Jennifer Nguyen, Lead Data Scientist
Canadian Analytics COE

Taking a peek under the hood

Interpreting black box models

Sun Life

Imagine…

You're asked to build a mod

Why interpretability matters

REASONS FOR INTERPRETABILITY

- Justify a prediction
- Instill transparency
- Improve your model

Performance vs. Interpretability

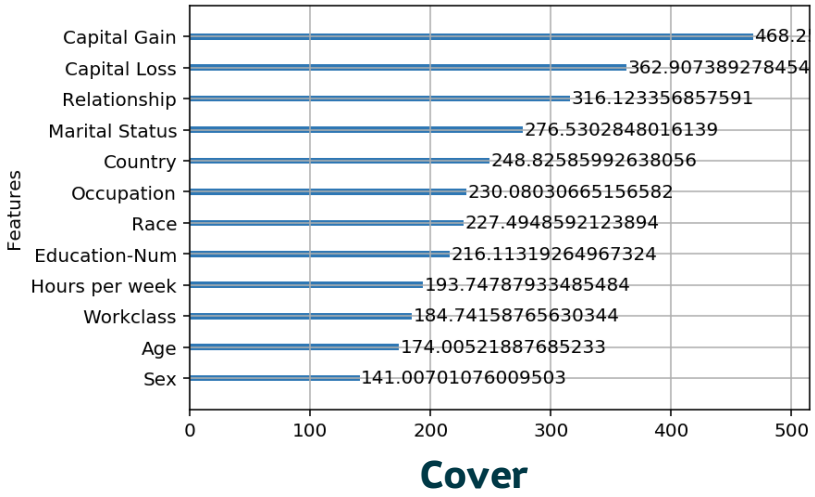
WHICH DO YOU CHOOSE?

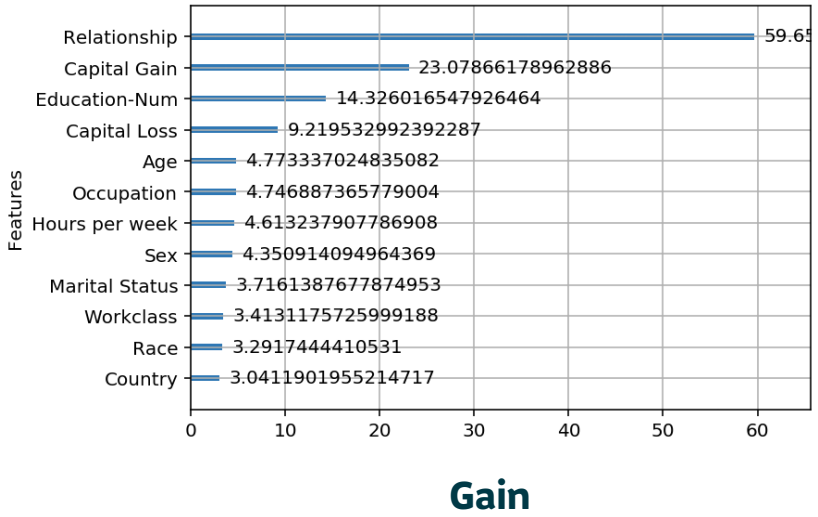
Performance

Interpretability

Existing methods are inconsistent

COMMON HEURISTICS





Source: Scott Lundberg, 2018

Solution: Build a model of the model

ADDITIVE FEATURE ATTRIBUTION MODELS

Let *f(x)* represent the original model and *g(x′)* be the corresponding explanation model

$$g(x') = \underbrace{\phi_0 + \sum_{i=1}^{M} \phi_i x'_i}_{\text{linear}}$$

Where x′ is a simplified version of *x*, ϕ_i's represents the feature importance and M is the number of features.

Desirable properties of g(x')

MISSING IN ACTION

If a feature is not present, then its importance should be zero.

$$x'_i = 0 \Rightarrow \phi_i = 0$$

M.I.A.

Suppose LeBron James is benched for the entire game, should he get any credit if his team wins?

Desirable properties of g(x')

LOCALLY ACCURATE

The explanation model should be equal to the original model for a given input *x*

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

Locally accurate actresses

Katy Perry and Zoey Deschanel have similar features, but obviously they are not the same person.

