SHAP: Interpreting ML Models with IML

March 7, 2020
SatRday

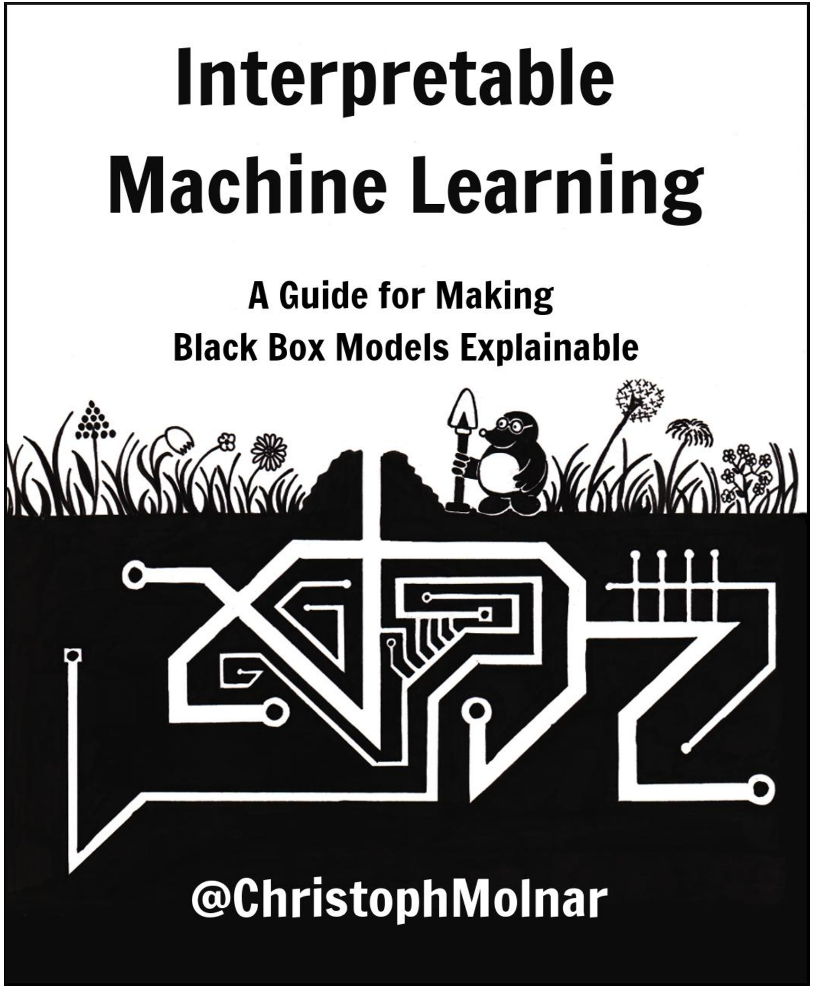
Who am I?

Drikus du Toit

Data Scientist
Decision Science
Capitec Bank

Resource

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.



What is ML interpretability and why is it important?

The Client

Applies for loan, gets rejected.

Client Questions:

- Why did I not get the loan?
- What should I do to improve my credit score?

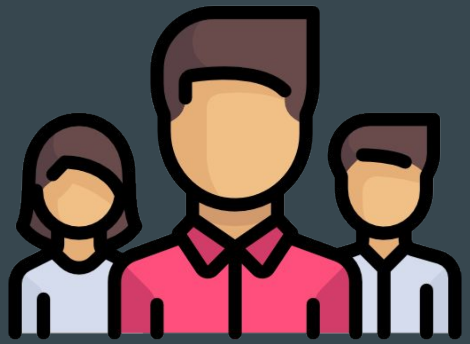
The Business

Build a credit default model. High accuracy usually goes with high complexity.

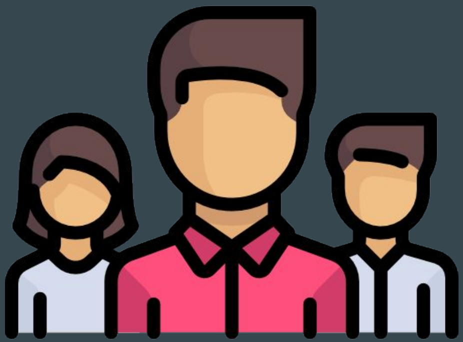
Business Questions:

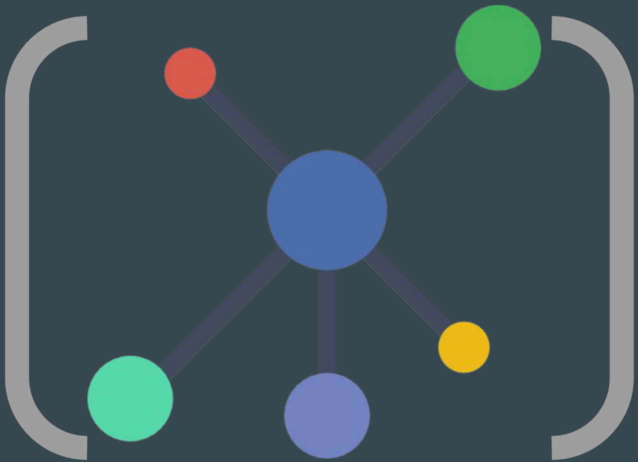
- Is there bias in our model?
- Do we understand our underlying data?
- What will cause the model to not perform as expected?
- Are we within regulatory framework?

Context..



Context..





Context..

	Interpretable	Accurate
Complex Model		
Simple Model		

Context..

	Interpretable	Accurate
Complex Model		✓
Simple Model		

Context..

	Interpretable	Accurate
Complex Model	✗	✓
Simple Model		

Context..

	Interpretable	Accurate
Complex Model	✗	✓
Simple Model	✓	

Context..

