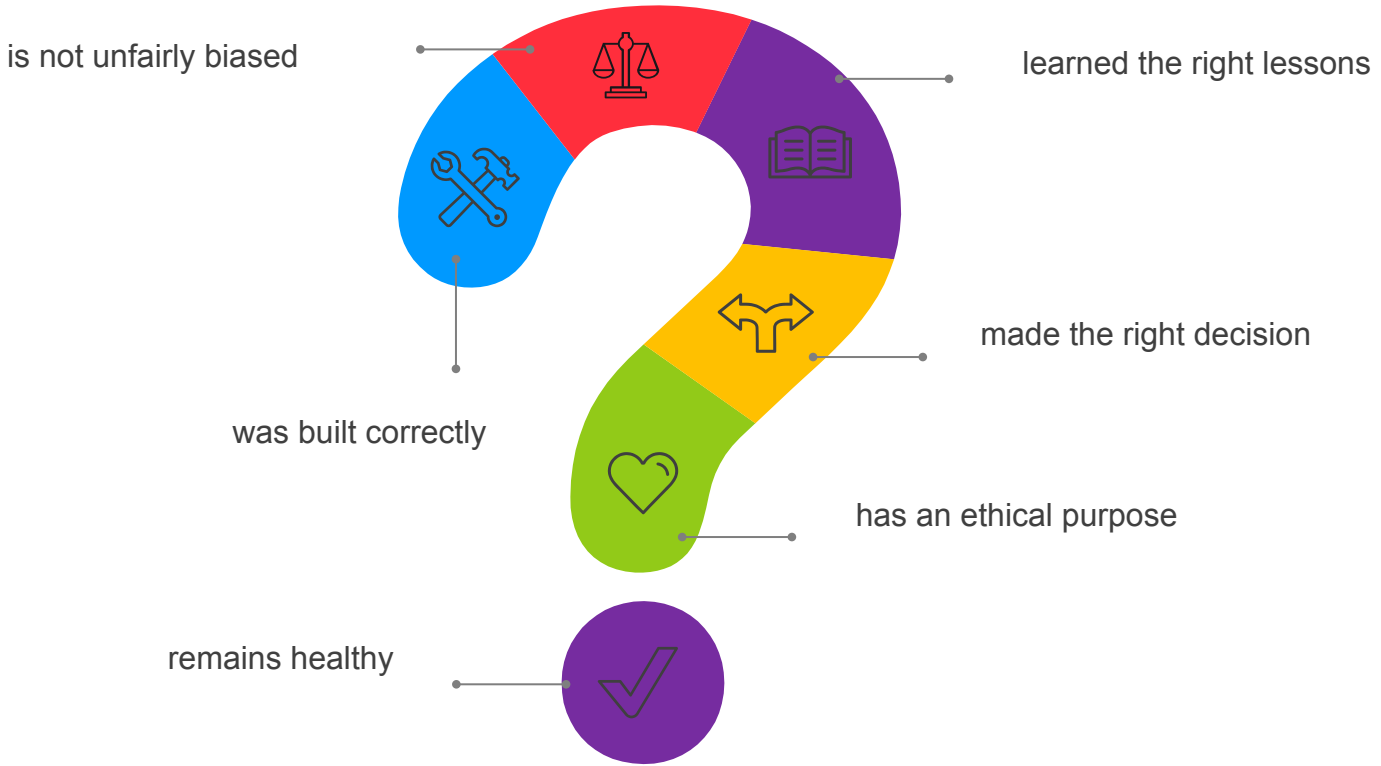
Rajiv Shah

A Quest for Interpretability

https://bit.ly/inter_workshop


Predictive Model Around Aggression

Trust: The big picture



Today, We are focusing on Just a SMall Part


Interpretable Predictive Model Around Aggression

Why Interpretability?

Why?

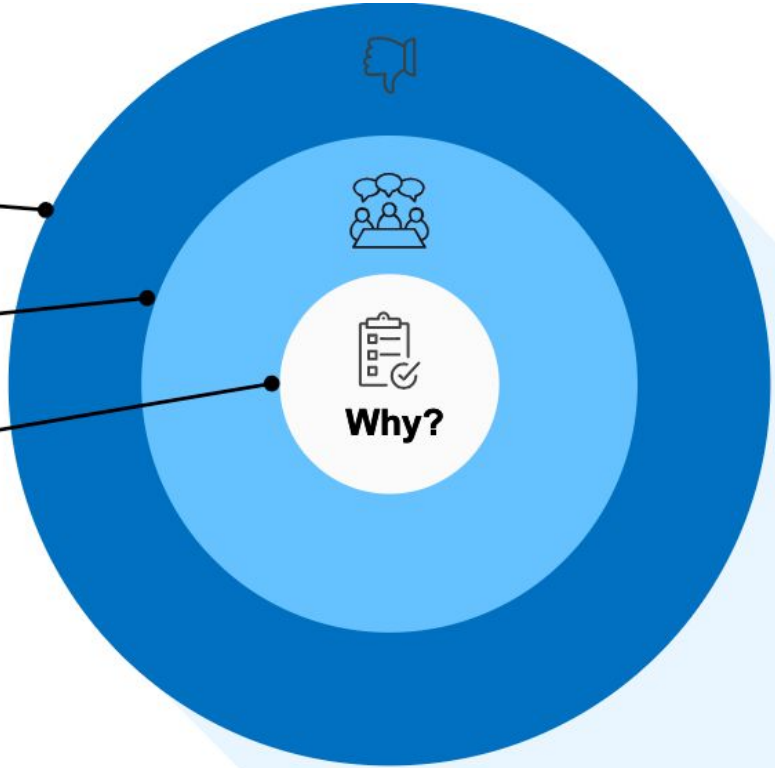
It's easy for things to go wrong

Why?

You need buy-in from human staff

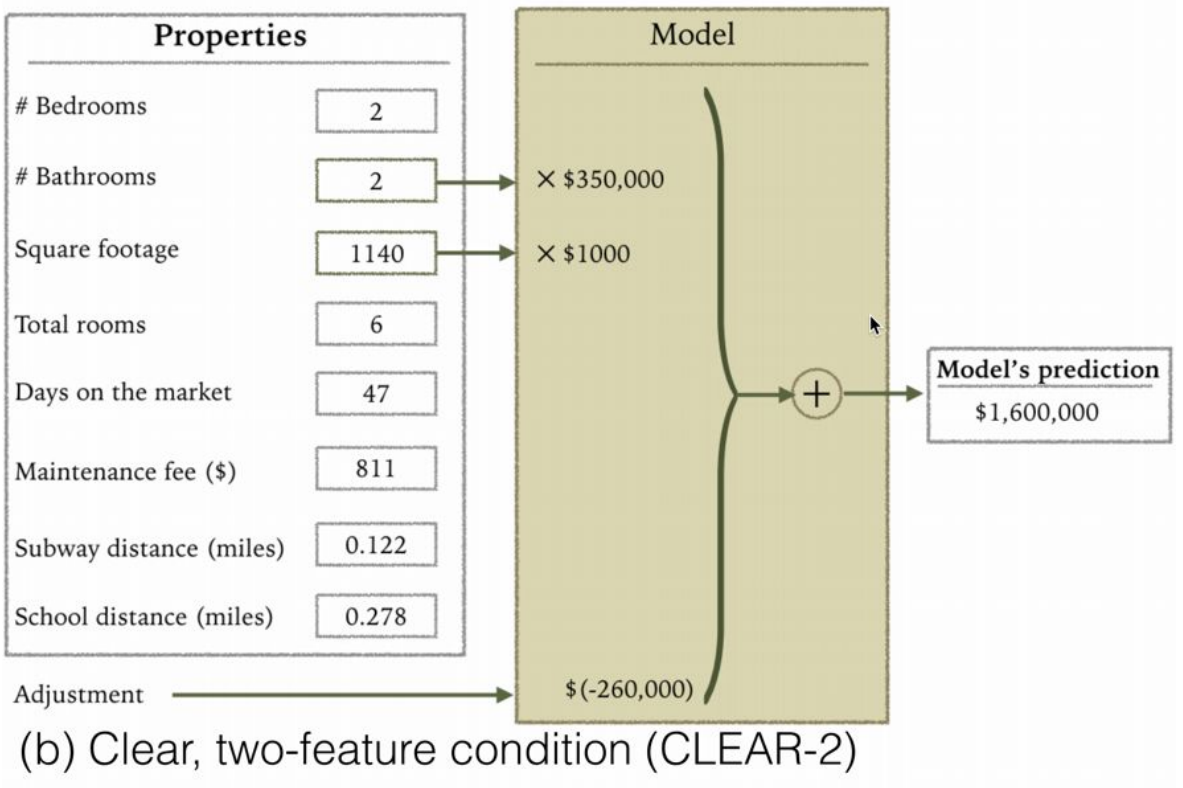
Why?

You need buy-in from regulators



An Understandable White Box Model

All the features and calculations are exposed

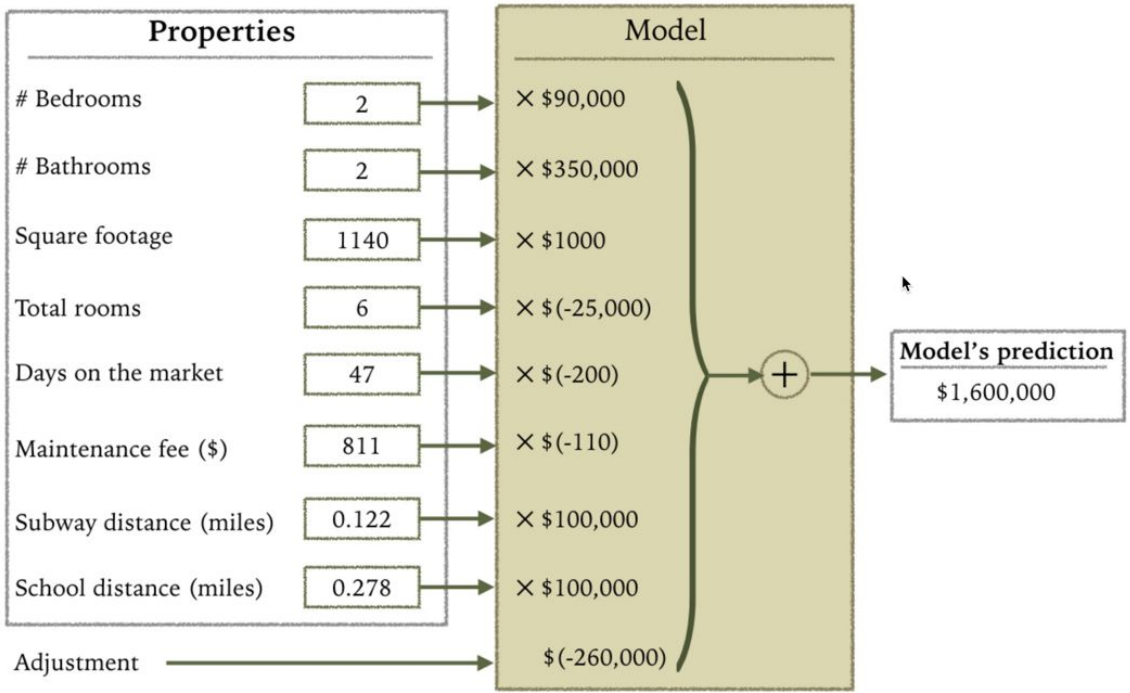


(b) Clear, two-feature condition (CLEAR-2)

Source: [Poursabzi-Sangdeh 2017]

An Understandable White Box Model? #$@&%*!

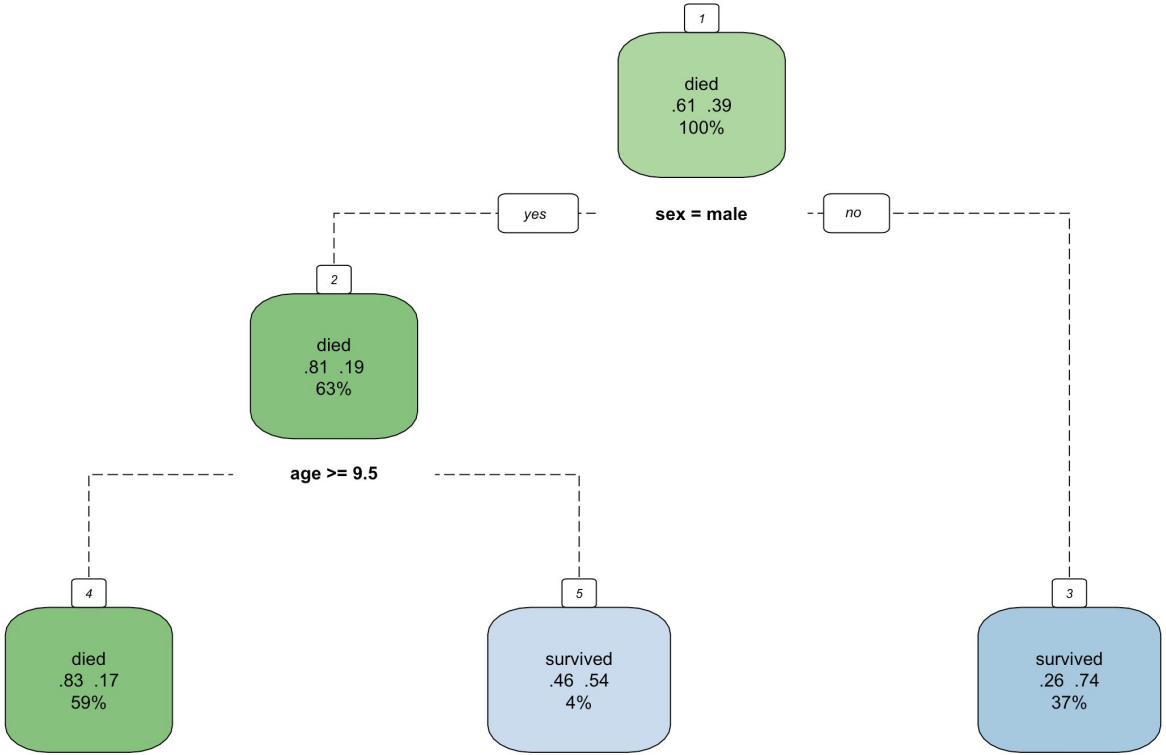
More features and correlated features make it difficult to understand



