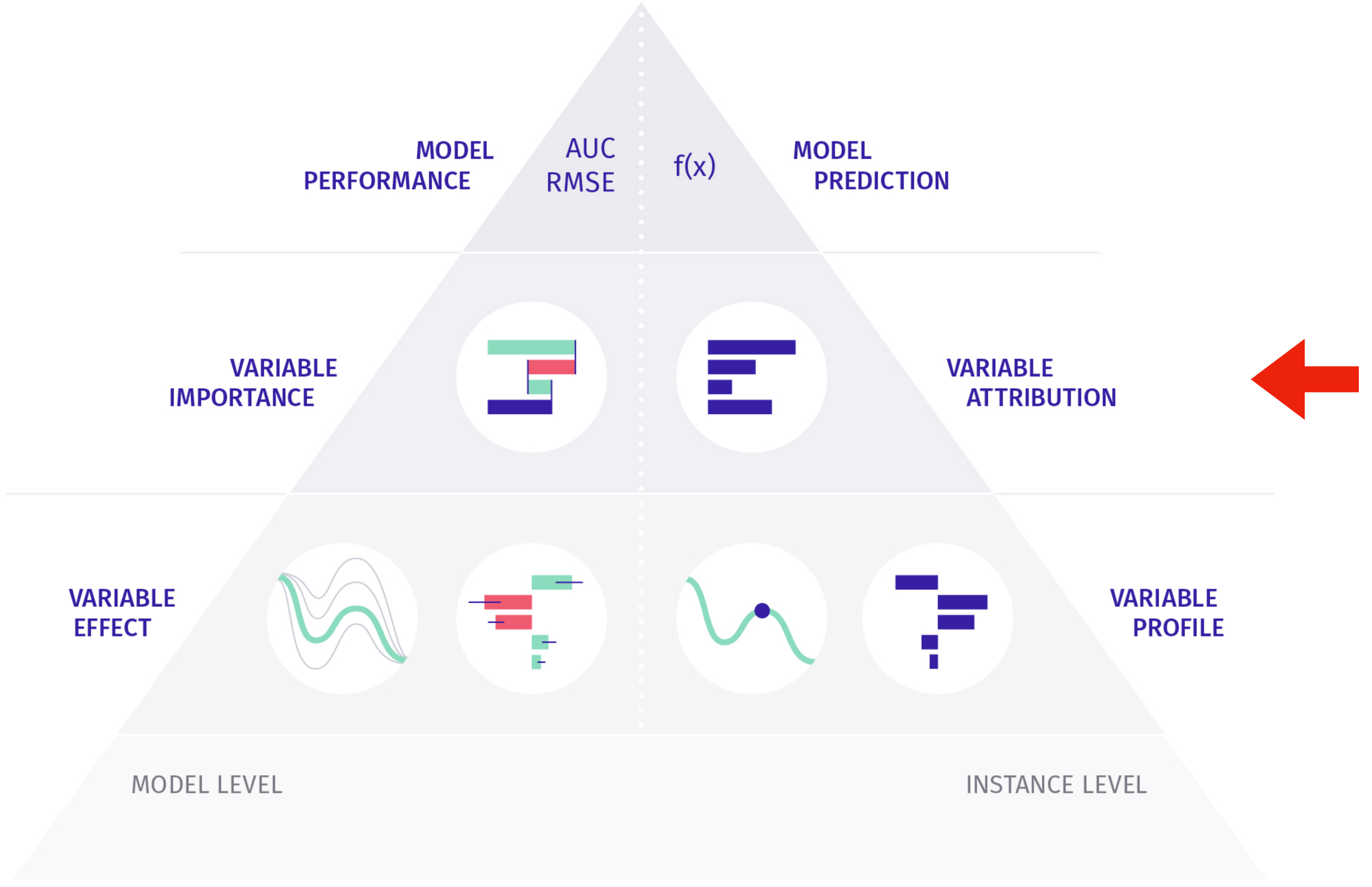

Situation 2. Channel contribution.
260000
240000
220000
y
900
1000
1100
1200
1300
1400
lidl_spend_leaflets
260000
240000
220000
y
1000
2000
3000
4000
5000
lidl_spend_newspapers
47

Day 2

Model exploration

Part 3

Instance level analysis - variable attributions


MODEL PERFORMANCE
AUC
RMSE
f(x)
MODEL PREDICTION
VARIABLE IMPORTANCE
VARIABLE ATTRIBUTION
VARIABLE EFFECT
VARIABLE PROFILE
MODEL LEVEL
INSTANCE LEVEL

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

TYTUŁ	CYTOWANE PRZEZ	ROK
A unified approach to interpreting model predictions S Lundberg, SI Lee arXiv preprint arXiv:1705.07874	2717	2017
Consistent individualized feature attribution for tree ensembles SM Lundberg, GG Erion, SI Lee arXiv preprint arXiv:1802.03888	350	2018
From local explanations to global understanding with explainable AI for trees SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, ... Nature machine intelligence 2 (1), 56-67	339	2020
Explainable machine-learning predictions for the prevention of hypoxaemia during surgery SM Lundberg, B Nair, MS Vavilala, M Horibe, MJ Eisses, T Adams, ... Nature biomedical engineering 2 (10), 749-760	240	2018
A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia SI Lee, S Celik, BA Logsdon, SM Lundberg, TJ Martins, VG Oehler, ... Nature communications 9 (1), 1-13	91	2018
Advances in neural information processing systems SM Lundberg, SI Lee, I Guyon New York: Curran Associates	91	2017
An unexpected unity among methods for interpreting model predictions S Lundberg, SI Lee arXiv preprint arXiv:1611.07478	71	2016
Explainable AI for trees: From local explanations to global understanding SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, ... arXiv preprint arXiv:1905.04610	70	2019