	Interpretable	Accurate
Complex Model	✗	✓
Simple Model	✓	✗

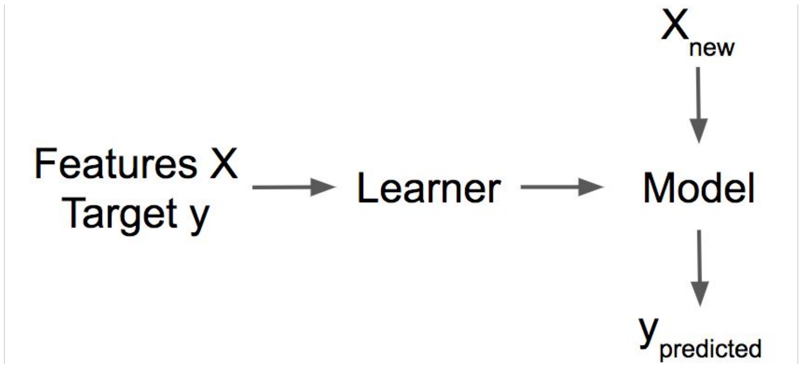
Context..

	Interpretable	Accurate
Complex Model	✗	✓
Simple Model	✓	✗

Context..

	Interpretable	Accurate
Complex Model	✗	✓
Simple Model	✓	✗

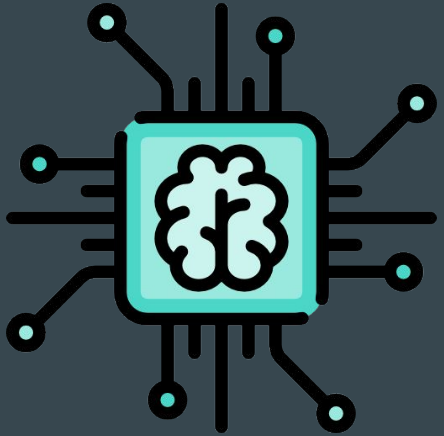
Machine Learning Model



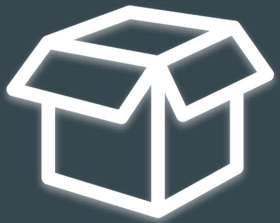
https://christophm.github.io/interpretable-ml-book/terminology.html

Importance of Interpretability (what or why)

1. Human curiosity and learning
2. Goal of science
3. Safety measures
4. Detecting bias
5. Manage social interactions
6. Debugged and audited



Interpretability Techniques



- **Reveal its internal mechanisms**
- Fully understood by looking at their parameters
- Also called interpretable models

- **Does not reveal its internal mechanisms**
- Cannot be understood by looking at their parameters (e.g. a neural network)

Black Box Models (interpretability techniques)

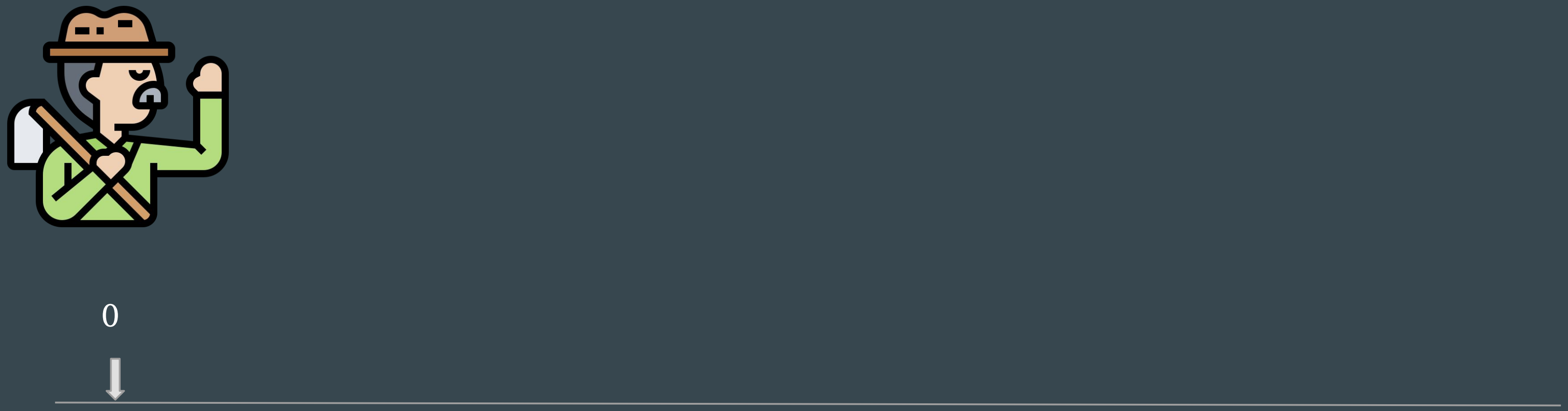
Model Agnostic Methods	Example Based Explanations
Partial Dependence Plot (PDP)	Counterfactual Explanations
Individual Conditional Expectation (ICE)	Adversarial Examples
Accumulated Local Effects (ALE) Plot	Prototypes and Criticisms
Feature Interaction	Influential Instances
Permutation Feature Importance	
Global Surrogate	
Local Surrogate (LIME)	
Scoped Rules (Anchors)	
Shapley Values	
SHAP (SHapley Additive exPlanations)	

Black Box Models (interpretability techniques)