(d) Clear, eight-feature condition (CLEAR-8)

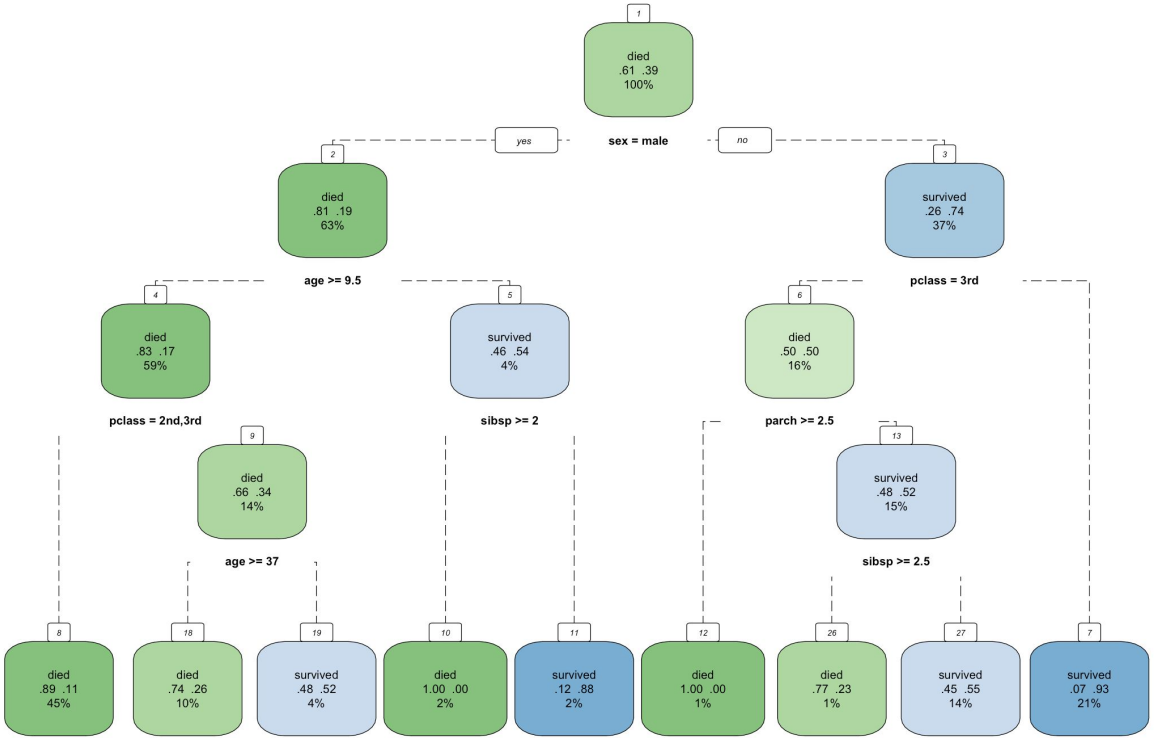
Source: Poursabzi-Sangdeh 2017

An Understandable White Box Model



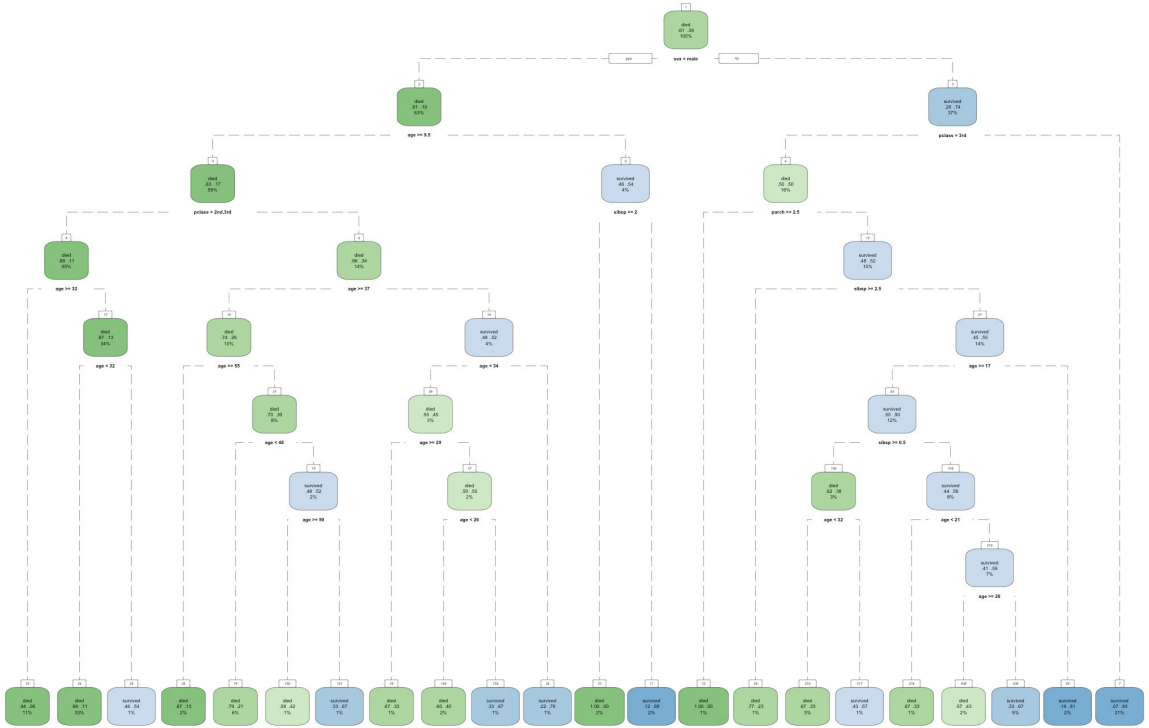
AUC = 0.74

An Understandable White Box Model?



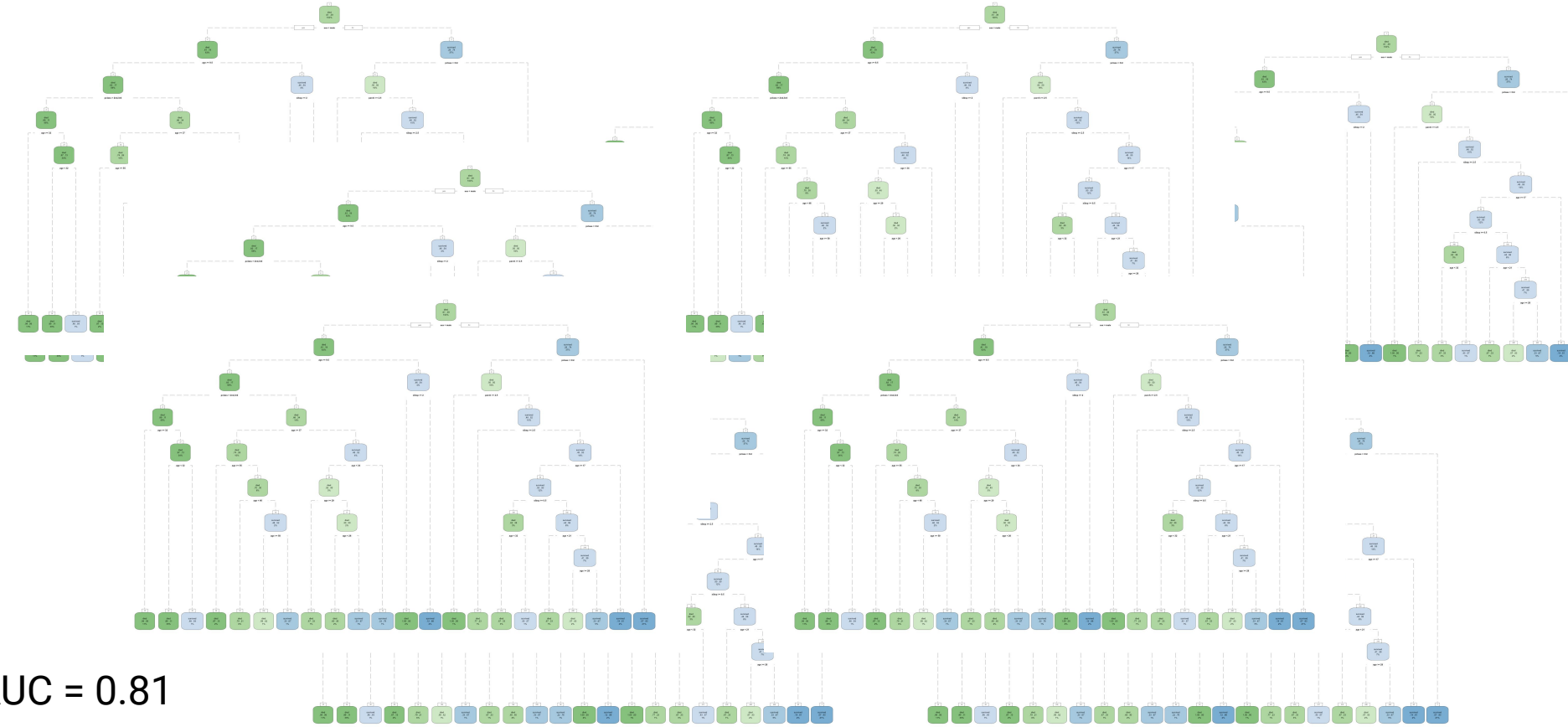
AUC = 0.78

An Understandable White Box Model? #$@&%*!



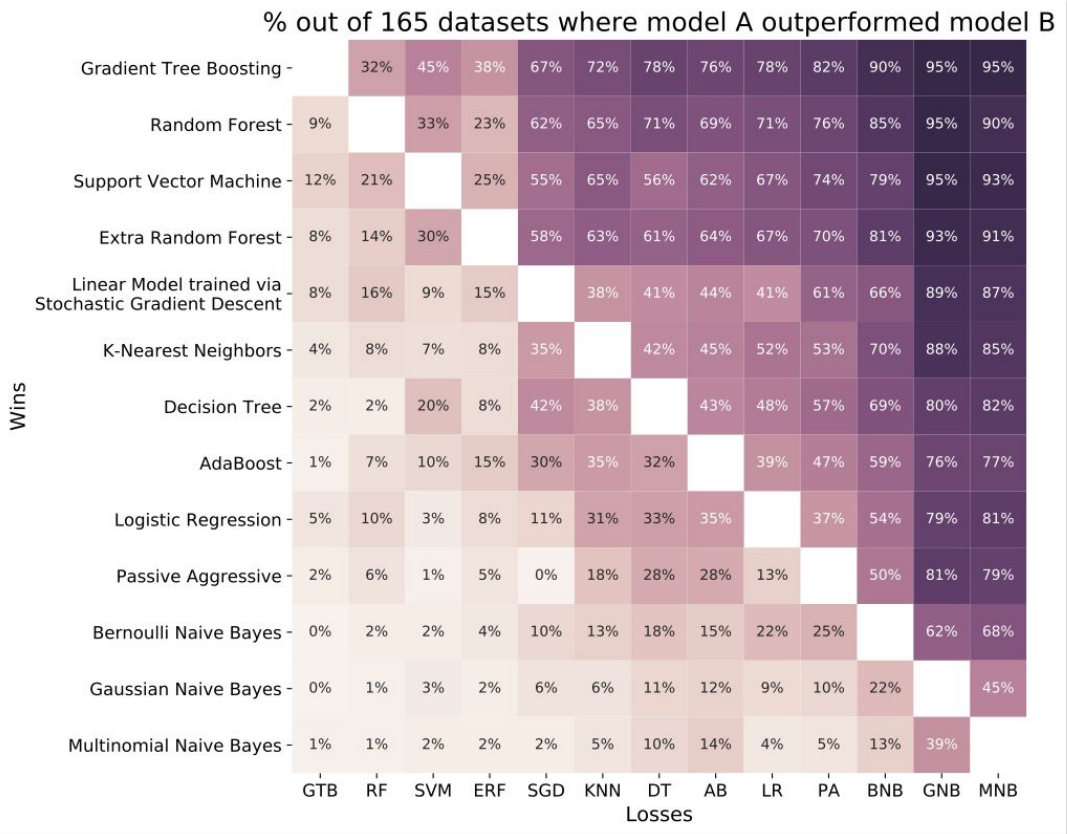
AUC = 0.79

Better Performance but too much to Comprehend



AUC = 0.81

There are so many algorithms to try



Source: [Olson 2018](#)
[Penn ML Benchmarks](#)

Algorithms matter

If the model is inaccurate, we are toast

Simple models != Accurate

Only very very simple models are human understandable

Further study:

- Interpretability in models with multicollinearity: Brieman
- Limits of human understanding: Poursabzi
- Simple models are unfair: Kleinberg
- In defense of the black box: Holm

There are tools that can explain **any** black box model

Model Agnostic Explanation Tools

Most impactful features - Feature importance

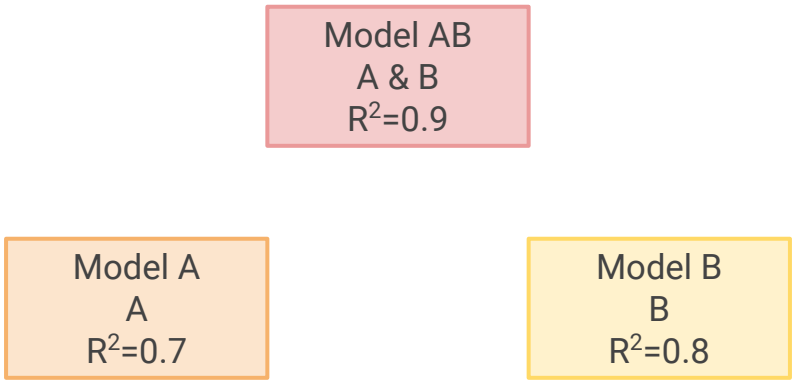
Directionality of the feature - Partial dependence

Explain a prediction - Explanation techniques (LIME, XEMP, SHAP . . .)

