Shapley values

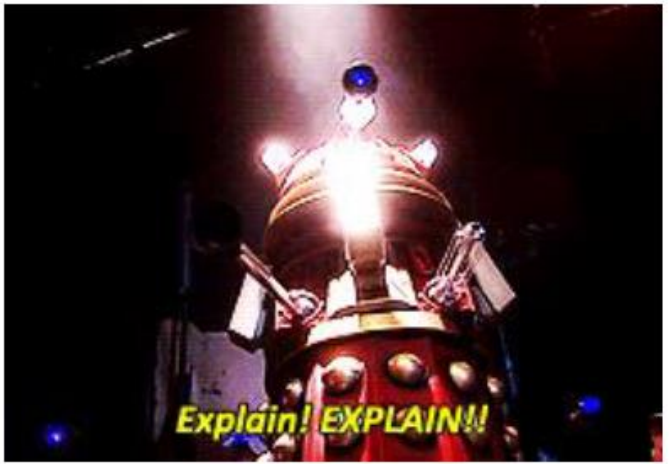
Aprendizaje Automático

Juan David Martínez
jdmartinev@eafit.edu.co

Agenda

- ML interpretable
- SHapley Additive exPlanations

Estado actual de ML



Usos

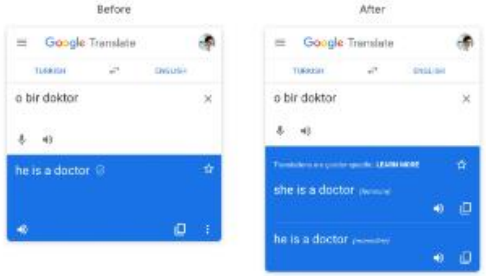
https://www.tesla.com/videos/autopilot-self-driving-hardware-neighborhood-long