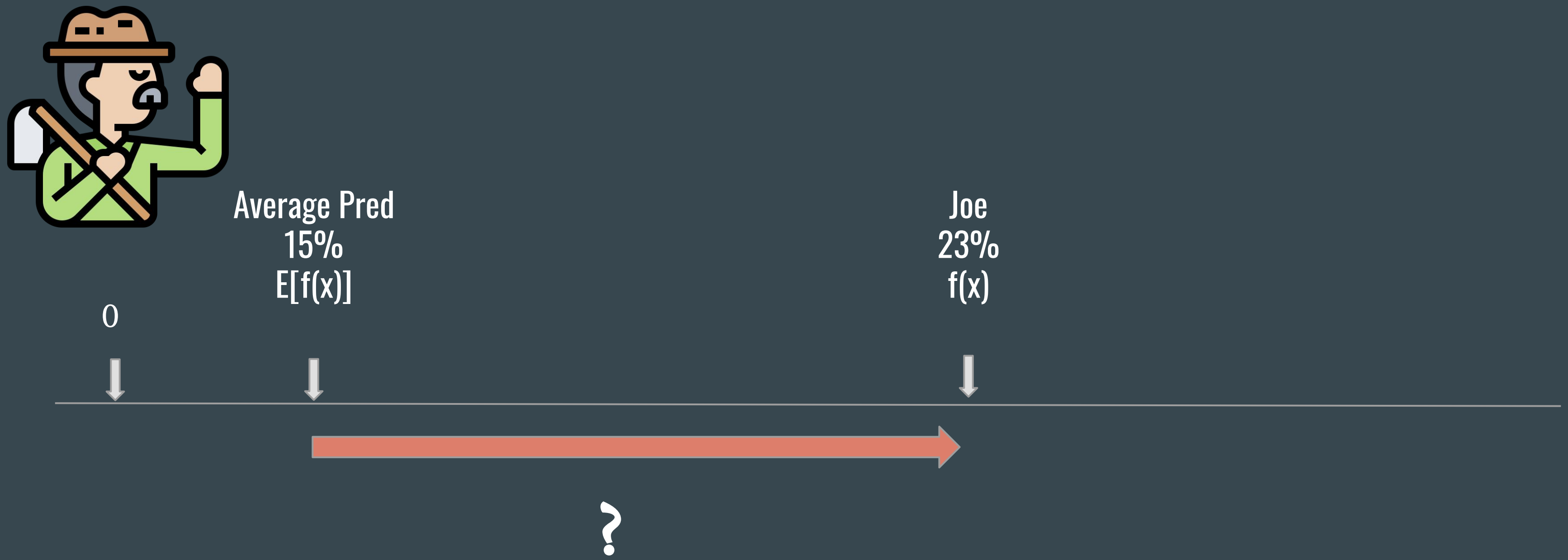
Model Agnostic Methods	**Example Based Explanations**
Partial Dependence Plot (PDP)	Counterfactual Explanations
Individual Conditional Expectation (ICE)	Adversarial Examples
Accumulated Local Effects (ALE) Plot	Prototypes and Criticisms
Feature Interaction	Influential Instances
Permutation Feature Importance	
Global Surrogate	
Local Surrogate (LIME)	
Scoped Rules (Anchors)	
Shapley Values	
SHAP (SHapley Additive exPlanations)	

