And why did you make this prediction, machine?

Sergey Yurgenson,
DataRobot

Kyiv, April 2017

DataRobot

Agenda

- Why do we need model interpretability?
- How can we achieve (some) model interpretability?

Why do we need interpretability?

- Subjective (Natural human suspicious of

anything new, untested.)

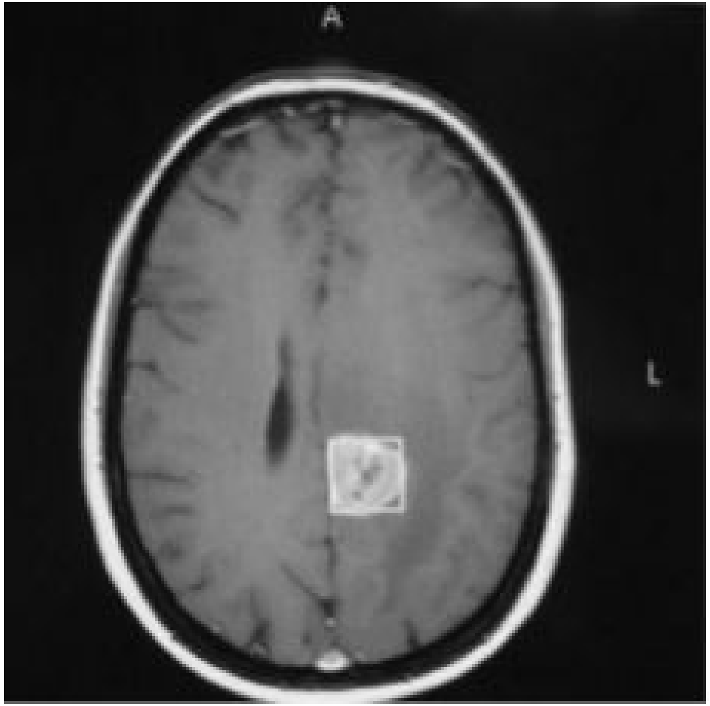
- Objective (medical field, criminal justice system, military...)
- Regulatory or legal requirements

Why do we need interpretability?

- Objective
    - Model has only advisory role, final decision is made by human
    - To improve business process (if we understand model inner-working we may implement some of findings as simple business rules)
    - Control for bias / discrimination
    - Knowledge extraction

Why do we need interpretability?

- Advisory role



Where is the tumor specifically?

Why do we need interpretability?

- To improve business process (if we understand model inner-working we may implement some of findings as business rules)
- Knowledge extraction

> “At the same time, Deep Patient is a bit puzzling. It appears to anticipate the onset of psychiatric disorders like schizophrenia surprisingly well. But since schizophrenia is notoriously difficult for physicians to predict, Dudley wondered how this was possible. He still doesn’t know. The new tool offers no clue as to how it does this. If something like Deep Patient is actually going to help doctors, it will ideally give them the rationale for its prediction, to reassure them that it is accurate and to justify, say, a change in the drugs someone is being prescribed. “We can build these models,” Dudley says ruefully, “but we don’t know how they work.””

MIT Technology Review

Intelligent Machines

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by Will Knight April 11, 2017

Why do we need interpretability?

- Control for bias / discrimination

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ProPublica

Why do we need interpretability?

- Regulatory or legal requirements
  - European Parliament adopted a set of comprehensive regulations for the collection, storage and use of personal information, the **General Data Protection Regulation**.
  - These regulations are planned to become effective in 2018.

Why do we need interpretability?

Article 22. Automated individual decision making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

   (c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Approaches

- Not that easy
  - “Although computer scientists are working on it, (Geoffrey) Hinton acknowledged that the challenge of opening the black box, of trying to find out exactly what these powerful learning systems know and how they know it, was “far from trivial—don’t believe anyone who says that it is.” “

THE NEW YORKER

ANNALS OF MEDICINE APRIL 3, 2017 ISSUE

A.I. VERSUS M.D.

What happens when diagnosis is automated?