Feature Importance

Age
Weight
Gender
Color
Breath Fire
# of Kills
Winged
# of Heads
Spiked tail
Demeanor
Children

Feature Importance



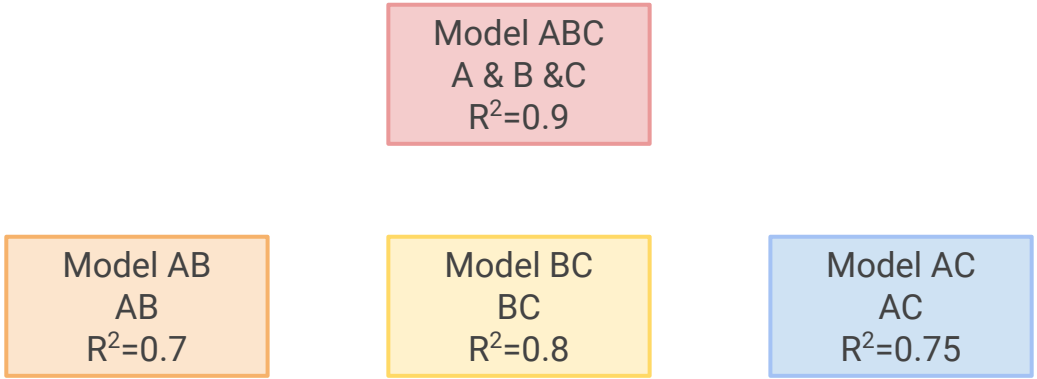
Build 3 different models based on different sets of features

Feature B is more important to the model

Ablation Methodology

Compare performance with and without the features

‘Leave it Out’ Feature Importance

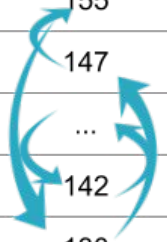


Build 4 different models based on ‘Leave it Out’ importance

Feature C is more important to the model

Permutation based Feature Importance

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24



Shuffle the feature (permute) which removes the signal within the same model

Source: Breiman 2001

Feature Impact Ranking:

1. # of Kills
2. # of Heads
3. Children
4. Age
5. Weight
6. Demeanor
7. Gender
8. Breath Fire
9. Color
10. Spiked tail
11. Winged

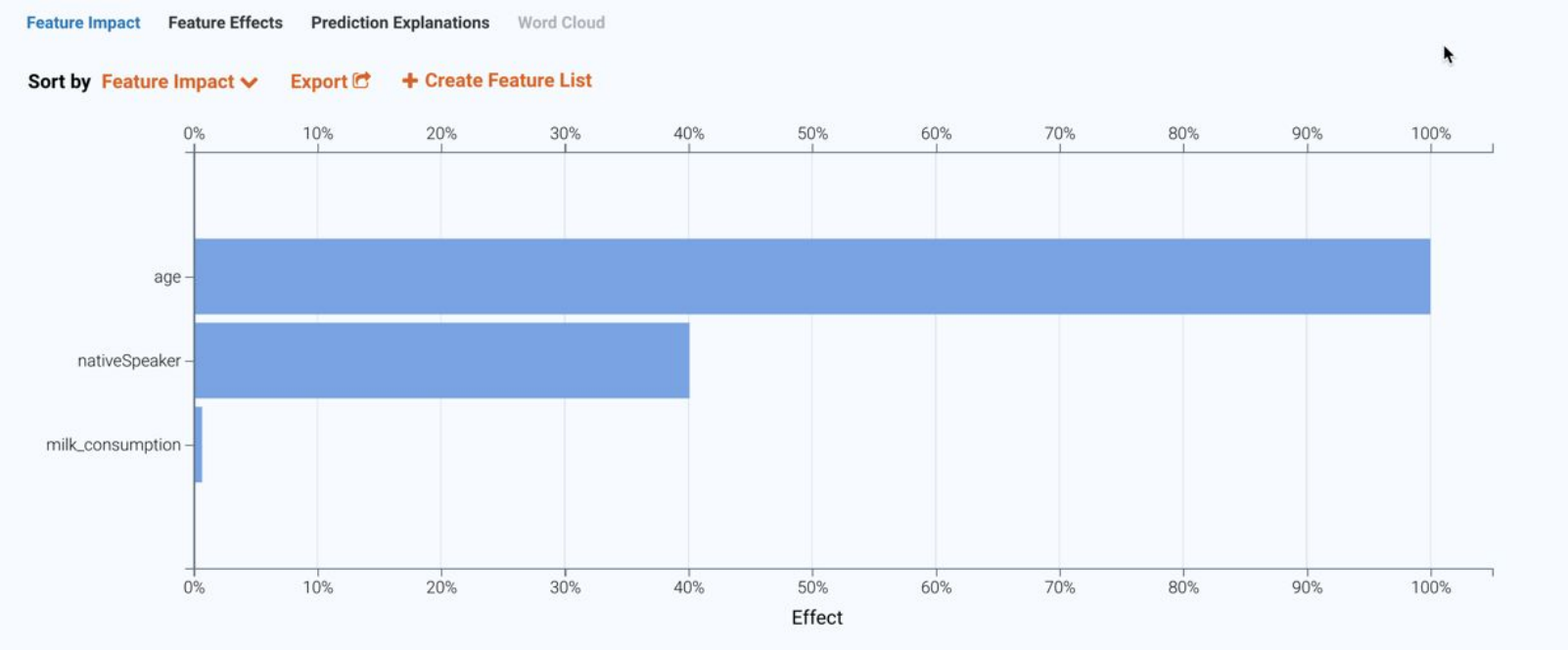
Feature impact has consequences, so you better get it right

What affects Reading?

Age

Milk Consumption

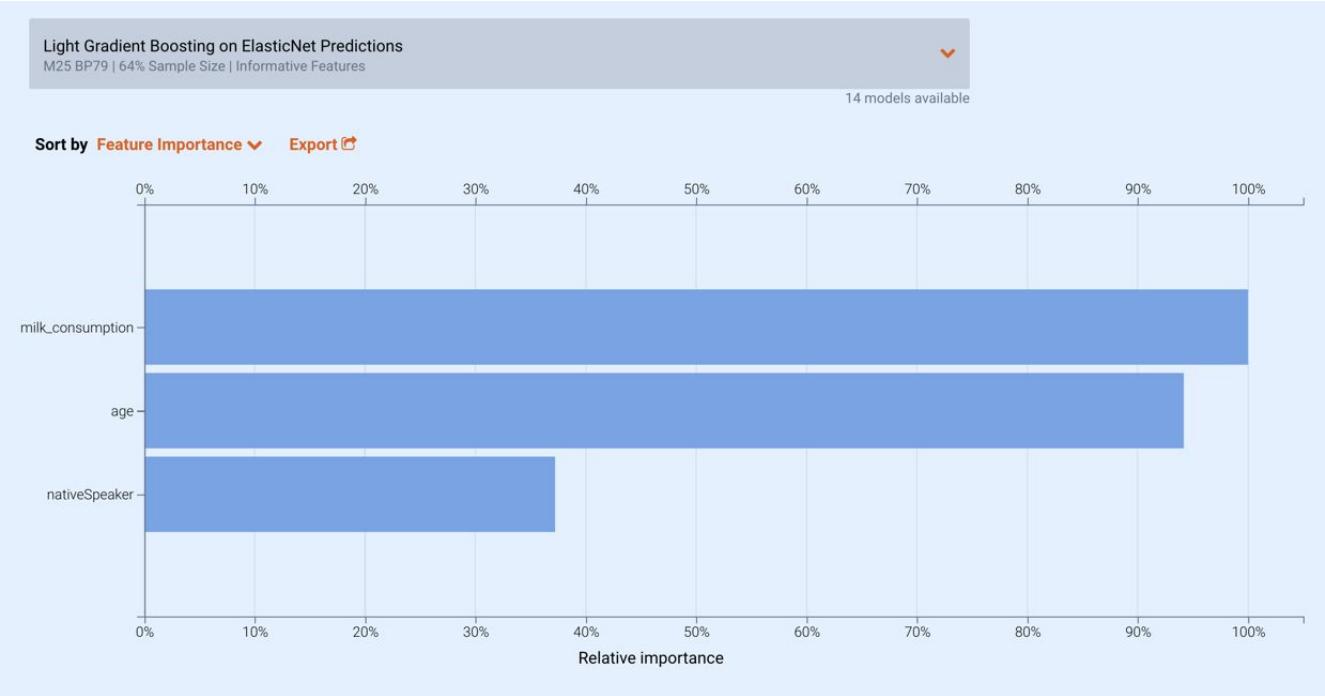
Permutation Based Variable Importance



PERMUTATION RECOGNIZES THAT AGE AFFECTS READING

Source: Strobl 2009

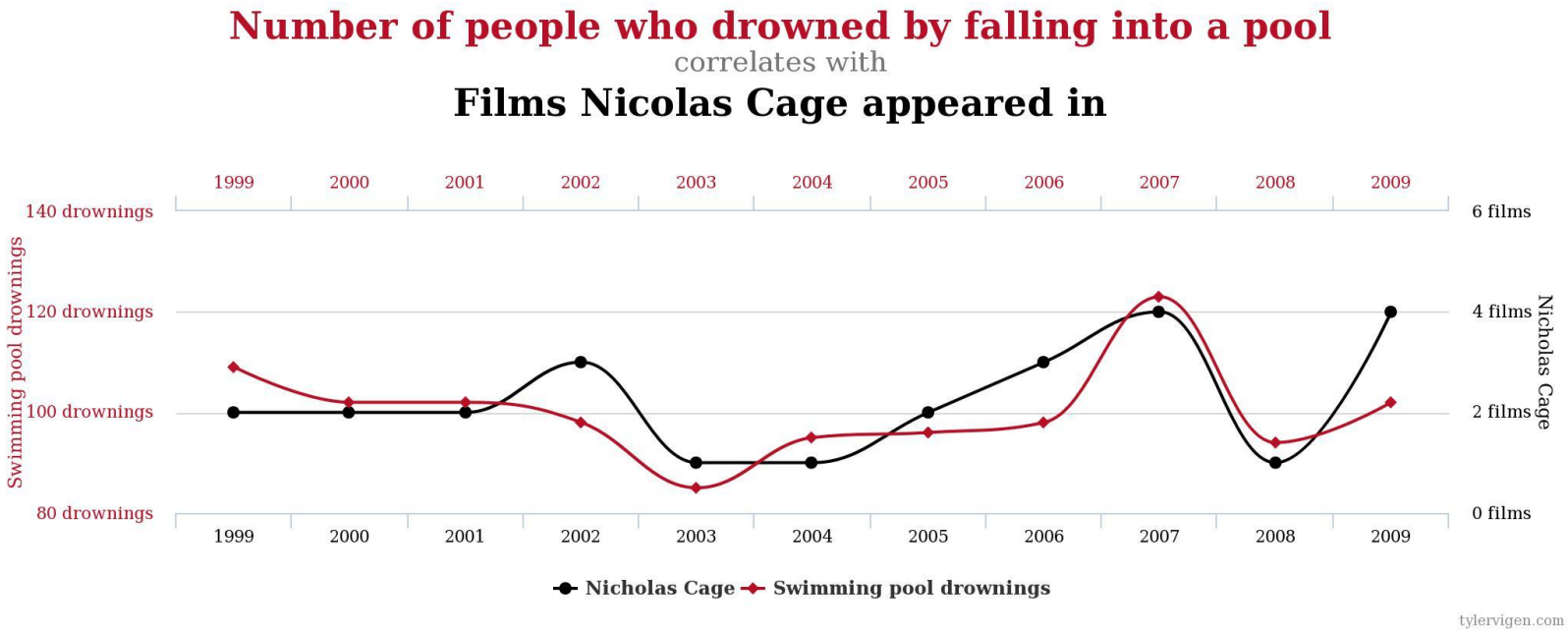
Split Based Variable Importance



Split Falls for Milk Consumption

Source: Strobl 2009

Don’t Fall for a Spurious Correlation

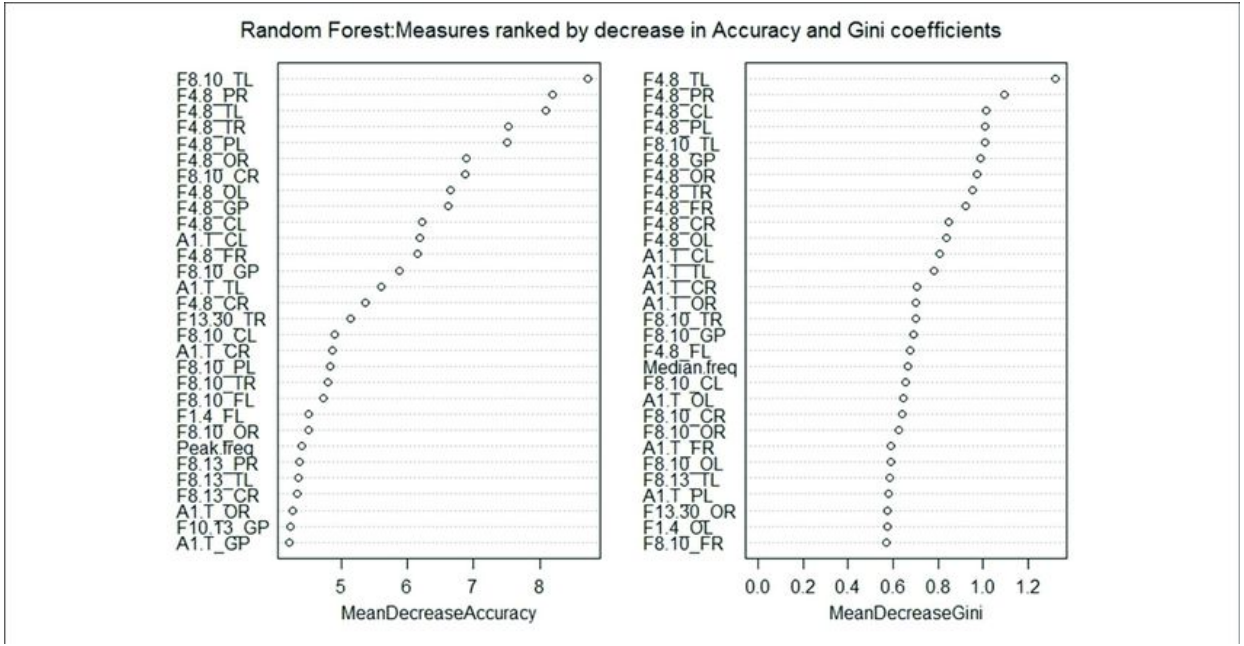


Machine Learning Can Identify Spurious Correlations

If your feature impact is wrong, you are toast.