Shapley Values


23 %

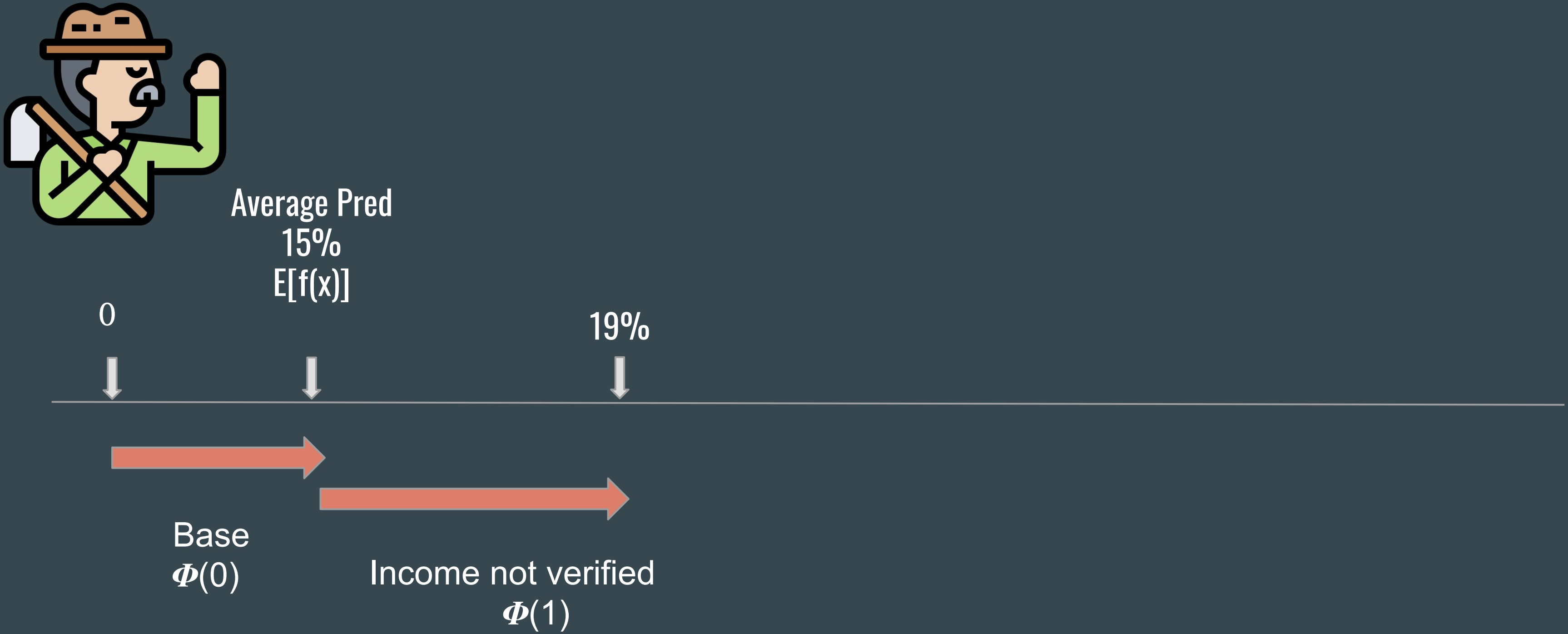

0


Average Pred
15%
0

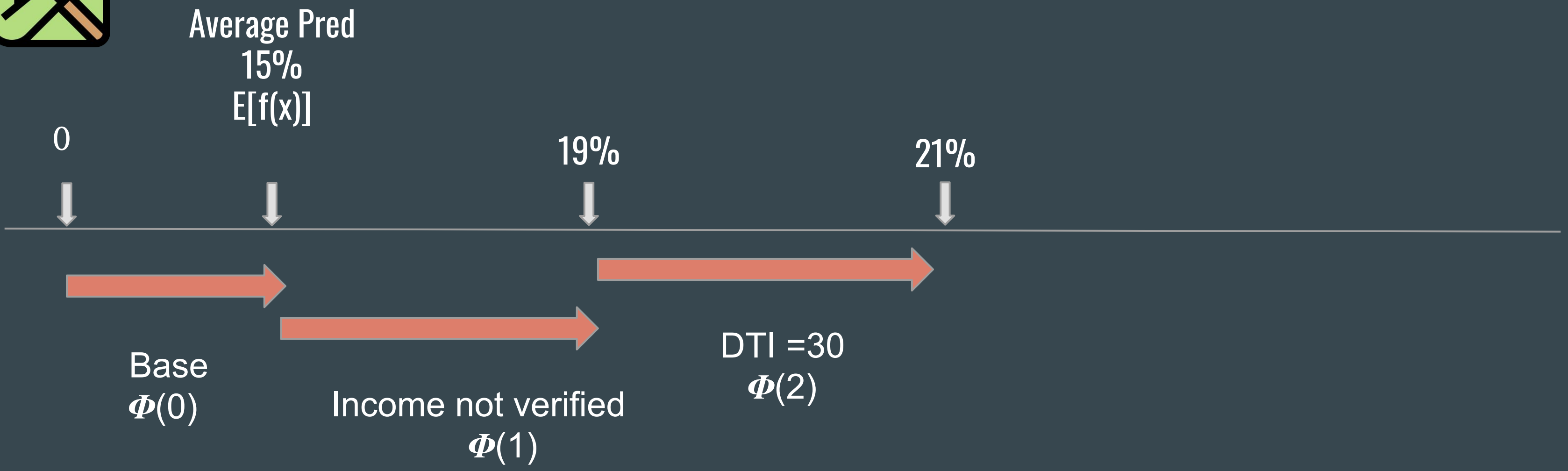


Scott Lundberg H2O World

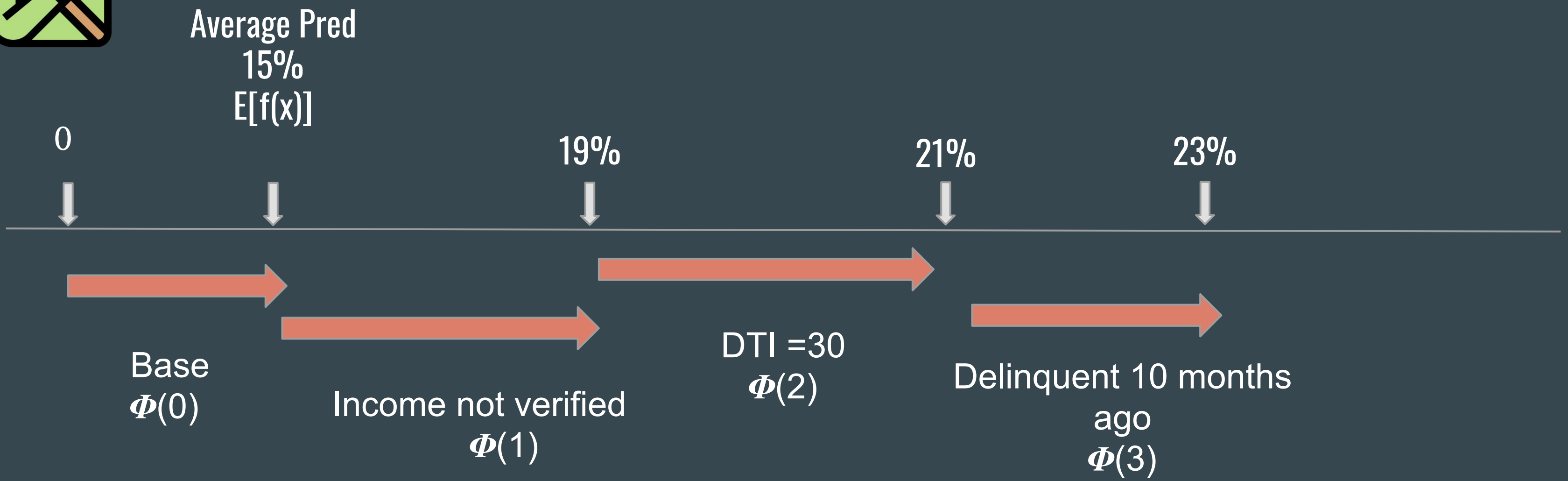

Average Pred
15%
E[f(x)]
0
Base
Φ(0)