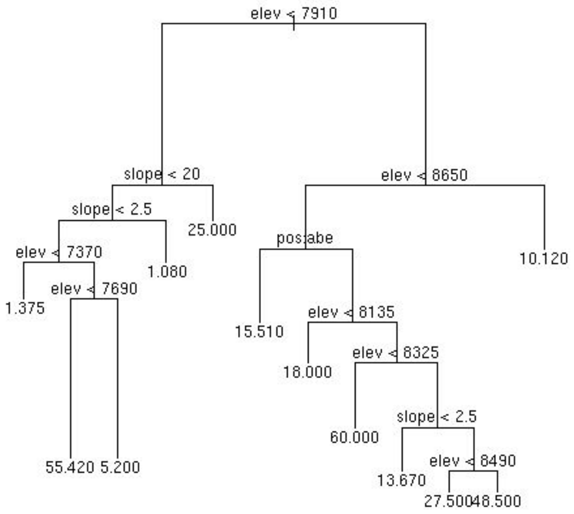
By Siddhartha Mukherjee

Approaches

- Some Approaches
  - Explainable models
  - Surrogate model
  - Local explanation
  - Compare with “typical result”

Approaches

- “Explainable” models
  - Decision trees
  - GLM
  - ...



$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

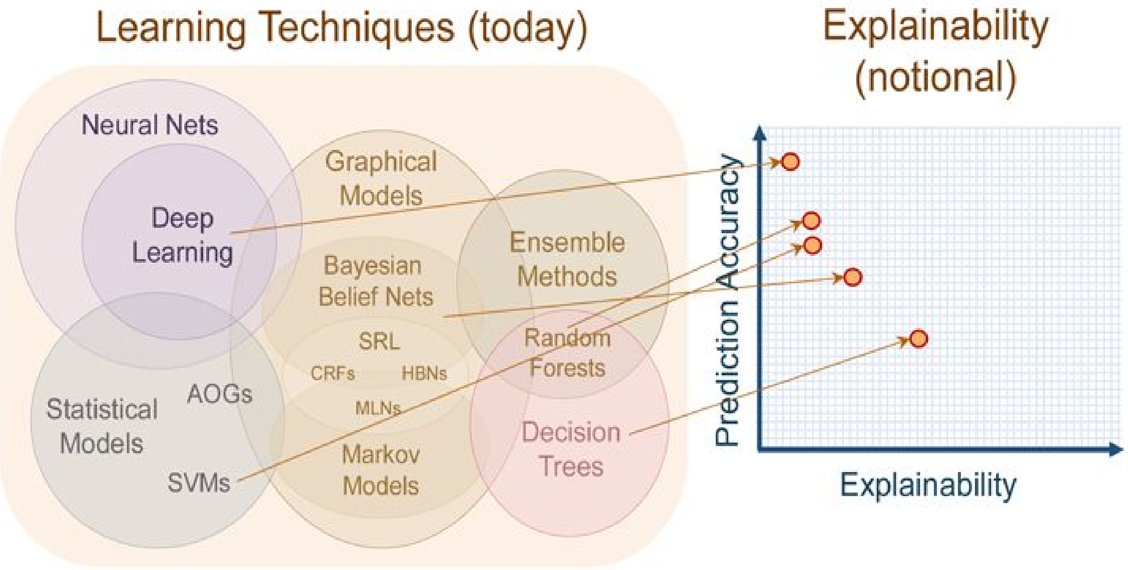
Approaches

Is Artificial Intelligence Permanently Inscrutable?

Despite new biology-like tools, some insist interpretation is impossible.

BY AARON M. BORNSTEIN
ILLUSTRATION BY EMMANUEL POLANCO
SEPTEMBER 1, 2016

NAUTILUS



WHAT VS. WHY: Modern learning algorithms show a tradeoff between human interpretability, or explainability, and their accuracy. Deep learning is both the most accurate and the least interpretable.