Desirable properties of g(x')

CONSISTENT

If a feature is at least as useful in a different model, its importance shouldn't decrease

Team	WAR	Salary
Athletics	6.9	$0.5M
Blue Jays	8.8	$4.3M

Consistent performance

Josh Donaldson's performance improved with the Blue Jays; and his consistency was reflected in his salary.

Source: Neil Payne, 2015, FiveThirtyEight

Ways to find g(x')

In other words, how do we solve for ϕ?

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

ADDITIVE FEATURE ATTRIBUTION MODELS

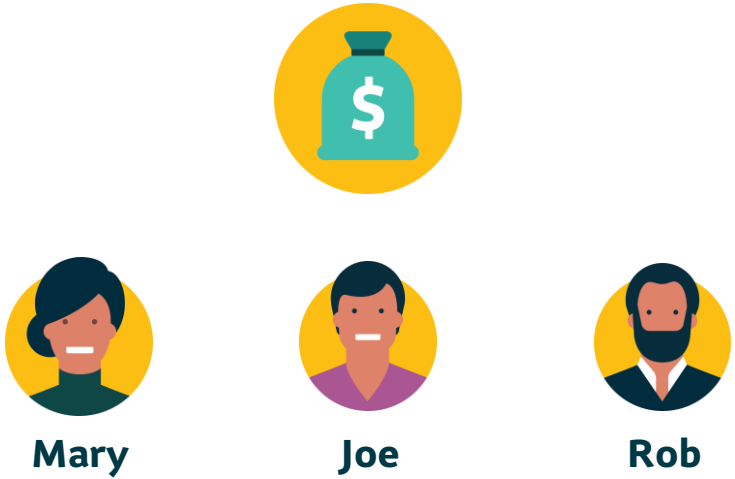
- LIME
- DeepLIFT
- Shapley Additive Explanations (SHAP)

ONLY SHAP MEETS ALL THREE PROPERTIES

SHAP origins

SHAP is inspired by a cooperative game theory concept: Shapley Values

Suppose a group of advisors work together to sell insurance. How should they share their profits?



Shapley Values

Shapley values assigns credit by calculating the average marginal contribution (MC) of each player over **all possible orderings** of the team.

Individual Shapley Values.

Mary

4.5

Joe

5

Rob

5.5

TOY EXAMPLE

Assume each sub-team contributes the following:

f(Mary) = 4 f(Joe) = 4 f(Rob) = 4

f(Mary, Joe) = 9 f(Mary, Rob) = 10 f(Joe, Rob) = 11

f(Mary, Joe, Rob) = 15

Order of Advisors	Mary's Marginal Contribution
Mary, Joe, Rob	$f(M)-f(\emptyset) = 4$
Mary, Rob, Joe	$f(M)-f(\emptyset) = 4$
Joe, Mary, Rob	$f(M,J)-f(J) = 9-4 = 5$
Joe, Rob, Mary	$f(M,J,R)-f(J,R) = 15-11 = 4$
Rob, Mary, Joe	$f(M,R)-f(R) = 10-4 = 6$
Rob, Joe, Mary	$f(M,J,R)-f(J,R) = 15-11 = 4$
Average MC	**4.5**

SHAP (SHapley Additive exPlanations)

SHAPLEY VALUES FOR FEATURES

Average over all feature orderings

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} [\, f_x(z') - f_x(z' \backslash i)]$$

Marginal contribution of feature *i*

Where z' is a binary feature vector and z'\i indicates absence of feature i

SHAP: Tips

What should I know when using SHAP?

TIPS

- Only implemented in Python
- Can be used for tree-based, kernel-based and neural network models
- Computationally intensive – sample your data, especially when calculating interaction values

RESOURCES

- SHAP Github repo with more examples
- SHAP paper by Lundberg et al.

TAKEAWAYS

- No longer need to compromise interpretability over performance
- Use a simple model to interpret a complex model (e.g., LIME, SHAP)
- SHAP provides a model-agnostic framework to measure feature importance, is locally accurate and consistent

Q&A

REACH OUT

www.jennguyen.ca

LinkedIn linkedin.com/in/jenguyen

Life's brighter under the sun