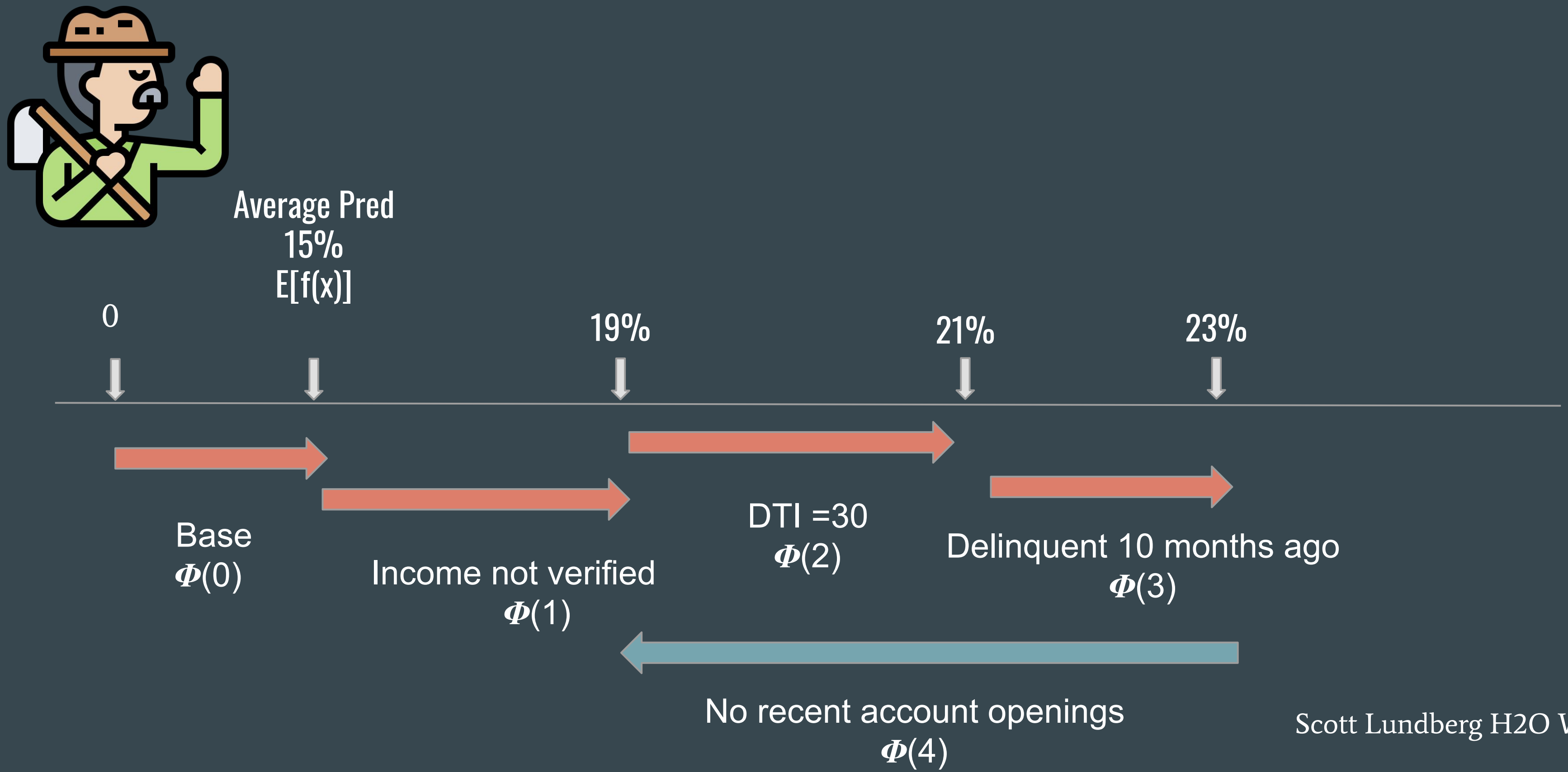


Scott Lundberg H2O World

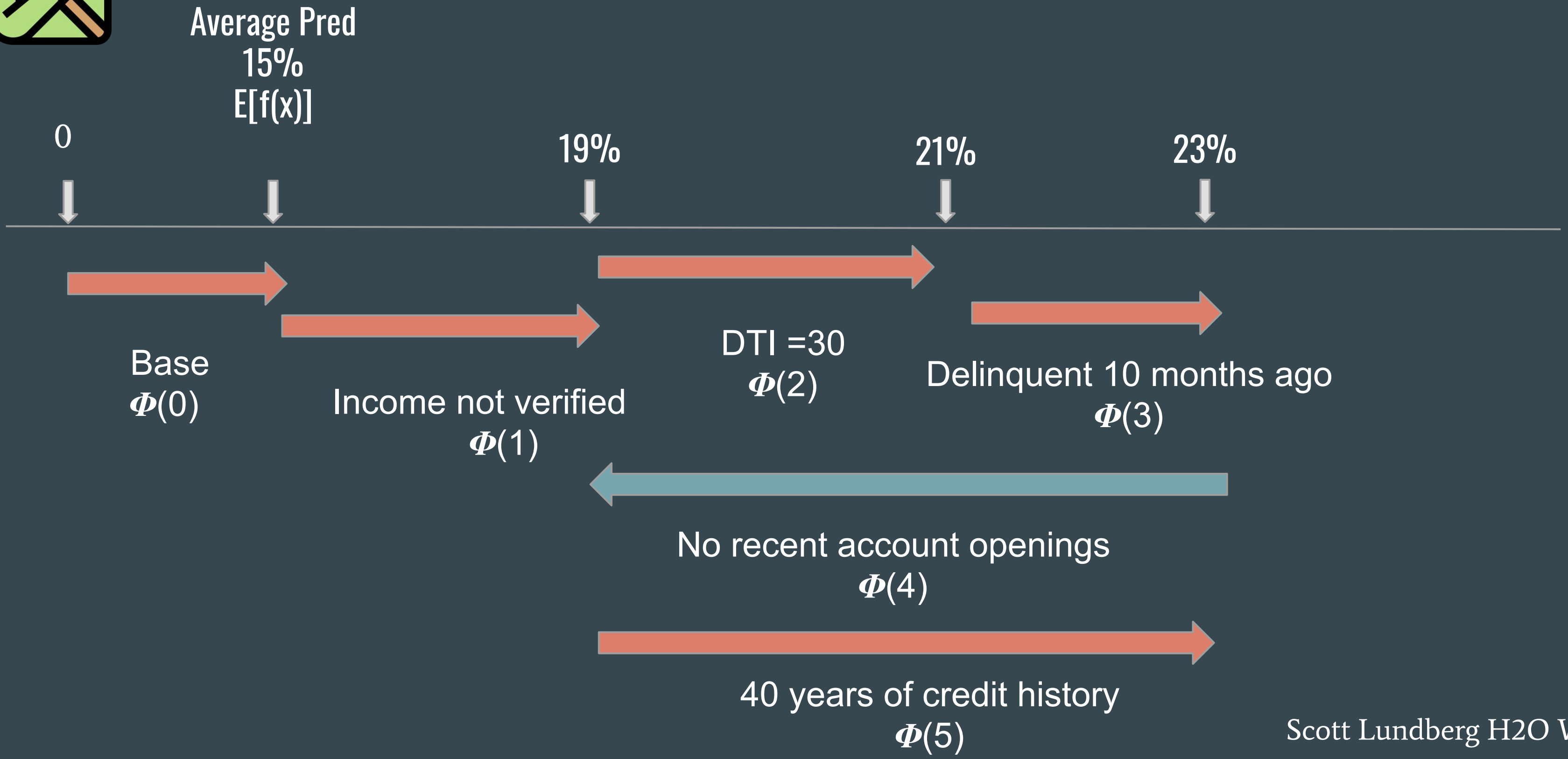


Scott Lundberg H2O World