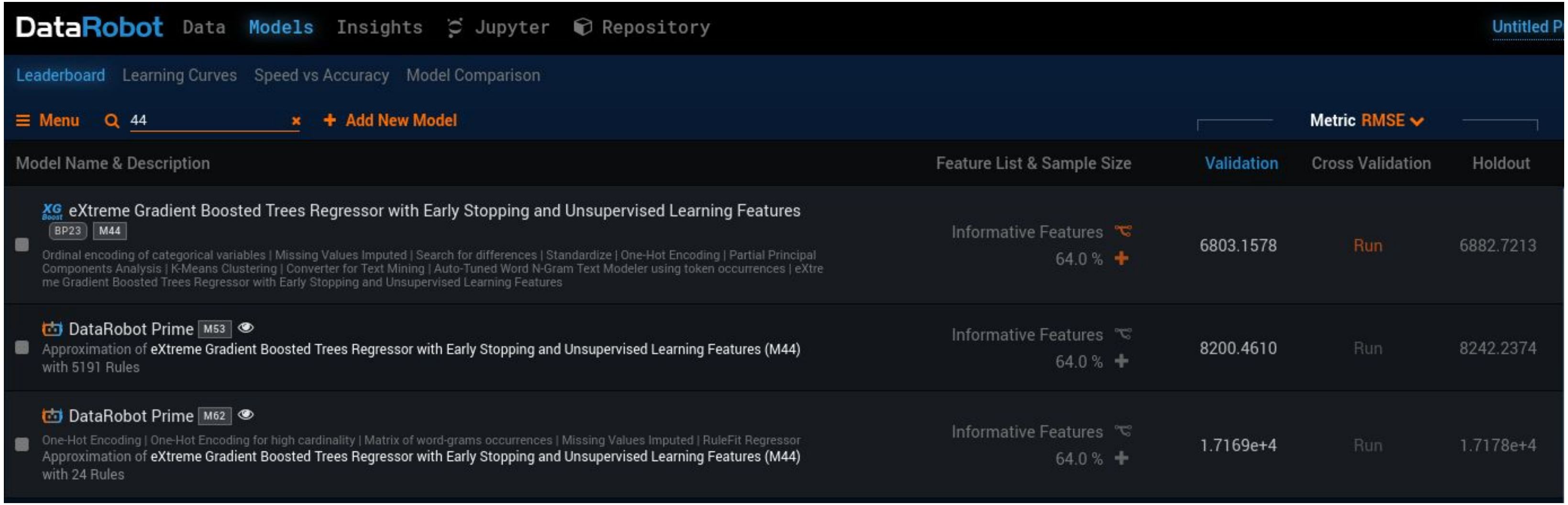
Darpa

Approaches

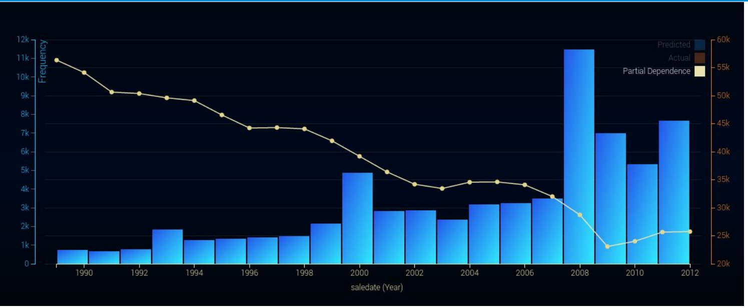
- Surrogate model
  - Create main model
  - Train second (“explainable”) model using predictions of the first model as targets