Definition 1 Additive feature attribution methods *have an explanation model that is a linear function of binary variables:*

Properties

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{1}$$

where $z' \in \{0,1\}^M$*,* M *is the number of simplified input features, and* $\phi_i \in \mathbb{R}$*.*

Property 1 (Local accuracy)

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{5}$$

The explanation model $g(x')$ *matches the original model* $f(x)$ *when* $x = h_x(x')$*, where* $\phi_0 = f(h_x(\mathbf{0}))$ *represents the model output with all simplified inputs toggled off (i.e. missing).*

Property 2 (Missingness)

$$x'_i = 0 \implies \phi_i = 0 \tag{6}$$

Missingness constrains features where $x'_i = 0$ *to have no attributed impact.*

The third property is *consistency*. Consistency states that if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.

Property 3 (Consistency) *Let* $f_x(z') = f(h_x(z'))$ *and* $z' \setminus i$ *denote setting* $z'_i = 0$*. For any two models* f *and* f'*, if*

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \tag{7}$$

for all inputs $z' \in \{0,1\}^M$*, then* $\phi_i(f', x) \geq \phi_i(f, x)$*.*

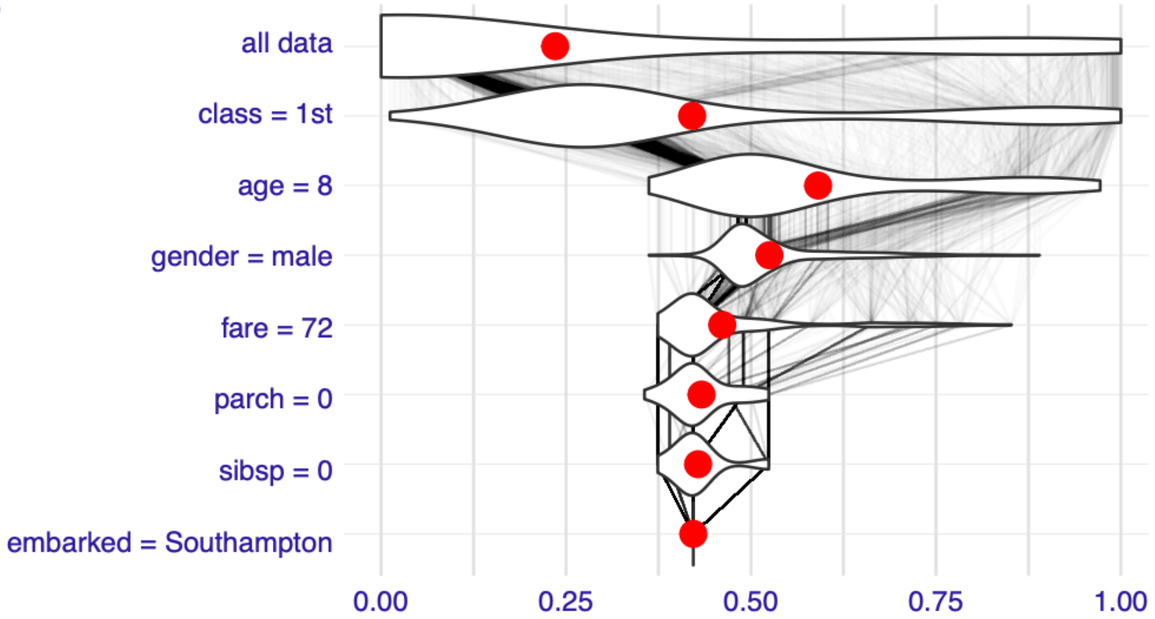
Theorem 1 *Only one possible explanation model g follows Definition 1 and satisfies Properties 1, 2, and 3:*

$$\phi_i(f, x) = \sum_{z' \subset x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{8}$$