R

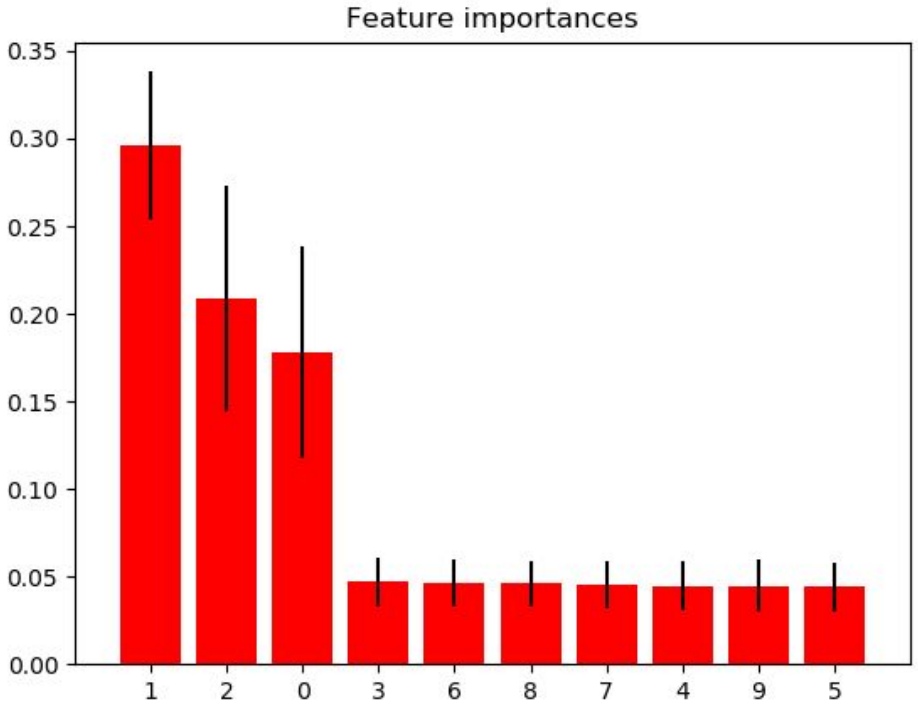
R randomforest shows both permutation and gini based importance



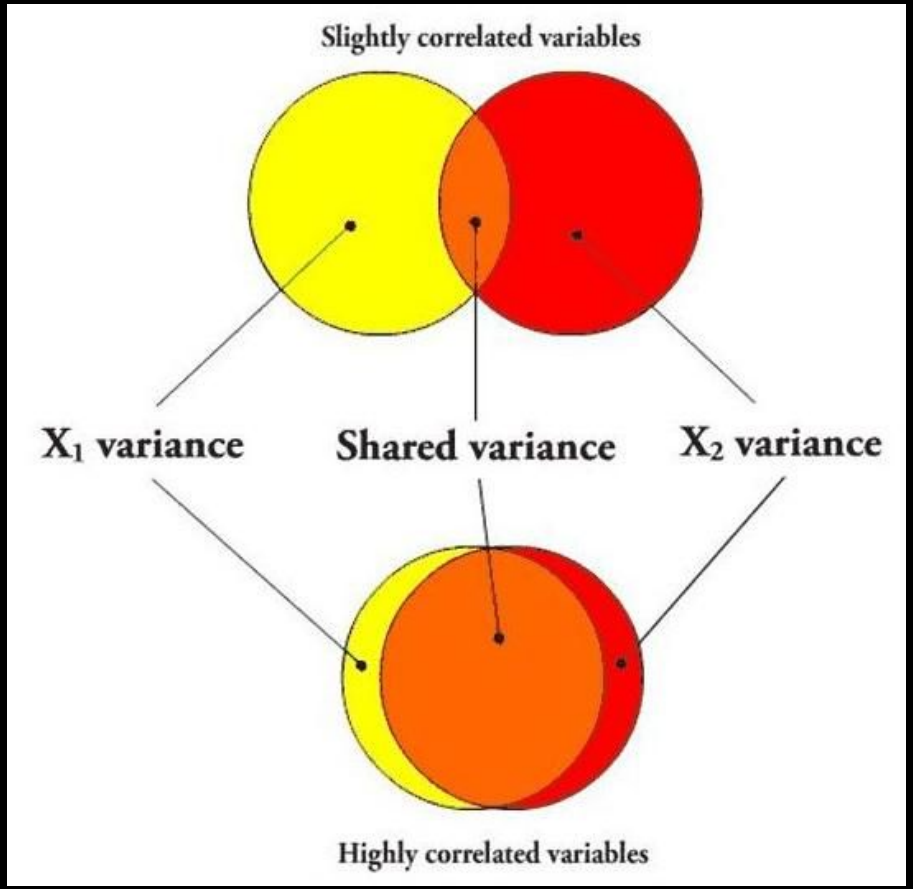
YEA, R SUPPORTS PERMUTATION!

Python

Python sklearn only uses gini for feature importance

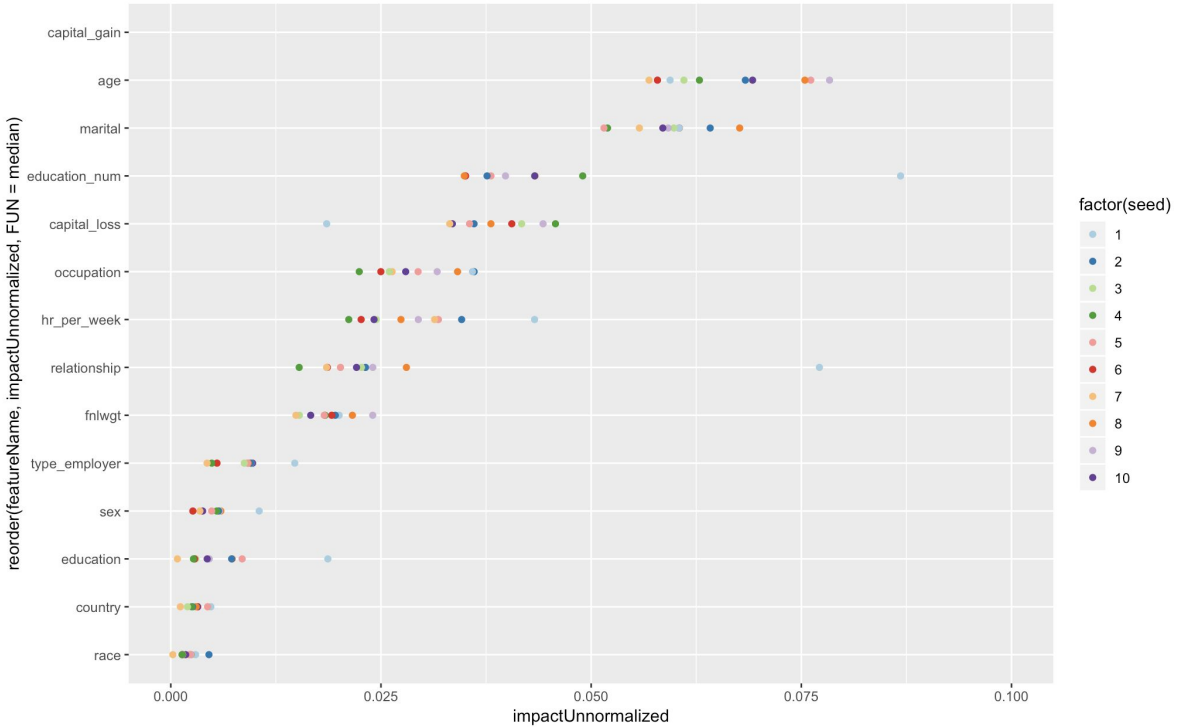


BOO!, PYTHON DOES NOT SUPPORT PERMUTATION!



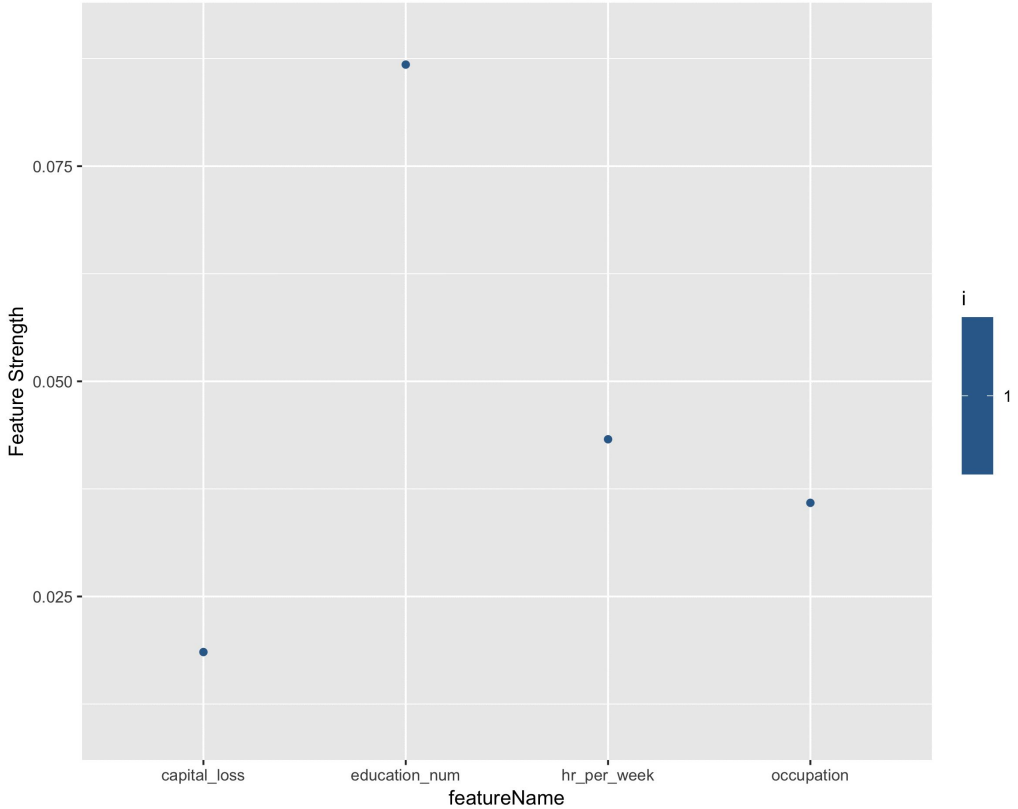
Multicollinearity

Run It Again



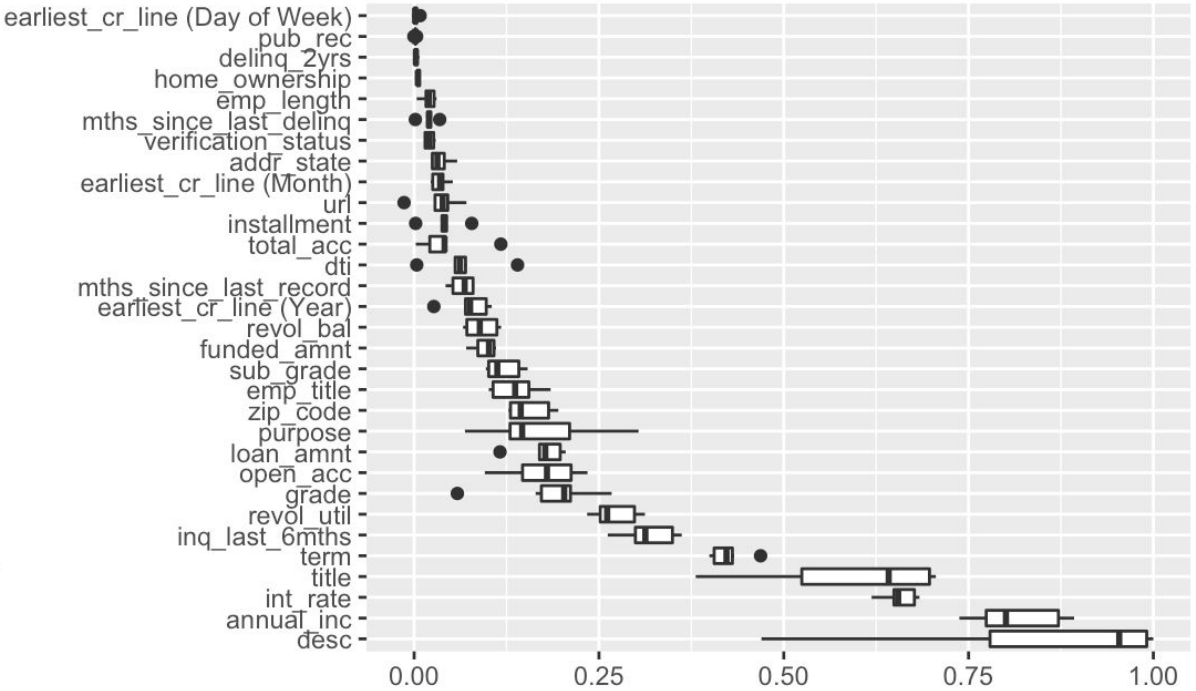
10 different models, 10 different feature importances