Approaches

- Surrogate models



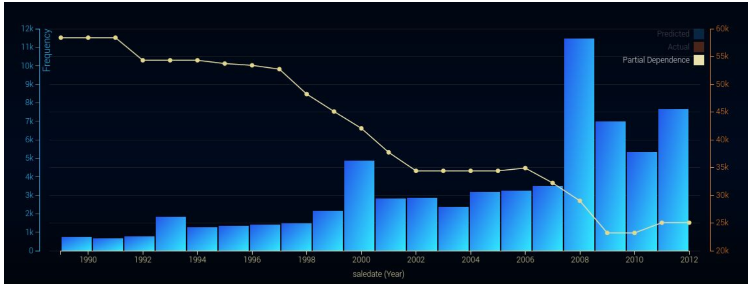
Approaches



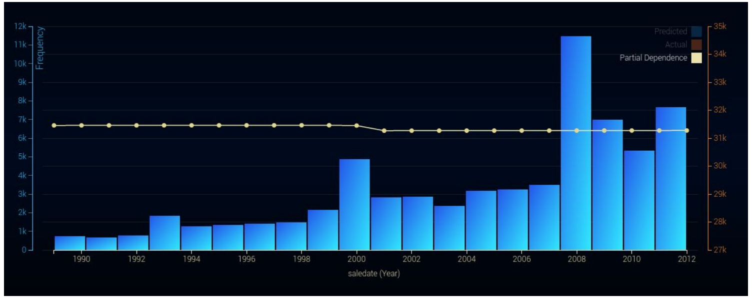
- Main model. Partial dependence plot

- Surrogate models

- 5191 rules

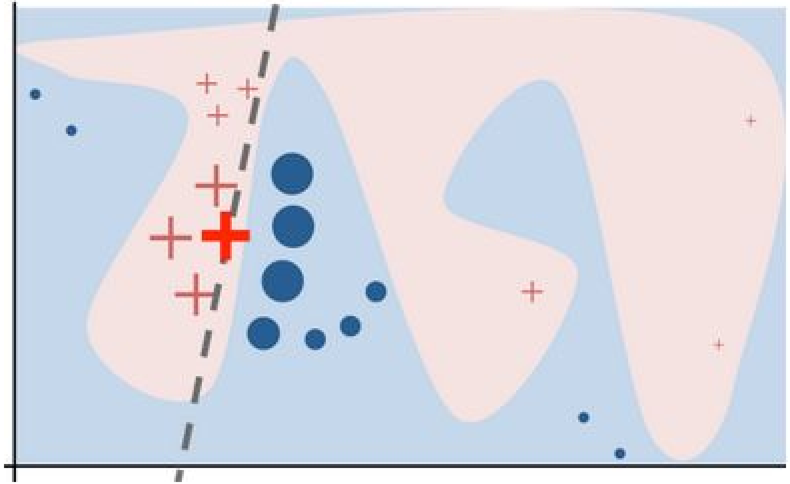


- 24 rules



Approaches

- Local explanation (LIME [Local Interpretable Model-Agnostic Explanation]
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin)