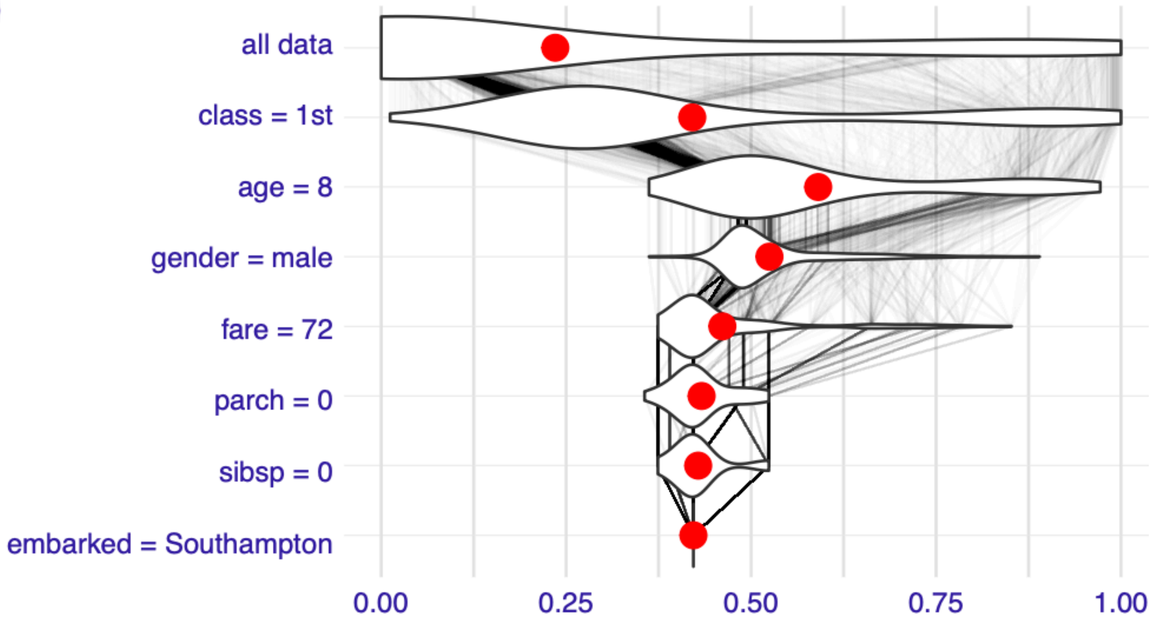
Additive shifts for non additive models

Model agnostic approach

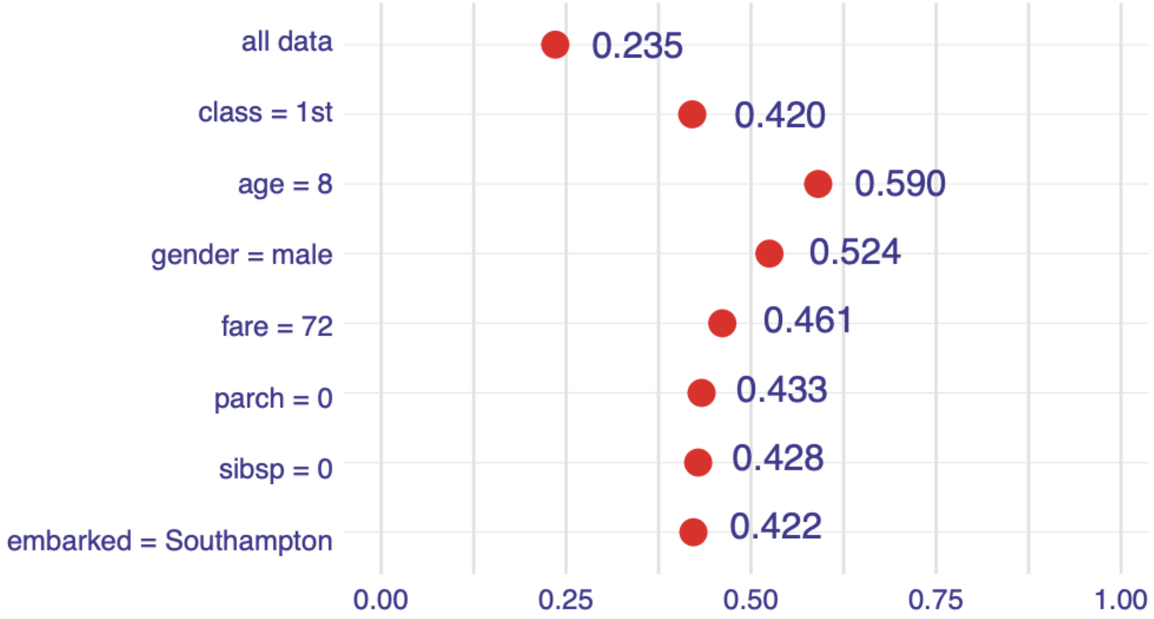
A)



A)