Multicollinearity affects Interpreting models



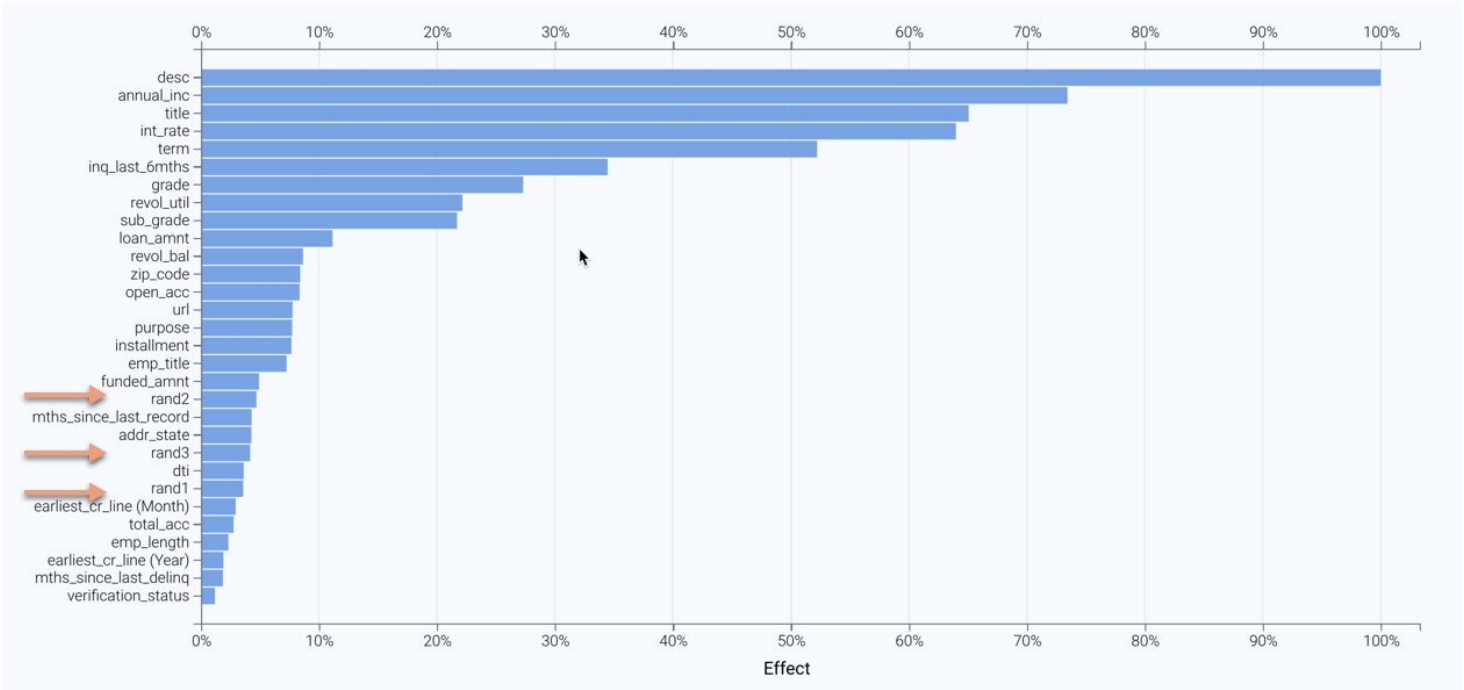
Features trade off against each other in different model runs

Pro Tip: Aggregate Feature Importance to Provide a Richer Understanding



This plots show how the ranking of feature importance varies across multiple model runs

Pro Tips: Add Random Features



Helps you understand the line between signal and noise

Permutation based importance is a good balance of computation and performance for any model

Further study:

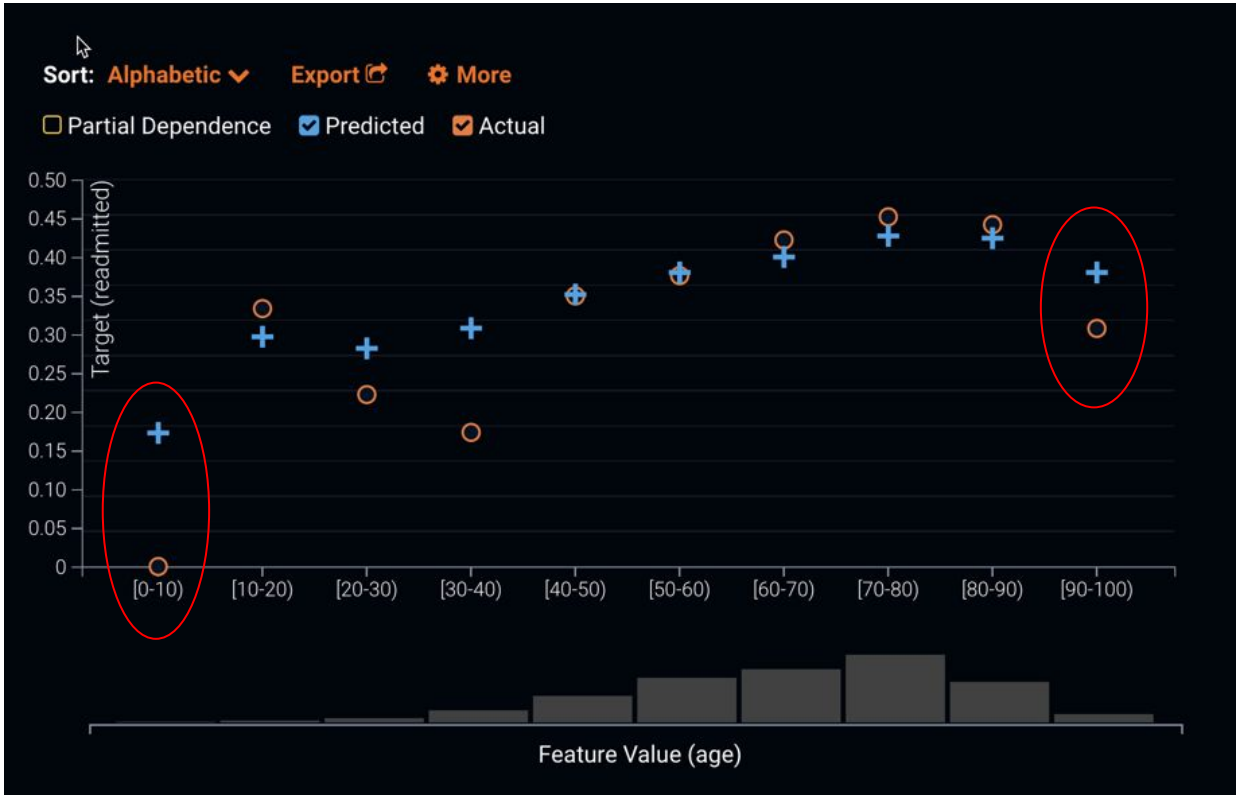
Studies on permutation based importance: Strobl 2008 and Lundberg 2018 and explained.ai and datadive . . . more advanced approaches - Party, Shap, and Boruta

Partial Dependence

Age

Weight

Effect of Age on our Target

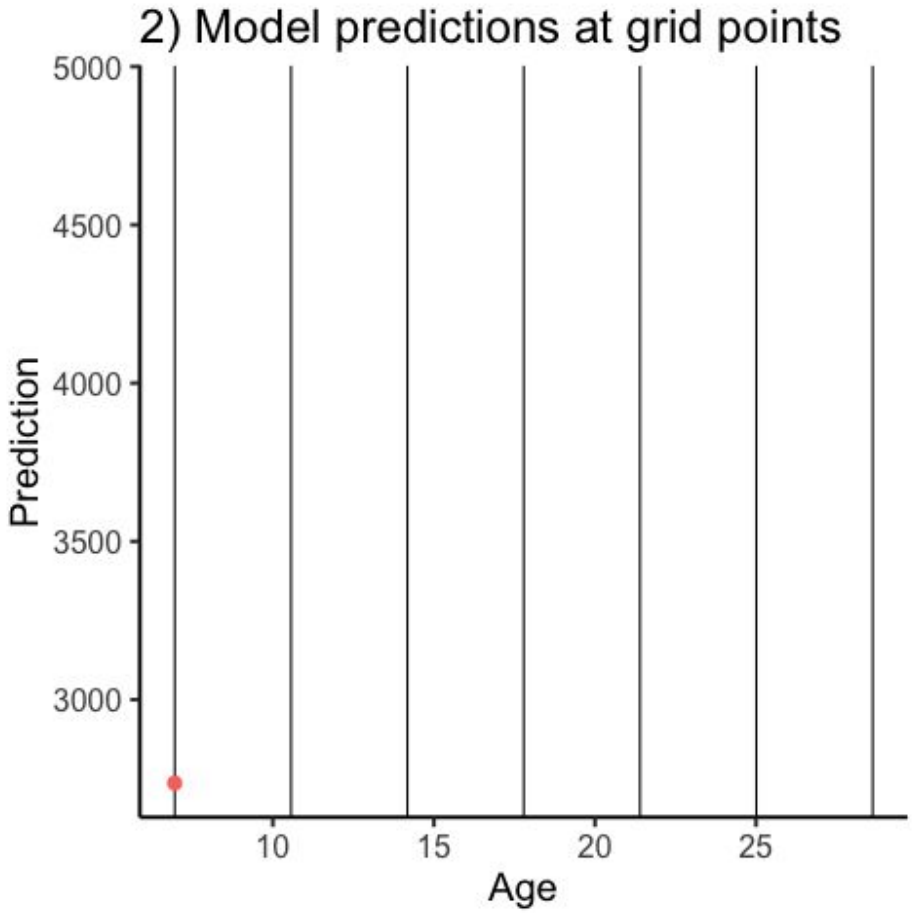


What is the average weight for each of these bins?

THIS PLOT DOES NOT ISOLATE THE EFFECT OF AGE

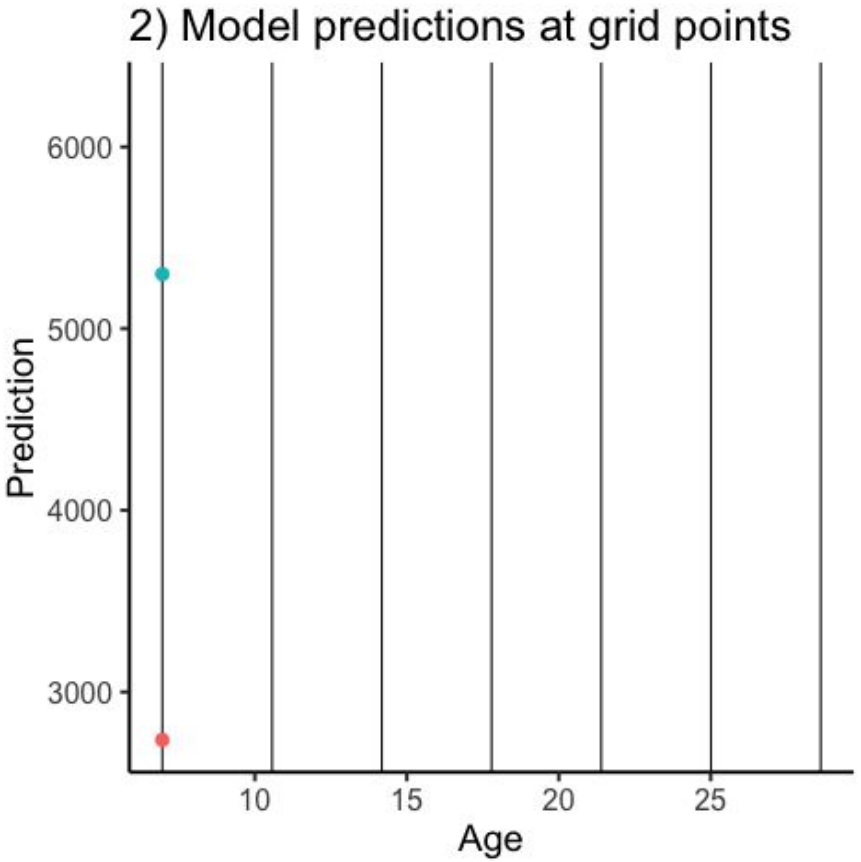
Calculating Partial Dependence

Start with an observation and get predictions for different values



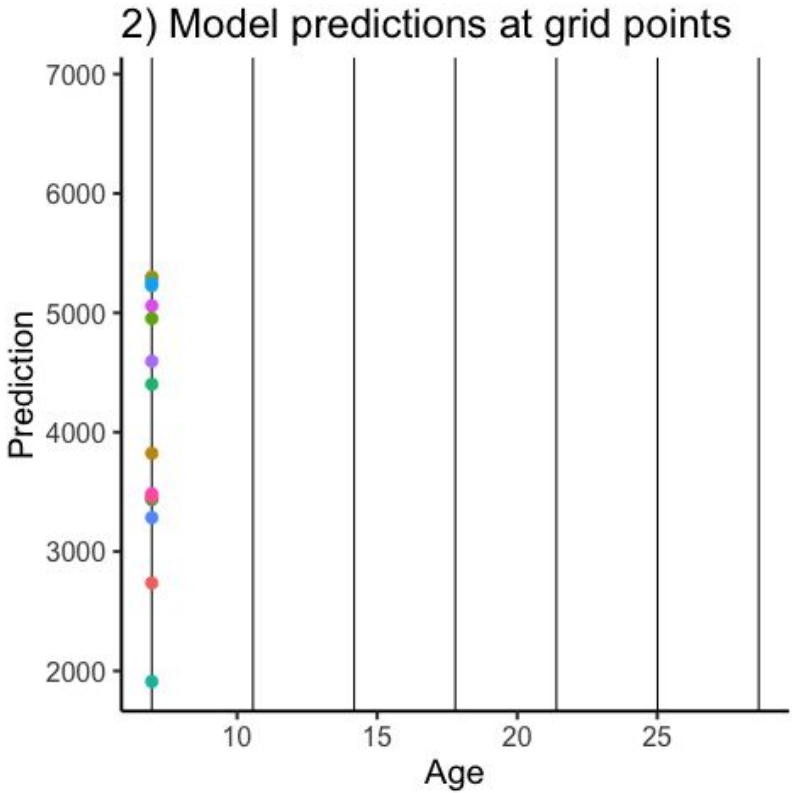
Calculating Partial Dependence

Start with another observation and get predictions for different values



How Partial Dependence is Calculated

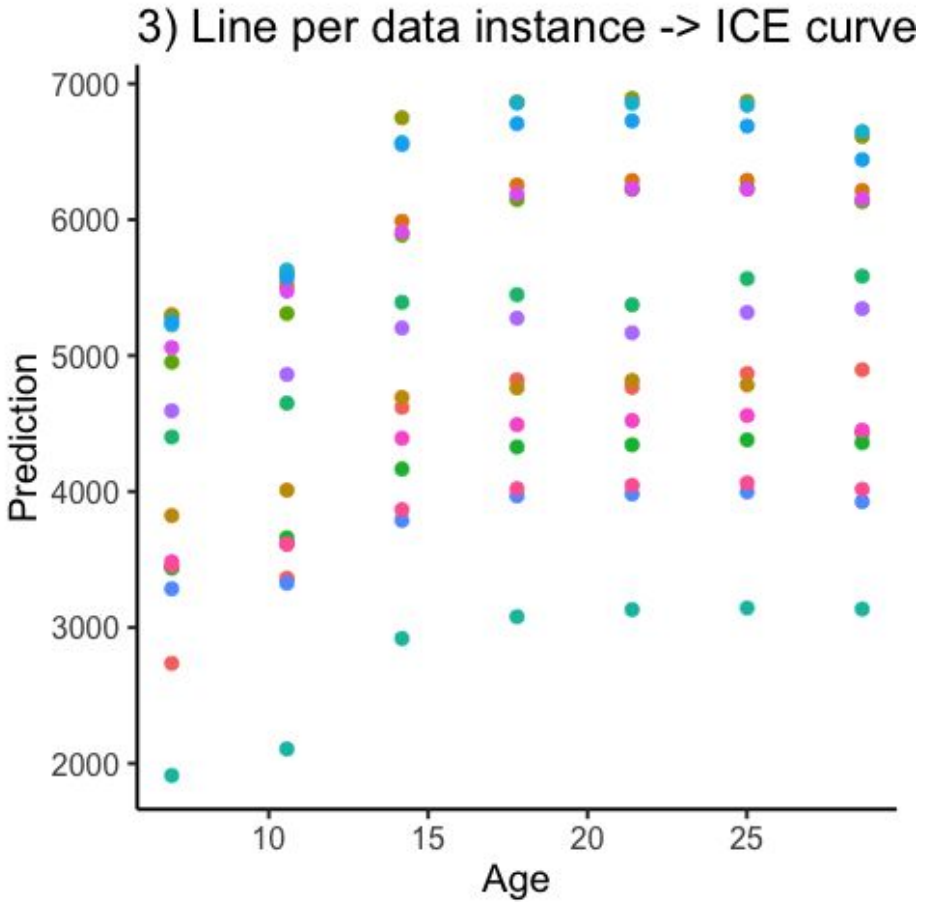
Start with a set of observations from our dataset