Approaches

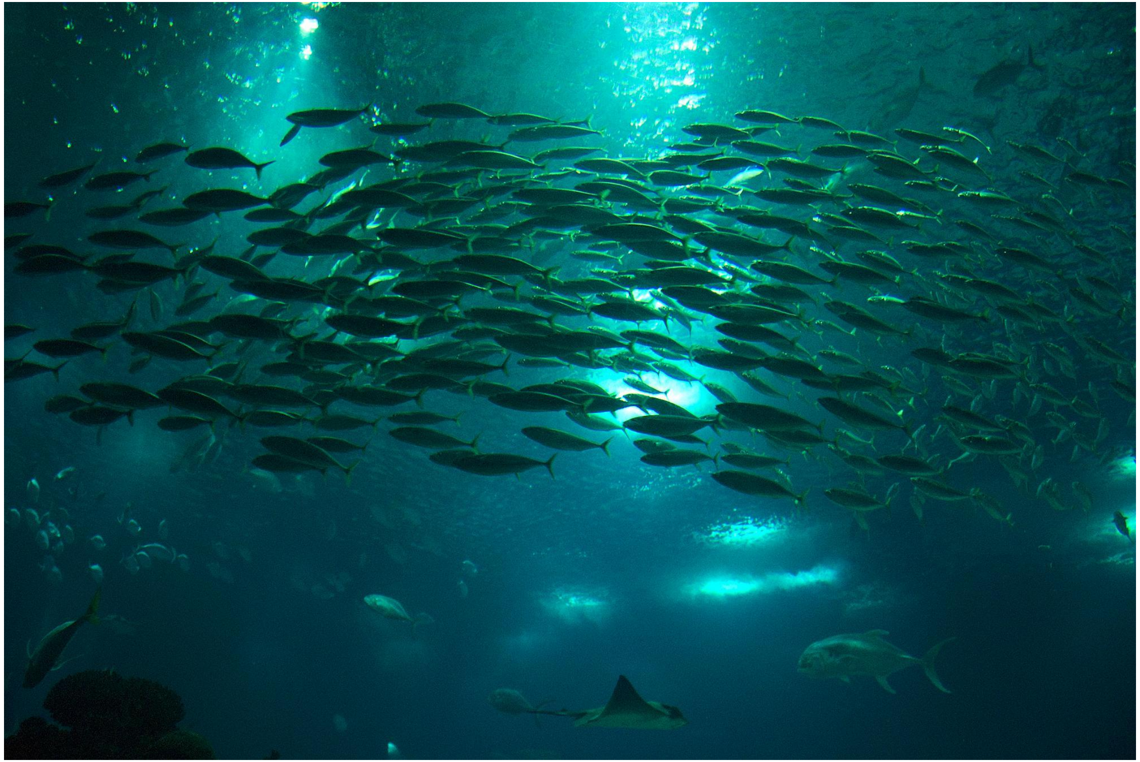


Photo by Michal Osmenda

Approaches

- Approach
  - Compare a prediction with “typical result”
  - Assume the difference can be explained by difference of predictors from “typical predictors”

Approaches

- Define what is a “typical” ?
  - The whole population
  - Middle 50% (25% to 75% quantiles)
  - Manually specified subpopulation
  - ...
- Find what would be the prediction when a predictor have “typical values”?
- Evaluate how difference of each predictor from “typical values” affect prediction
- Select strongest factors and communicate to the user

Approaches

- House price = f (size, N_bedrooms, N_bathrooms, Territory,...)
- Calculate “effect” of each variable

size	N_bedrooms	N_bathrooms	Territory
1600	3	2	5

- Typical values : Numerical – random sample
- Typical values :Categorical – each unique value with appropriate weights

Territory						
1	2	3	4	5	6	7
20	50	130	500	100	100	100
0.02	0.05	0.13	0.5	0.1	0.1	0.1

Approaches

size	N_bedrooms	N_bathrooms	Territory
1600	3	2	5

size	N_bedrooms	N_bathrooms	Territory	Prediction
1600	3	2	1	185000
1600	3	2	2	220000
1600	3	2	3	190000
1600	3	2	4	270000
1600	3	2	5	200000
1600	3	2	6	195000
1600	3	2	7	235000

Approaches

Territory	Prediction	Weight
1	1850	0.02
2	2200	0.05
3	1900	0.13
4	2700	0.5
5	2000	0.1
6	1950	0.1
7	2350	0.1

- Average prediction for "typical" value of territory = 1850*0.02 + 2200*0.05 + 1900*0.13 + 2700*0.5 + 2000*0.1 + 1950*0.1 + 2350*0.1 = 2374
- Effect of territory = 2000 - 2374 = - 374

Approaches


Advanced AVG Blender
M17+18+23+26+27+...
M66
Informative Features
63.97 %
0.8686
0.8801
0.8487
Blueprint
Lift Chart
Model X-Ray
Feature Impact
Model Info
Model Log
ROC Curve
Coefficients
Word Cloud
DataRobot Prime
Deploy Model
Predict
Reason Codes
Blueprint JSON
Prediction App
Download
Get top 3 reasons for low high predictions:
COMPUTE & DOWNLOAD
0.0
0.206
0.5
0.771
1.0
Validation Predictions
Preview of reasons for predictions from validation data within highlighted area(s):
ID
PREDICTION
REASONS
310 0.937 +++ Sex = "female" +++ Pclass = 1 ++ Embarked = "C"
215 0.924 +++ Pclass = 1 +++ Sex = "female" ++ Name = "Newell, Miss....
337 0.923 +++ Pclass = 1 +++ Sex = "female" ++ Fare = 134.5
94 0.065 --- Sex = "male" -- Name = "Coxon, Mr.... -- Age = 59

Approaches

- Approach is model-agnostic
- Explanations are model specific
- Reflects model inner-working
  - Does not take into account variable interactions
  - Multicollinear features may be treated differently depending on model

Questions

Questions?

Thank you