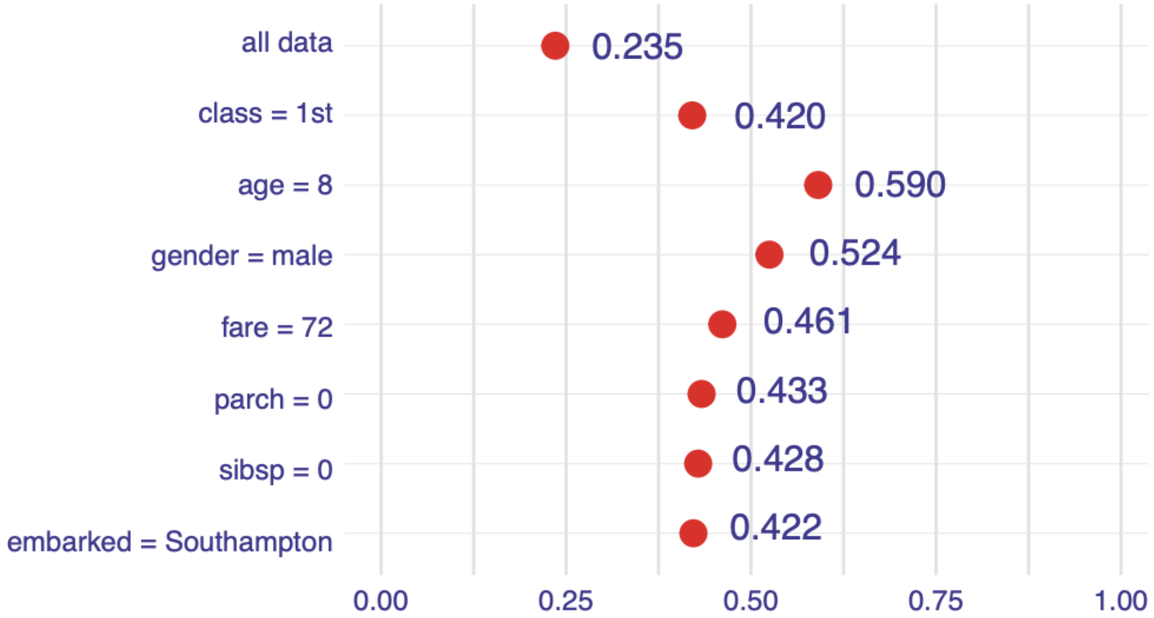

all data
class = 1st
age = 8
gender = male
fare = 72
parch = 0
sibsp = 0
embarked = Southampton
0.00
0.25
0.50
0.75
1.00

B)

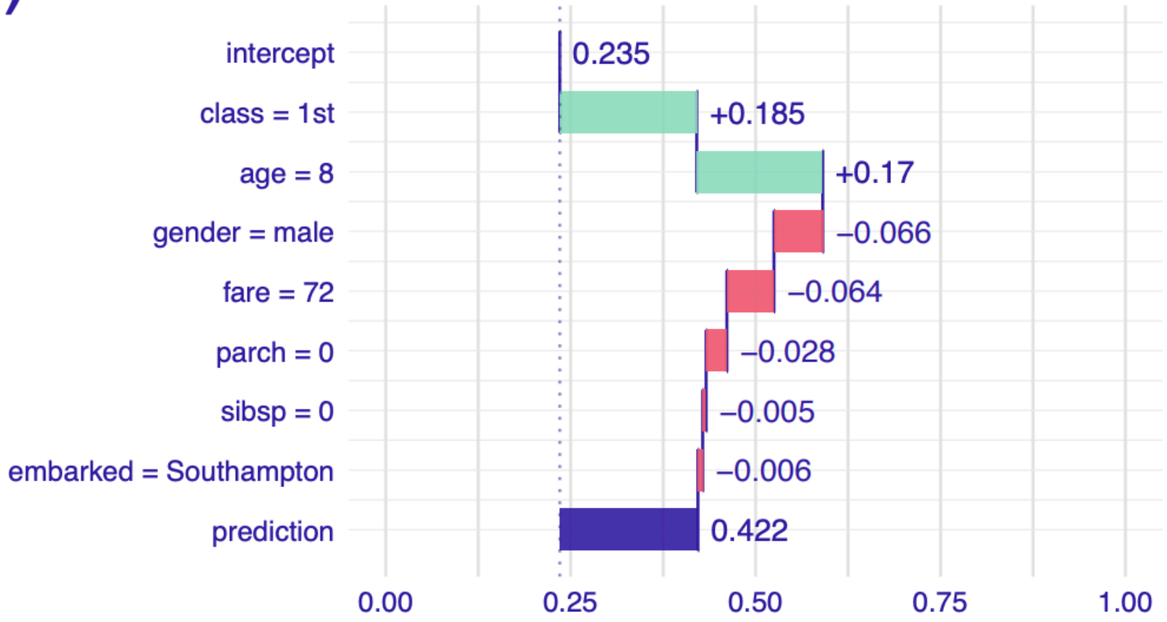

all data
class = 1st
age = 8
gender = male
fare = 72
parch = 0
sibsp = 0
embarked = Southampton
0.235
0.420
0.590
0.524
0.461
0.433
0.428
0.422
0.00
0.25
0.50
0.75
1.00

C)