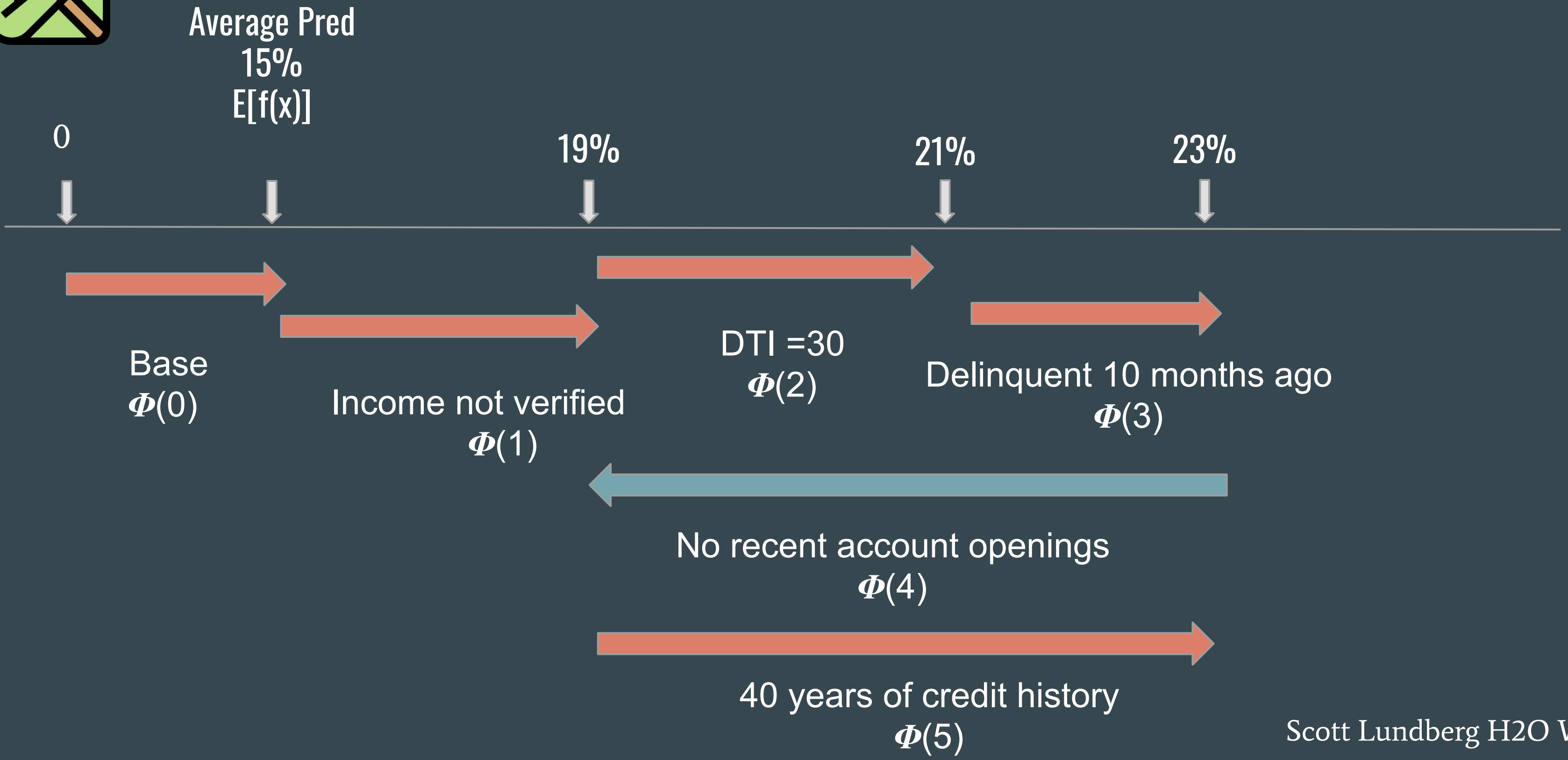

Average Pred
15%
E[f(x)]
0
19%
21%
23%
Base
Φ(0)
Income not verified
Φ(1)
DTI =30
Φ(2)
Delinquent 10 months ago
Φ(3)