Source: Christoph Molnar

How Partial Dependence is Calculated

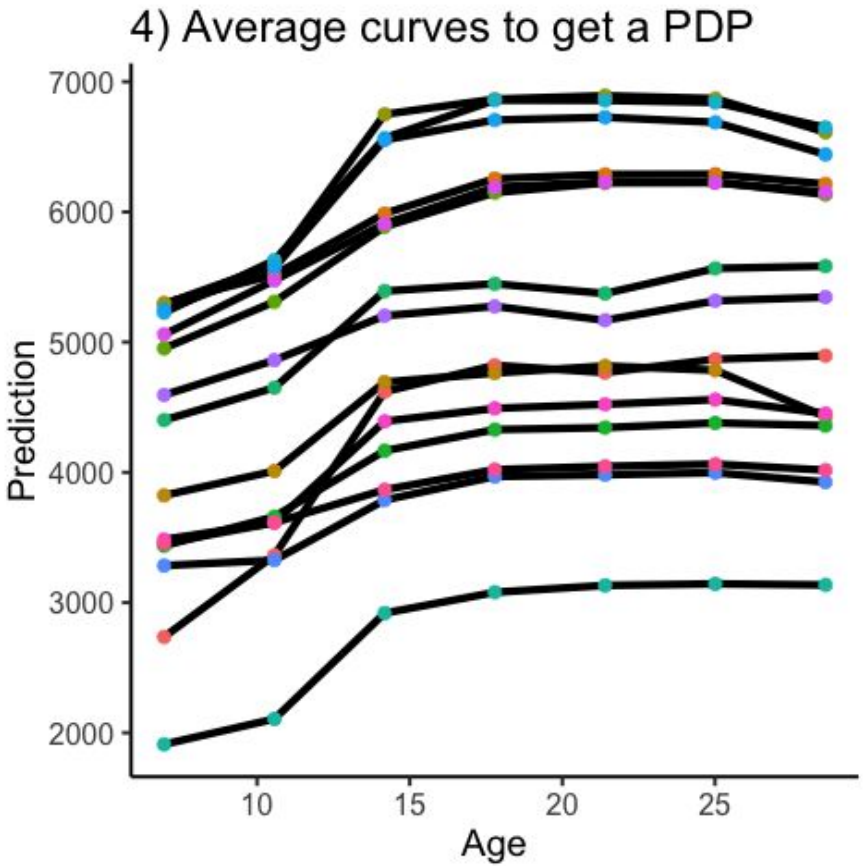
Get a line per instance



Source: [Christoph Molnar]

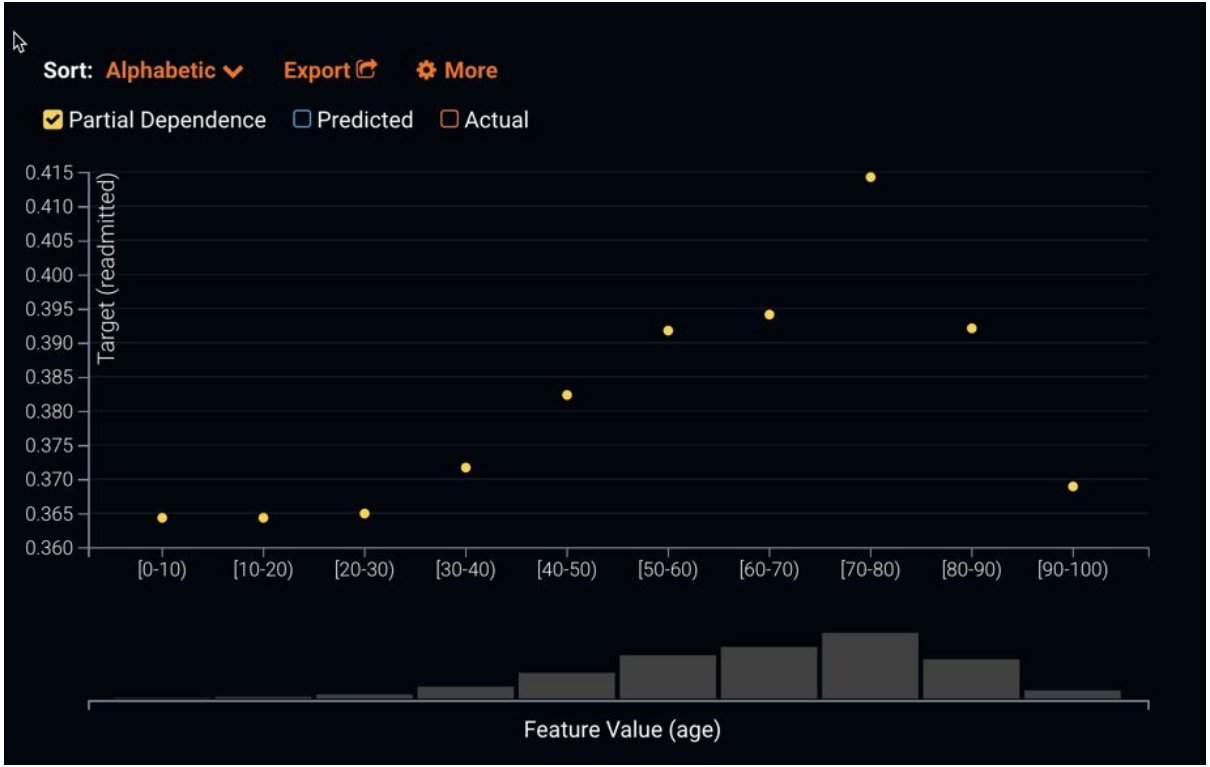
How Partial Dependence is Calculated

Average the curves to get the partial dependence curve

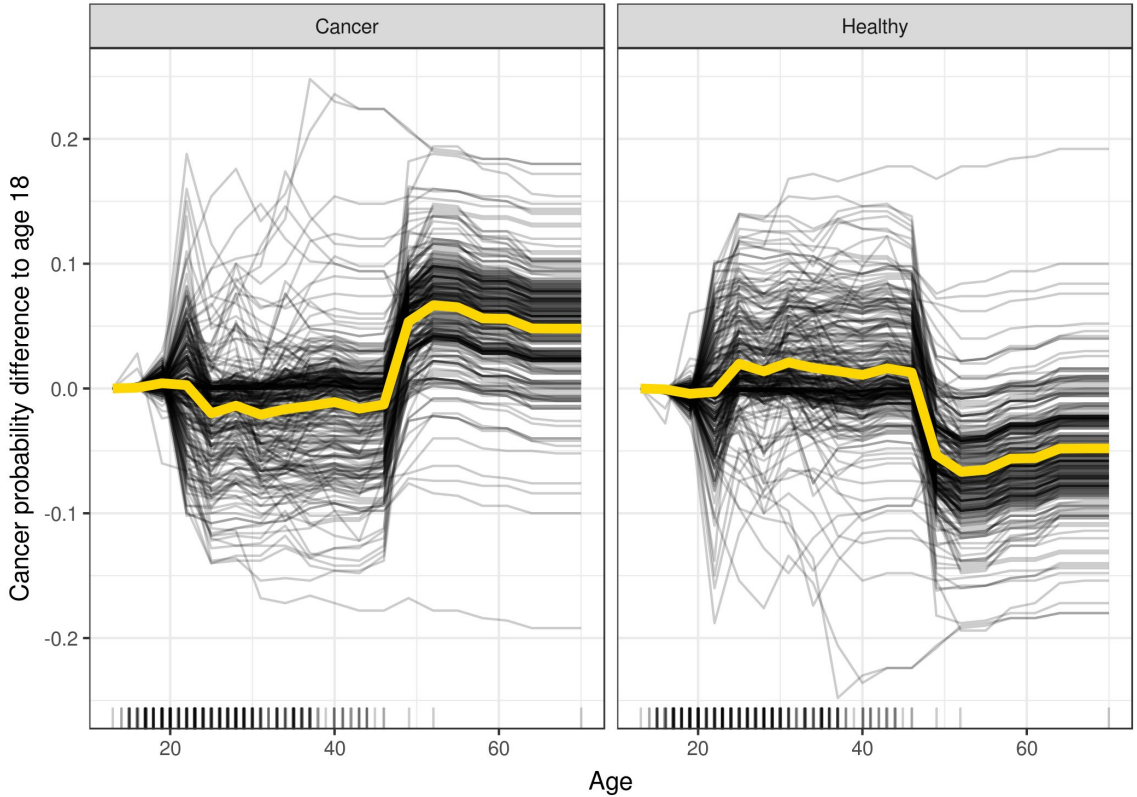


Source: [Christoph Molnar]

Partial Dependence to Isolate the Effect of Age



ICE Plots



Individual Conditional Expectation plots draw one line per instance

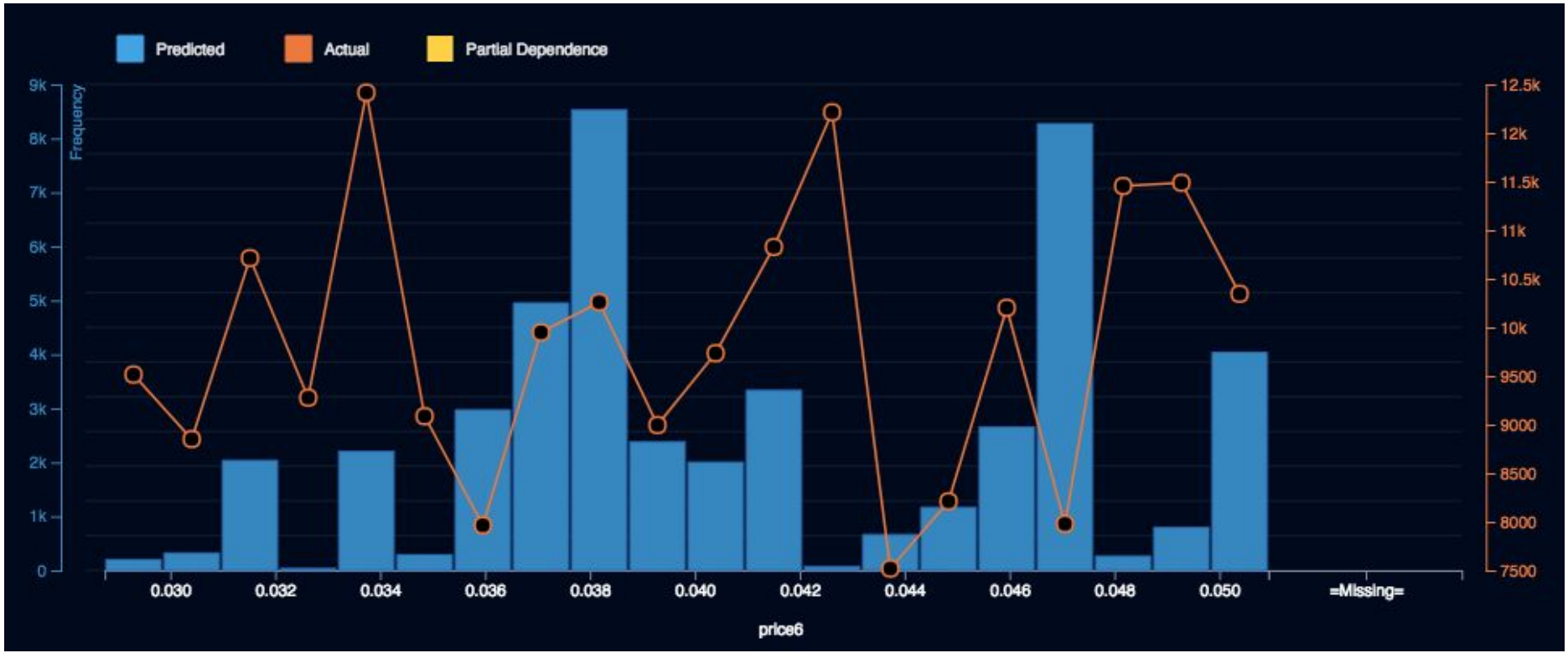
Partial Dependence to show Price Elasticity

Effect of price on sales of orange juice

Features include:

- store location
- date
- coupons
- advertising
- prices for 10 other brands

Change in Price Affects Sales?



Ahh, Price does affect Sales!