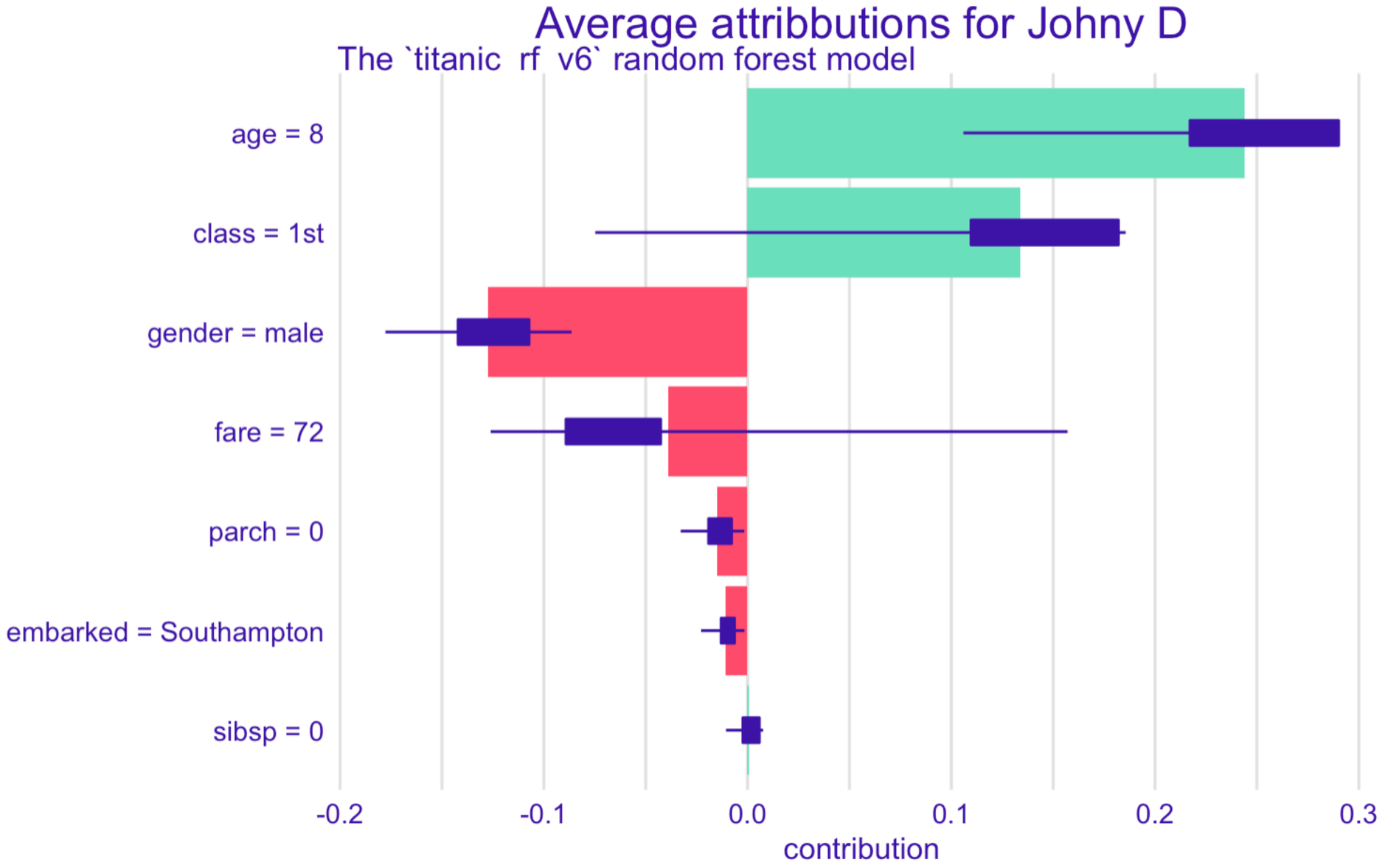
B)


all data
class = 1st
age = 8
gender = male
fare = 72
parch = 0
sibsp = 0
embarked = Southampton
0.235
0.420
0.590
0.524
0.461
0.433
0.428
0.422
0.00
0.25
0.50
0.75
1.00

C)


intercept
class = 1st
age = 8
gender = male
fare = 72
parch = 0
sibsp = 0
embarked = Southampton
prediction
0.235
+0.185
+0.17
−0.066
−0.064
−0.028
−0.005
−0.006
0.422
0.00
0.25
0.50
0.75
1.00

Take an average effect



2.1 LIME

Great unification!

To find ϕ, LIME minimizes the following objective function:

$$\xi = \arg\min \ L(f, g, \pi_{x'}) + \Omega(g). \tag{2}$$

2.2 DeepLIFT

DeepLIFT was recently proposed as a recursive prediction explanation method for deep learning [8, 7]. It attributes to each input x_i a value $C_{\Delta x_i \Delta y}$ that represents the effect of that input being set to a reference value as opposed to its original value. This means that for DeepLIFT, the mapping $x = h_x(x')$ converts binary values into the original inputs, where 1 indicates that an input takes its original value, and 0 indicates that it takes the reference value. The reference value, though chosen by the user, represents a typical uninformative background value for the feature.

DeepLIFT uses a "summation-to-delta" property that states:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta o} = \Delta o, \tag{3}$$

2.3 Layer-Wise Relevance Propagation

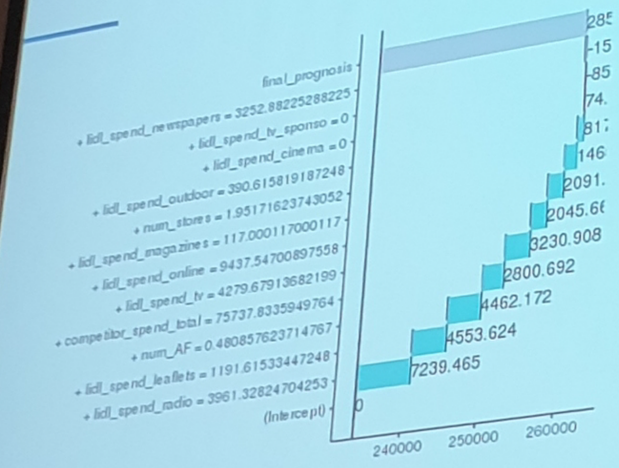
The *layer-wise relevance propagation* method interprets the predictions of deep networks [1]. As noted by Shrikumar et al., this menthod is equivalent to DeepLIFT with the reference activations of all neurons fixed to zero. Thus, $x = h_x(x')$ converts binary values into the original input space, where 1 means that an input takes its original value, and 0 means an input takes the 0 value. Layer-wise relevance propagation's explanation model, like DeepLIFT's, matches Equation 1.