Scott Lundberg H2O World

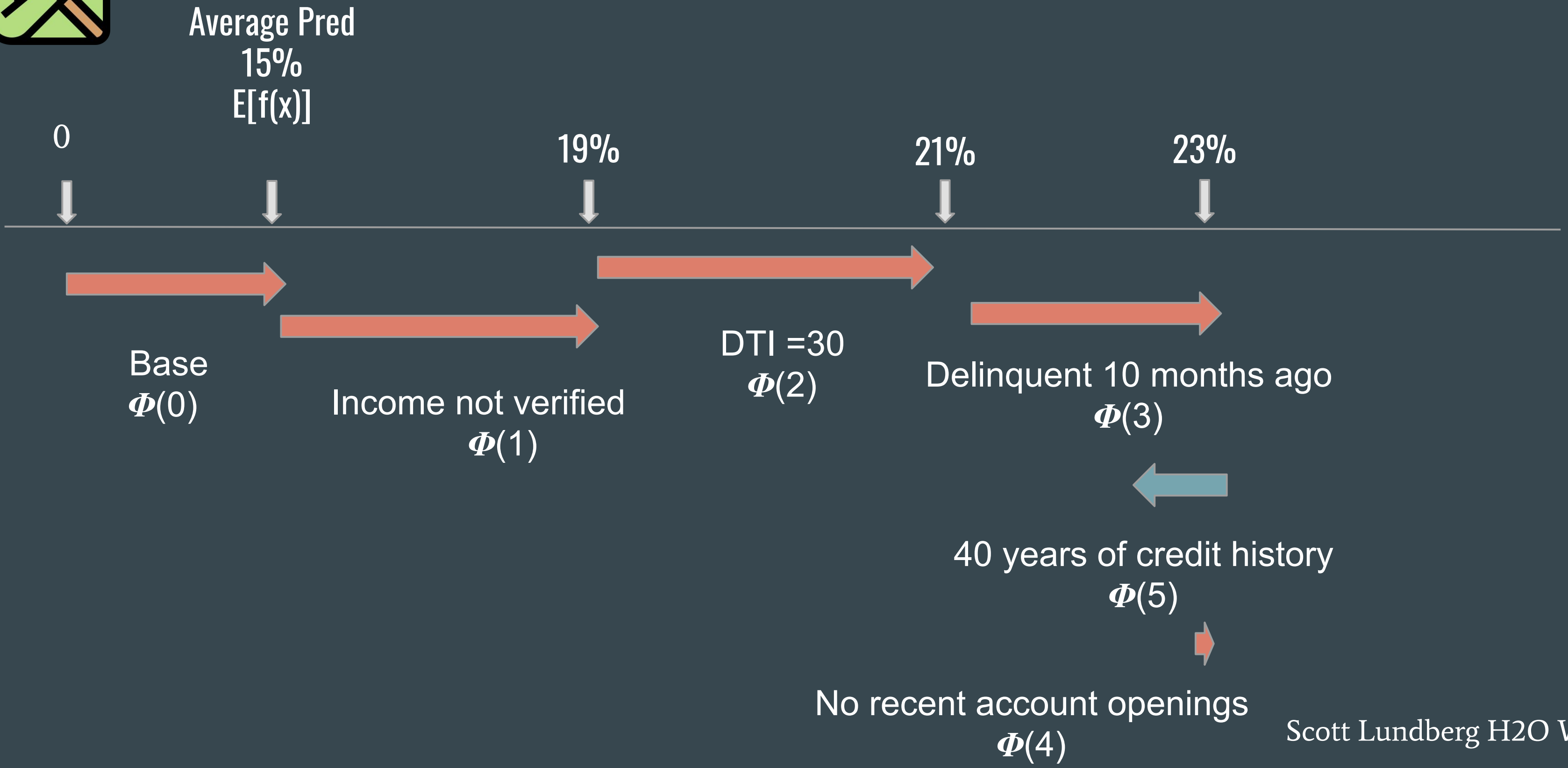

Average Pred
15%
E[f(x)]
0
19%
21%
23%
Base
$\Phi(0)$
Income not verified
$\Phi(1)$
DTI =30
$\Phi(2)$
Delinquent 10 months ago
$\Phi(3)$
No recent account openings
$\Phi(4)$
40 years of credit history
$\Phi(5)$

What about the order?