$3.46

$4.23

Predicted
Actual
Partial Dependence
Frequency
9k
8k
7k
6k
5k
4k
3k
2k
1k
0
10.35k
10.3k
10.25k
10.2k
10.15k
10.1k
10.05k
10k
9950
9900
9850
9800
9750
0.030
0.032
0.034
0.036
0.038
0.040
0.042
0.044
0.046
0.048
0.050
=Missing=
price6

Partial dependence is a best practice for understanding the features in your model

Further study:

- Friedman, 2001 on PDP
- Goldstein, 2013 on ICE Plots

Predictions

Prediction: 9.1

Explanations:
1. # of Past Kills (+0.8)
2. Color (+0.3)
3. Gender (-0.2)

Predictions & Explanations

Charge Nurse

Nurse Susan

Floor

1st

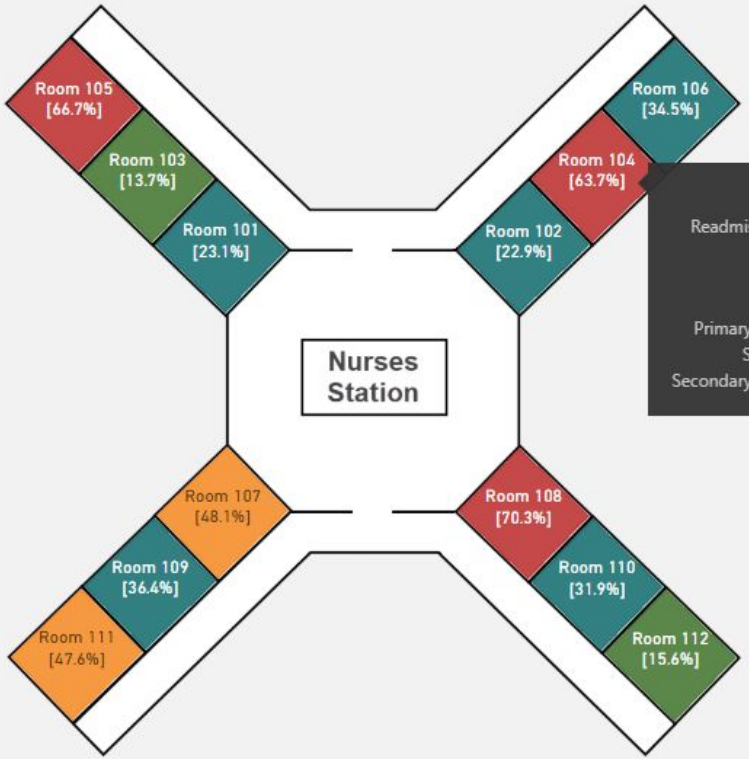
3
Urgent Risk Patients

2
High Risk Patients

5
Moderate Risk Patients

2
Low Risk Patients

Floor Map with Readmission Probability by Room

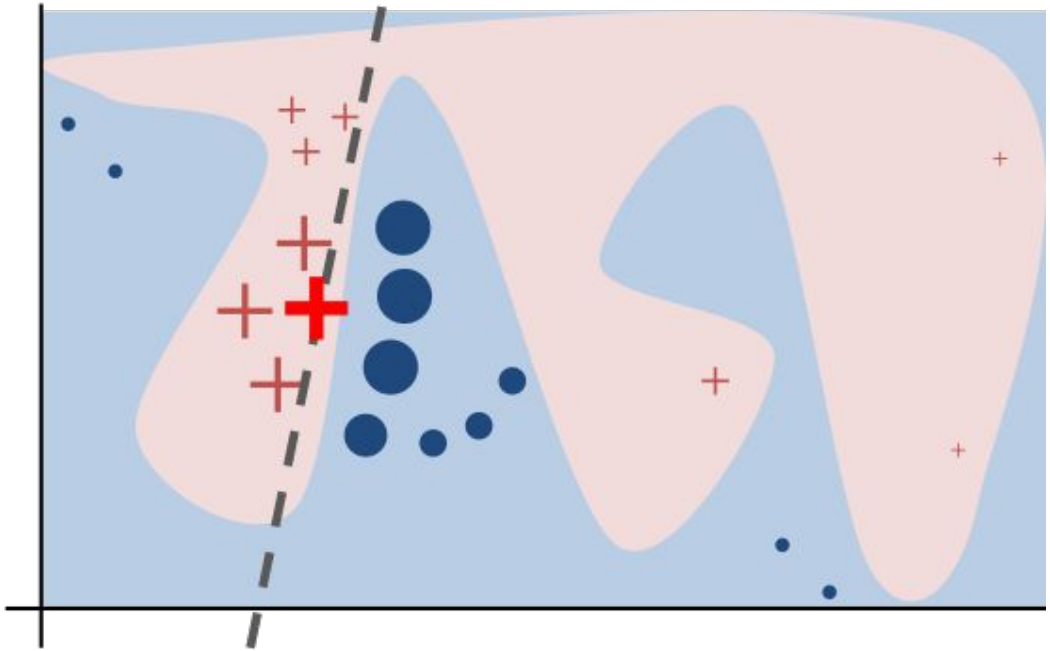


Location	**Room 104**
Readmission Probability	**63.7%**
Patient Name	**Lester Briones**
Primary Factor	**Primary Diagnosis: Abdominal pain, unspecified site**
Primary Factor Strength	**High**
Secondary Factor	**Medical Specialty: Unspecified**
Secondary Factor Strength	**Medium**

Explanation Methods:
Local Interpretable Model-Agnostic Explanations (LIME)

For any prediction:

LIME gives you an ordered list of the most important features for that prediction



SPEND SOME TIME WITH LIME

Prediction: 9.1

Explanation (1)

1. # of Past Kills (+0.8)
2. Color (+0.4)
3. Gender (-0.2)

Explanation (2)

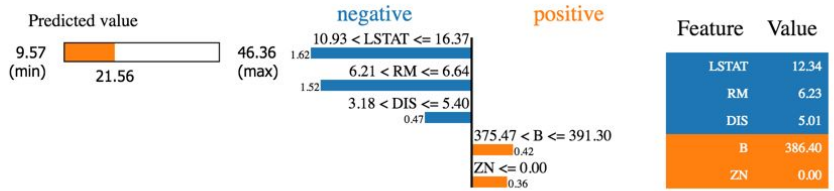
1. Gender (+0.5)
2. Breath Fire (+0.3)
3. Weight (+0.1)

Explanations should be Identical

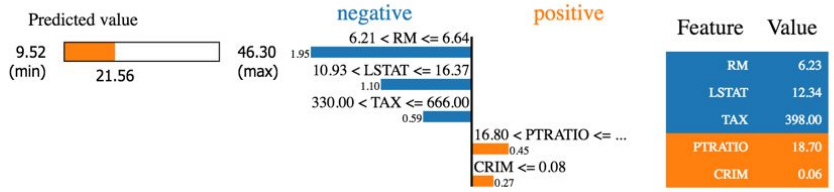
Identical Explanations

```
In [56]: i = 13
exp = explainer.explain_instance(test[i], rf.predict, num_features=5)
print (exp)
exp.show_in_notebook(show_table=True)
i = 13
exp = explainer.explain_instance(test[i], rf.predict, num_features=5)
print (exp)
exp.show_in_notebook(show_table=True)
```

```
Intercept 24.51223189433816
Prediction_local [21.66593303]
Right: 21.55850000000006
<lime.explanation.Explanation object at 0x1a2acd6860>
```




```
Intercept 24.84927186407495
Prediction_local [21.91763542]
Right: 21.55850000000006
<lime.explanation.Explanation object at 0x1a1c040c50>
```



Same Data, Same Model . . . Two Different Explanations!!

Prediction: 9.1

Explanations:
1. # of Past Kills (+0.8)
2. Color (+0.4)
3. Gender (-0.2)

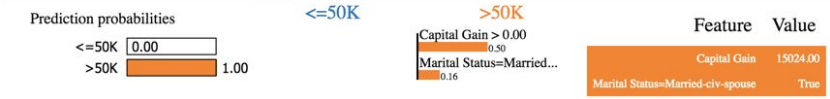
Prediction: 2.4

Explanations:
1. # of Past Kills (+0.8)
2. Color (+0.4)
3. Gender (-0.2)

EXPLANATIONS SHOULD HAVE FIDELITY TO THE DATA

Local Fidelity

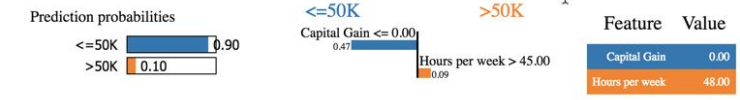
```
In [22]: np.random.seed(1)
         i = 1653
         exp = explainer.explain_instance(test[i], predict_fn, num_features=2)
         exp.show_in_notebook(show_all=False)
```

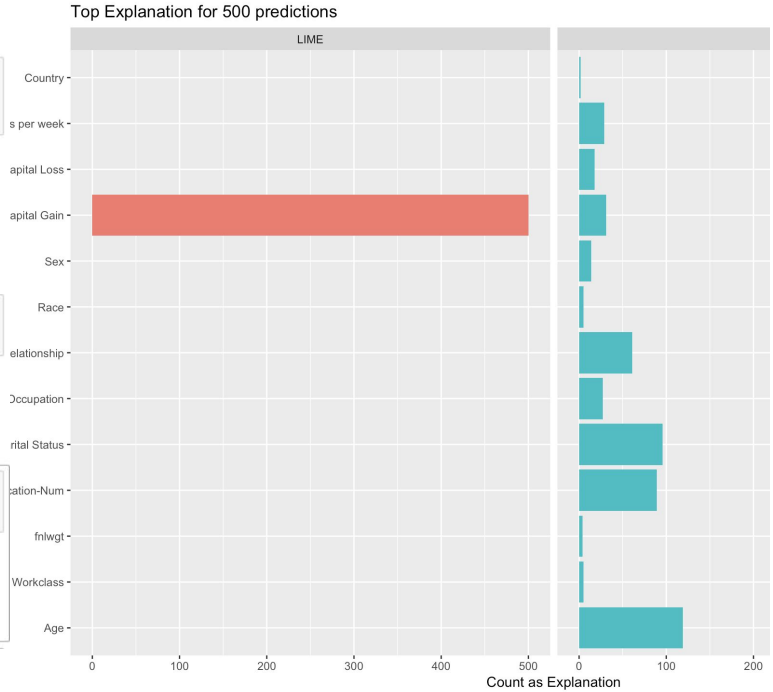


Note that capital gain has very high weight. This makes sense. Now let's see an example where the person has a capital gain below the mean:

```
In [23]: i = 59
         exp = explainer.explain_instance(test[i], predict_fn, num_features=2)
         exp.show_in_notebook(show_all=False)
```


```
In [24]: i = 26
         exp = explainer.explain_instance(test[i], predict_fn, num_features=2)
         exp.show_in_notebook(show_all=False)
```





LIME EXPLANATIONS AREN'T RESPONSIVE TO THE DATA

Anyone relying on LIME is toast

Your Model or a Surrogate Model?

Your model

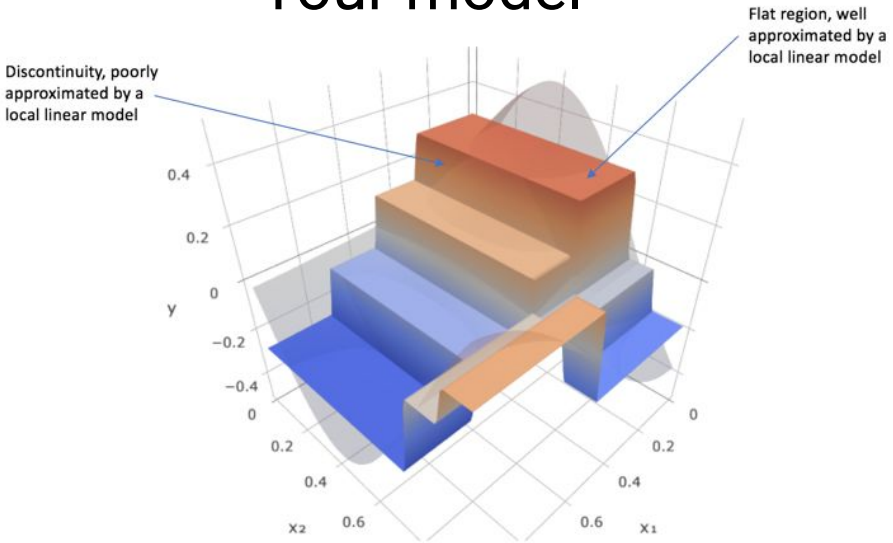
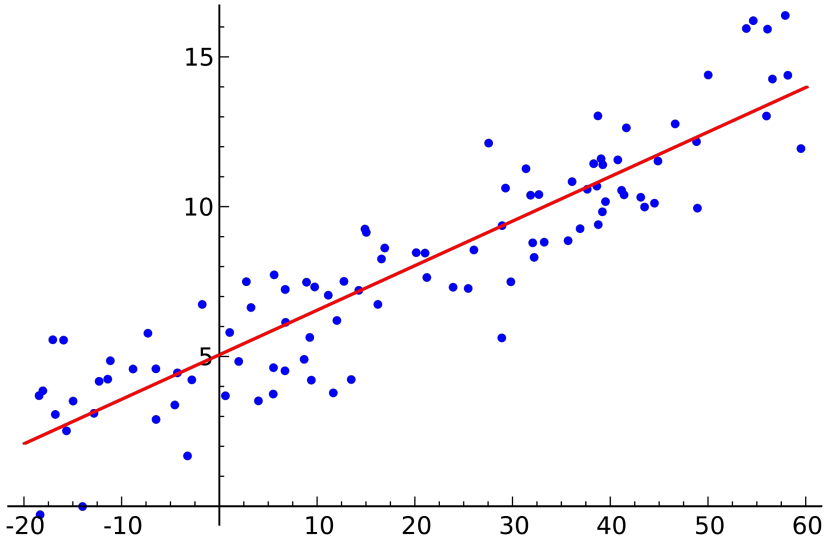