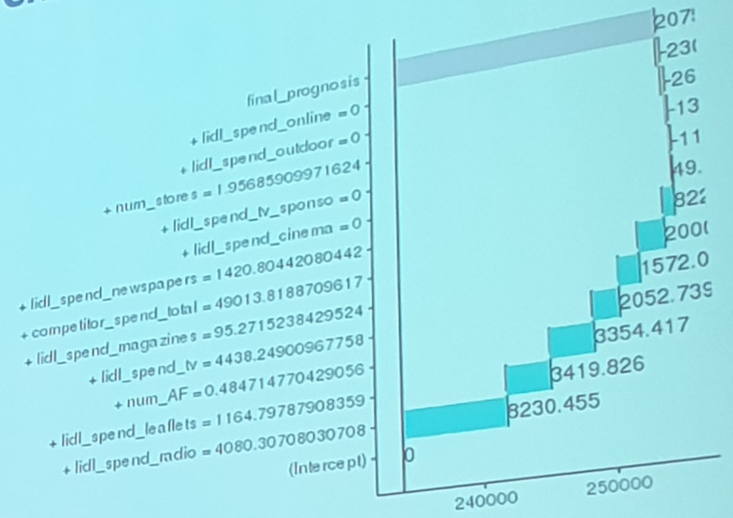
2.4 Classic Shapley Value Estimation

Three previous methods use classic equations from cooperative game theory to compute explanations of model predictions: Shapley regression values [4], Shapley sampling values [9], and Quantitative Input Influence [3].

Shapley regression values are feature importances for linear models in the presence of multicollinearity. This method requires retraining the model on all feature subsets $S \subseteq F$, where F is the set of all features. It assigns an importance value to each feature that represents the effect on the model prediction of including that feature. To compute this effect, a model $f_{S \cup \{i\}}$ is trained with that feature

Situation 2. Marketing Mix Model. Channel contribution.