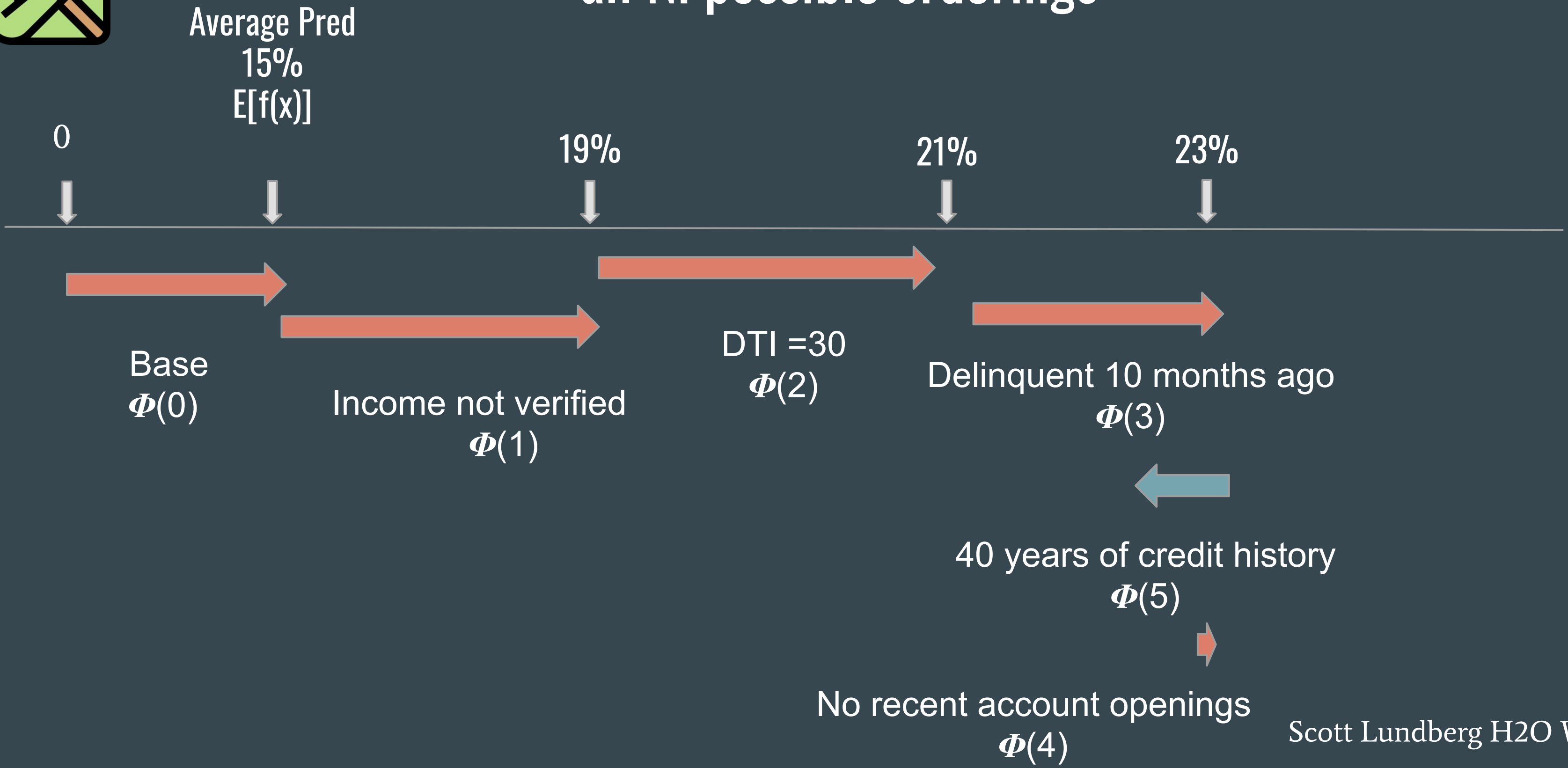
Scott Lundberg H2O World

What about the order?



Scott Lundberg H2O World

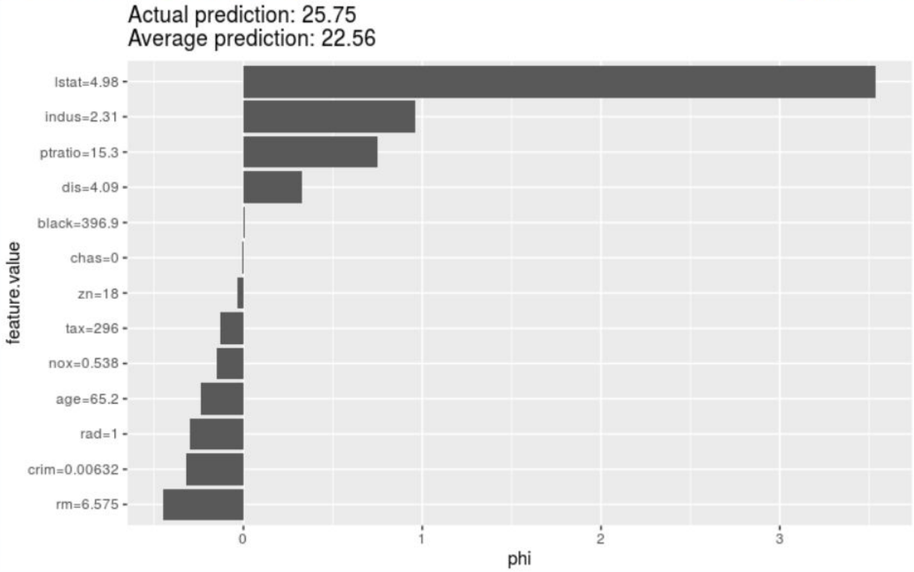
Shapley values results from averaging over all N! possible orderings



Scott Lundberg H2O World

Examples & Interpretation

```r
# First we fit a machine learning model on the Boston housing data

set.seed(42)
library("iml")
library("randomForest")

data("Boston", package = "MASS")
rf =randomForest(medv ~ ., data = Boston, ntree = 50)
X = Boston[-which(names(Boston) == "medv")]
mod = Predictor$new(rf, data = X)

# Then we explain the first instance of the dataset with the Shapley method:

x.interest = X[1,]
shapley = Shapley$new(mod, x.interest = x.interest)

# plot

plot(shapley)
```



SHAP (Shapley Additive Explanations)

KernelSHAP

An alternative, kernel-based estimation approach for Shapley values inspired by local surrogate models

TreeSHAP

An efficient estimation approach for tree-based models

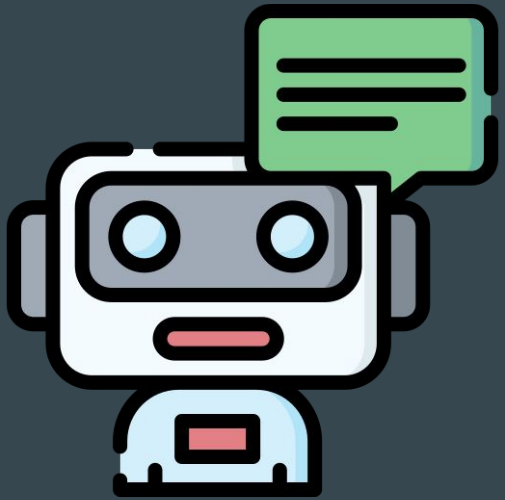
SHAP comes with many global interpretation methods based on aggregations of Shapley values

The future of interpretability

The focus will be on model-agnostic interpretability tools

Robots and programs will explain themselves

Ethical Issues



Resource

Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv Preprint arXiv:1706.07269. (2017). [2]

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. “Examples are not enough, learn to criticize! Criticism for interpretability.” Advances in Neural Information Processing Systems (2016). [3]

Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning,” no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017)

Package iml

https://www.youtube.com/watch?v=ngOBhhINWb8

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions.” Advances in Neural Information Processing Systems. 2017.

Resource

Icons:

Icons made by Smashicons from www.flaticon.com
Icons made by Freepik from www.flaticon.com
Icons made by monkik from www.flaticon.com
Icons made by DinosoftLabs from www.flaticon.com
Icons made by Vitaly Gorbachev from www.flaticon.com
Icons made by Flat Icons from www.flaticon.com
Icons made by photo3idea_studio from www.flaticon.com
Icons made by srip from www.flaticon.com
Icons made by Eucalyp from www.flaticon.com
Icons made by ultimatearm from www.flaticon.com

Questions?