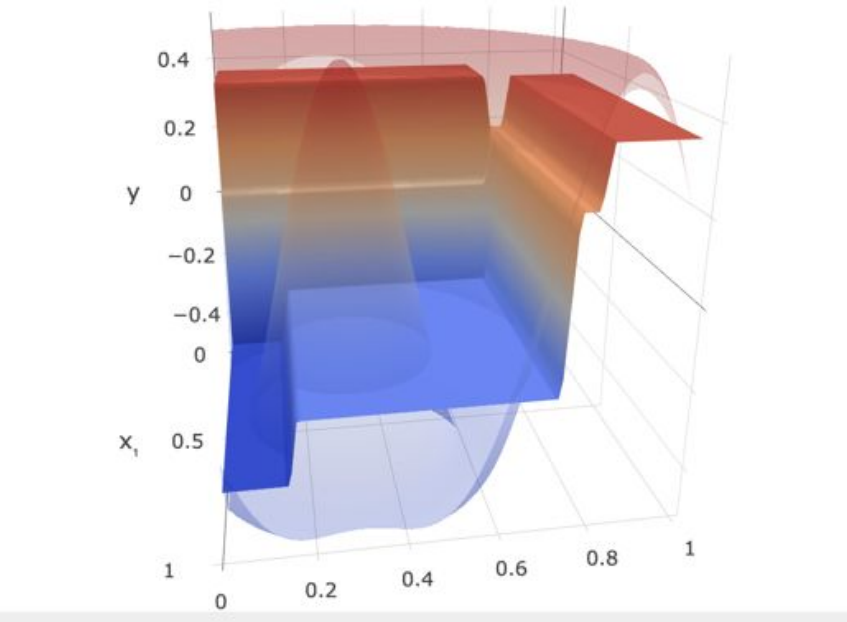
Image source: http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html

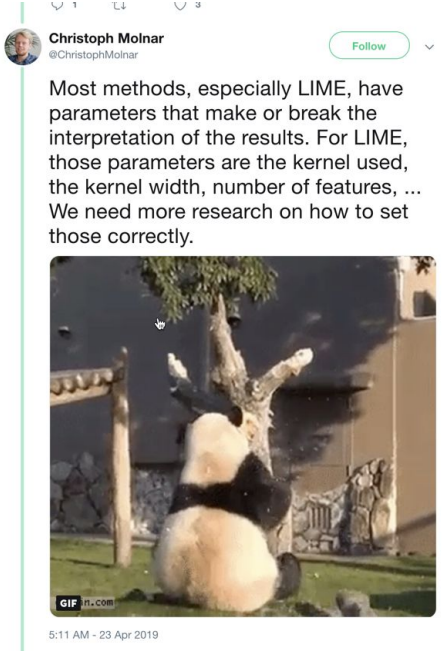
LIME's surrogate model



SURROGATE MODELS ARE APPROXIMATIONS

What is local?

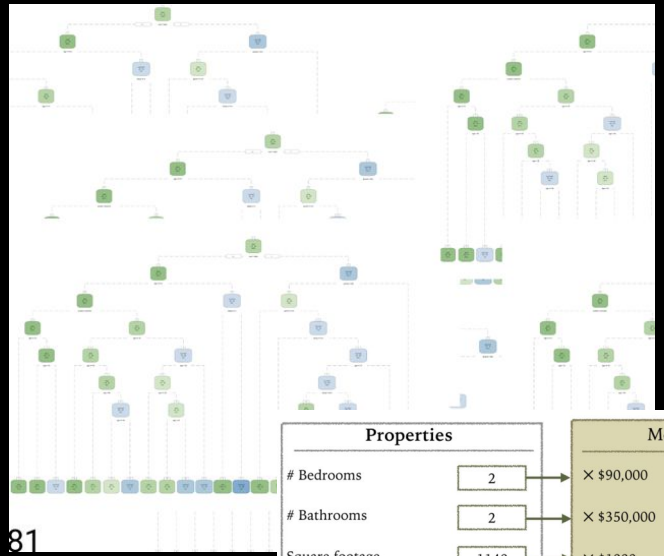


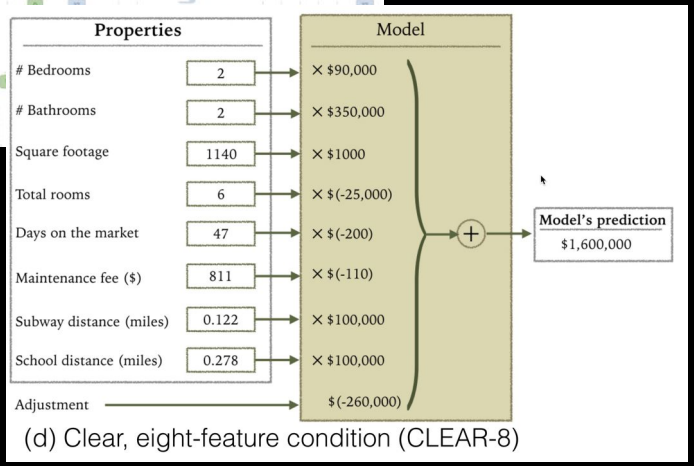


Source: Christoph Molnar

UGH!

LIME HAS HYPERPARMETERS TO DECIDE HOW TO SET





Explanation methods should be Model Agnostic

	Time in seconds	
	LIME	**XEMP**
Boston Housing (100 explanations)	43	0.3
Adult (1000 explanations)	423	3

Explanations should be Fast

Don’t use LIME

Lots of tradeoff in explanation approaches

Further study:

- Explanations affect fairness: Dodge
- Shap: Lundberg
- Live and Breakdown: Biecek

Use This!
Model Agnostic Explanation Tools

What are features driving the model? Feature importance

How is a specific feature driving? Partial dependence

Let’s explain some examples? Prediction explanation techniques (LIME, SHAP, XEMP . . .)

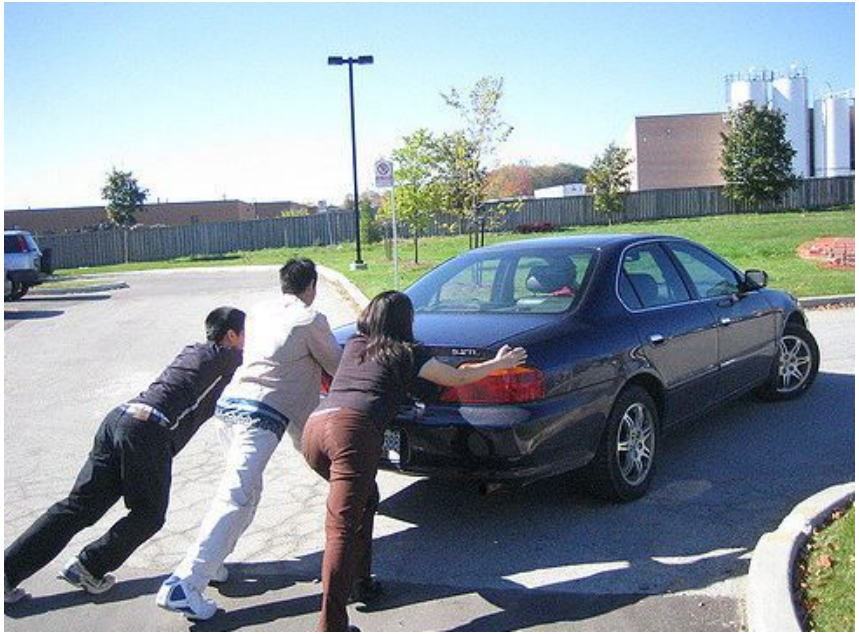
Question time

rajiv.shah@datarobot.com
@rajcs4

https://bit.ly/inter_workshop

Additional Material if you have more than 1 hour

Shapley Values

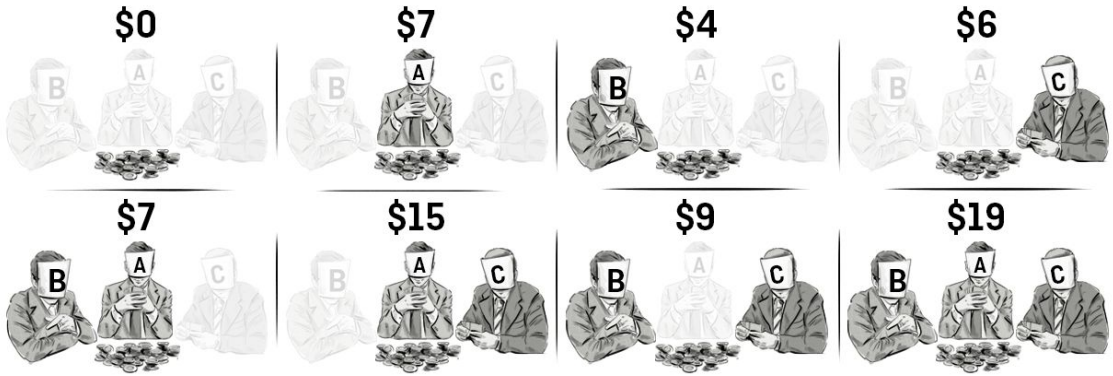


Beautiful Idea from Game Theory . . . Won some Awards

Calculating Shapley Values

Derived from game theory: For cooperating games, how much is the contribution for each player

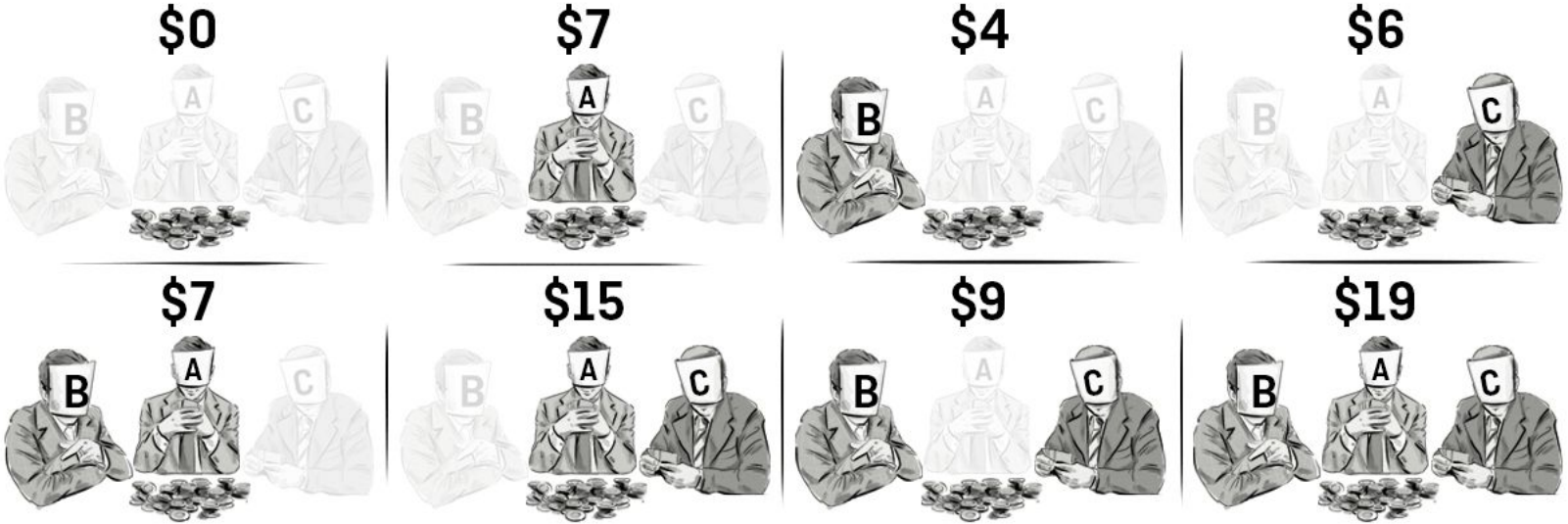
We can apply this to identify the contribution of each feature



1. Calculate the contribution in every scenario
2. Calculate each player's marginal contribution
3. The average is the Shapley Value

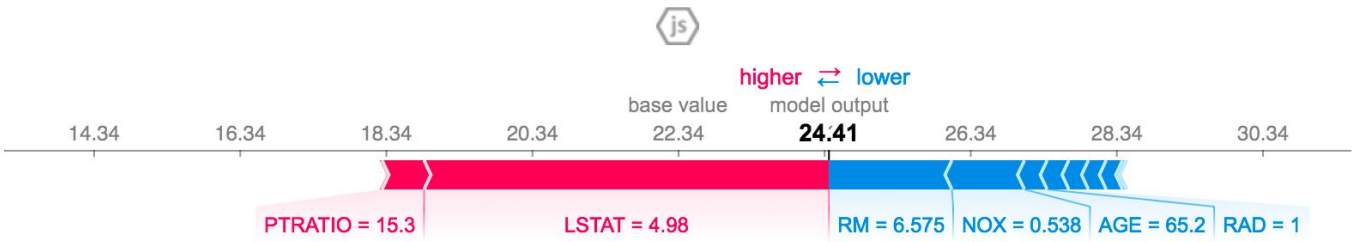
	A	B	C
ABC	7	0	12
ACB	7	4	8
BAC	3	4	12
BCA	10	4	5
CAB	10	3	6
CBA	9	4	6
	7.67	3.17	8.17

Shapley Values



UGH!!, DOESNT SCALE

Shapley Values



1, Strumbelj approximation that uses permutation (iml)

2. Shap Kernel uses uses a specially-weighted local linear regression to estimate SHAP values for any model.

3. Tree Shap is fast and accurate for tree based models

Approximating Shapley Values: Strumbelj

Strumbelj approximation:

Compare the value of interest to a permuted value repeatedly. This gets you the Shapley values.

You see this in the iml R package for getting Shapley values

An approximation

1. Repeat for M times.
2. Pick a feature j from the instance x_i.
3. Generate synthetic instances x_L and x_U using x_{ij} pivot.
4. Estimate individual contribution $\phi_{ijm} = \hat{f}(x_L) - \hat{f}(x_U)$
5. Estimate average contribution of feature j at the prediction of the i-th subject as: $\phi_{ij}(x) = \frac{1}{M}\sum_{m=1}^{M}\phi_{ijm}$

Approximating Shapley Values: Shap Kernel

Kernel SHAP uses a specially-weighted local linear regression to estimate SHAP values for any model.

The intuition is we can use least squares point estimate to get the mean values, i.e., the Shapley values

Better consistency than LIME

It is slow

Approximating Shapley Values: Tree Shap

A fast and exact algorithm to compute SHAP values for trees and ensembles of trees.