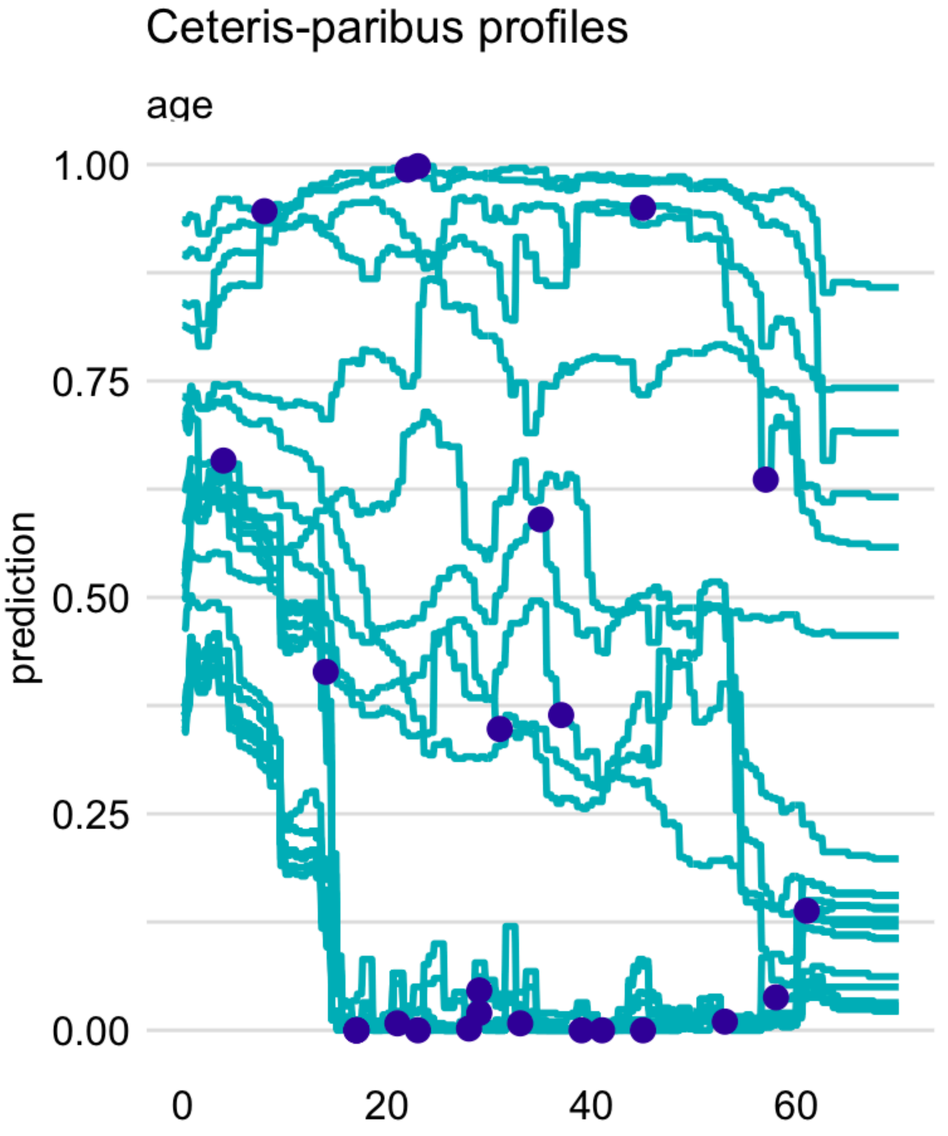


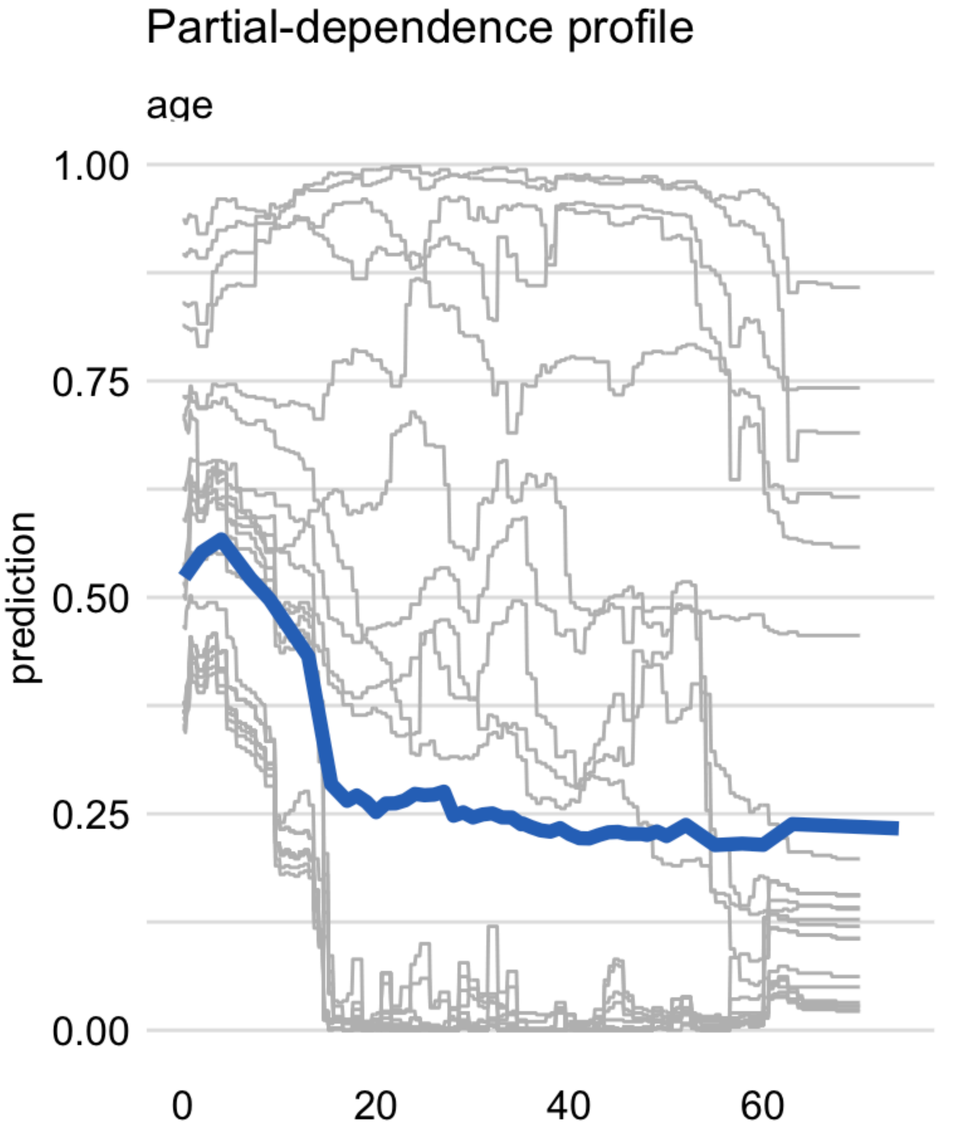
Day 2

Model exploration

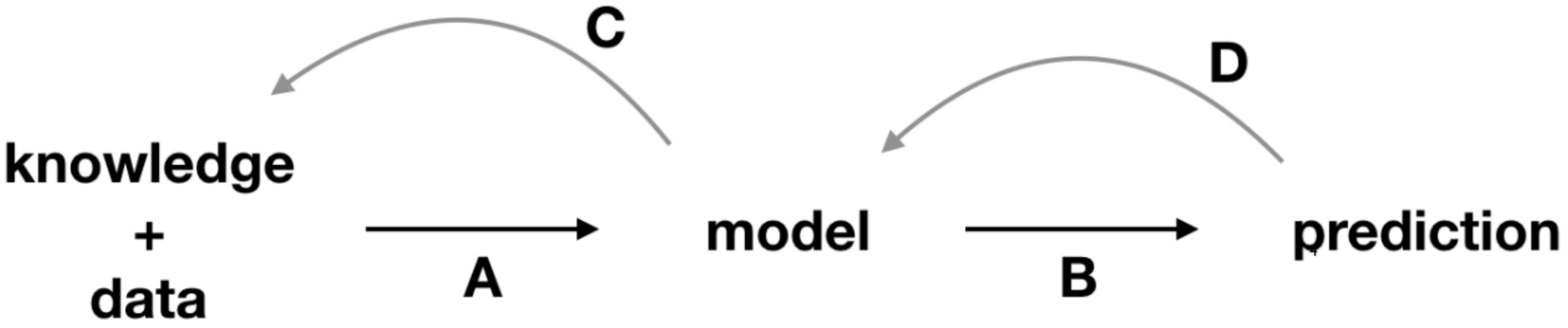
Part 4

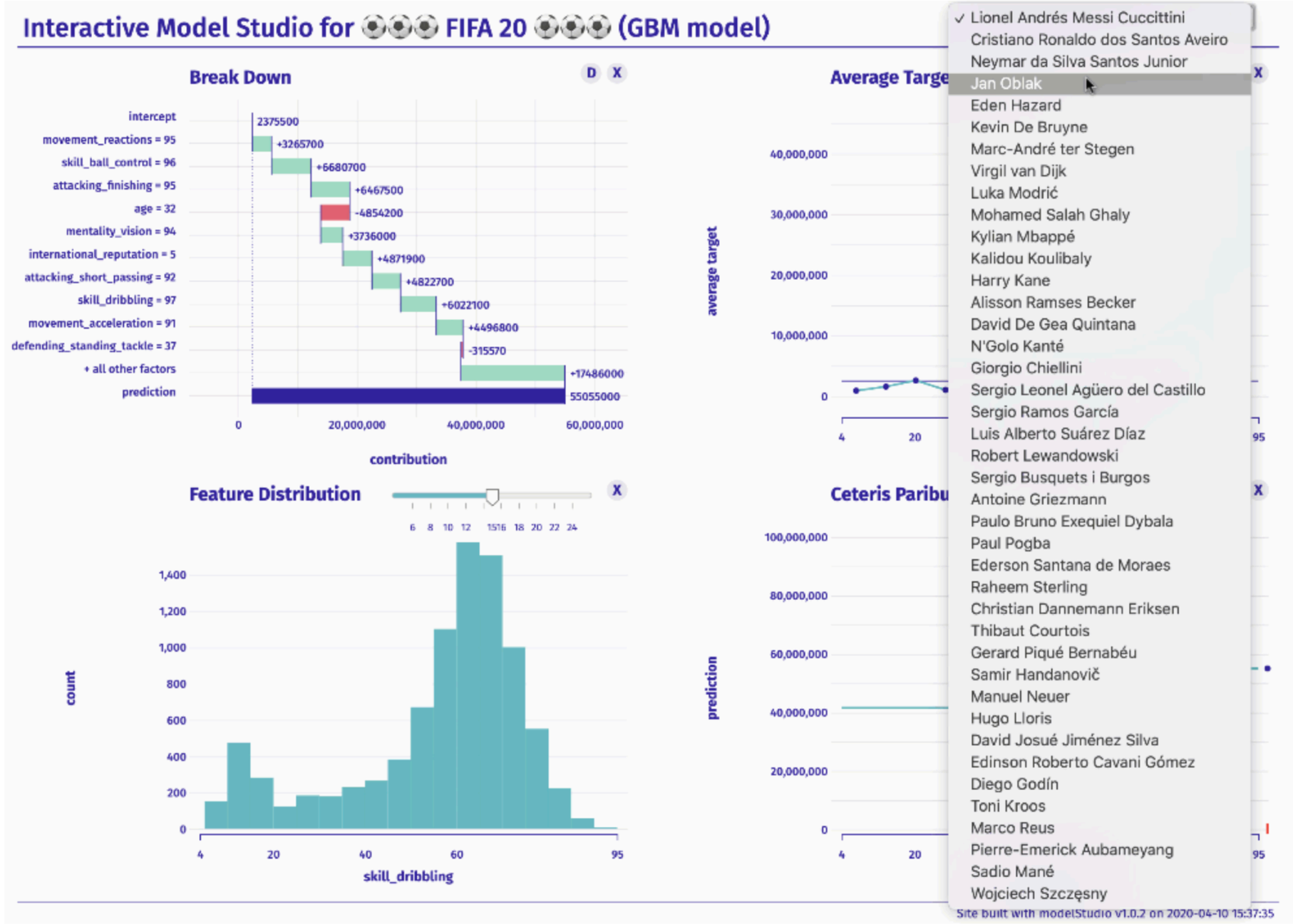
Instance level analysis - variable profile





http://ema.drwhy.ai/partialDependenceProfiles.html


C
D
knowledge
+
data
A
model
B
prediction



http://modelstudio.drwhy.ai/

Introduction to Responsible Machine Learning

with mlr3 and DALEX

Przemyslaw Biecek

17-18.04.2021

https://www.linkedin.com/in/pbiecek/