NYPost

MIT Technology Review

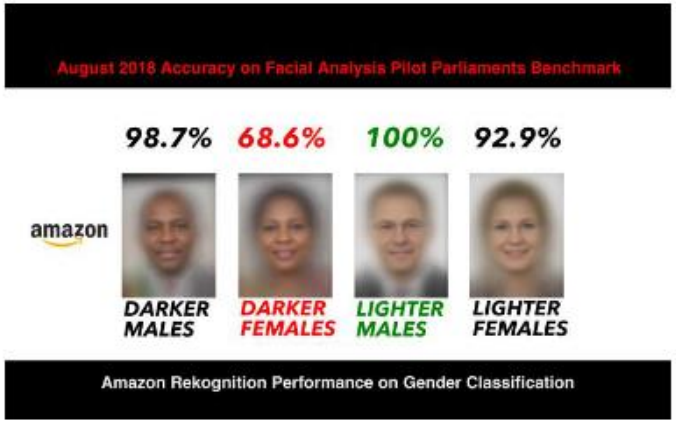
DeepMind

DeepMind

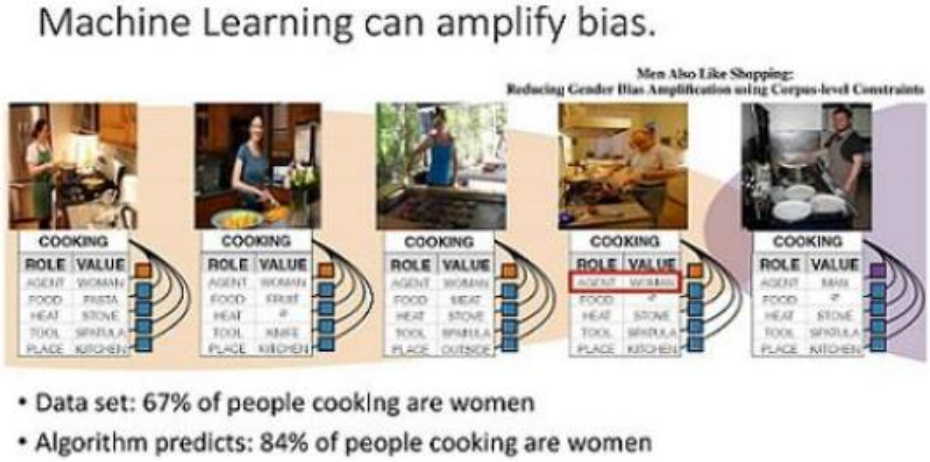


Problemas

Sesgos de los algoritmos



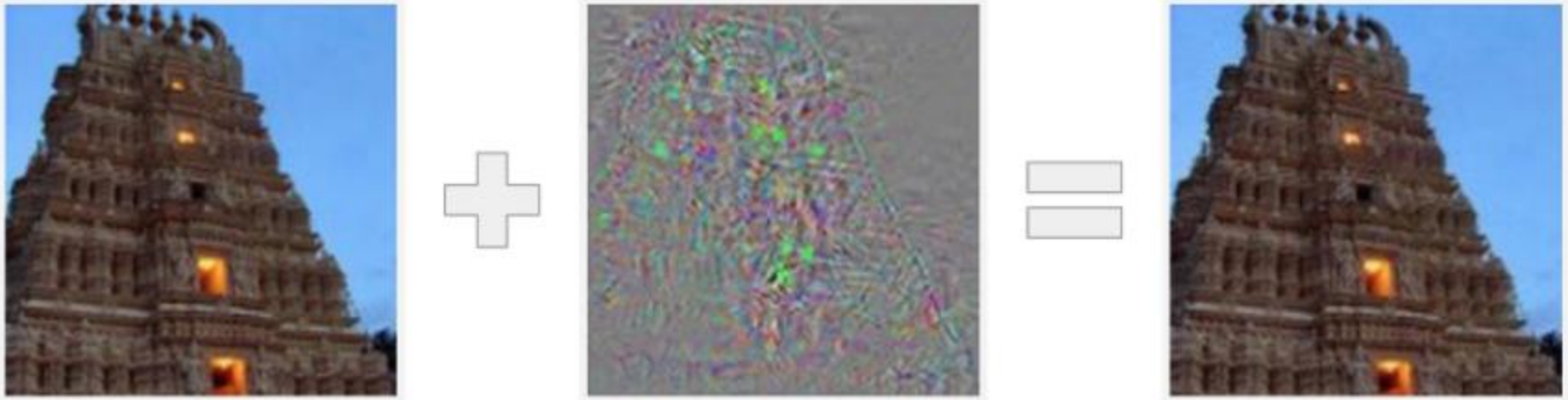
https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced



https://www.infoq.com/presentations/unconscious-bias-machine-learning/

Problemas

Ejemplos adversarios



Original image

Temple (97%)

Perturbations

Adversarial example

Ostrich (98%)

ML interpretable

Tenemos varios problemas:

- No confiamos en los modelos
- No sabemos qué pasa en casos extremos
- Los errores pueden ser costosos/nocivos
- ¿Los modelos cometen errores similares a los de los humanos?
- ¿Cómo cambiamos el modelo si no da los resultados esperados?

Una forma de lidiar con estos problemas es a través de la interpretabilidad

Shapley Additve exPlanations - SHAP

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

Shapley Additve exPlanations - SHAP

$$\phi_i = \sum_{S \subseteq \{1,\ldots,p\}\{i\}} \frac{|S|!(p-|S|-1)!}{p!} [val(S \cup \{i\}) - val(S)]$$

Shapley values para juego de 2 participantes

kaggle

First	Second	Third
$10,000	$7,500	$5,000

P1 P2

$$C_{12} = 10{,}000$$
$$C_{1} = 7{,}500$$
$$C_{2} = 5{,}000$$
$$C_{0} = 0$$

Valores de las coaliciones

Contribución marginal esperada = Shapley values

P1

$C_{12} = 10{,}000$
$C_1 = 7{,}500$
$C_2 = 5{,}000$
$C_0 = 0$

P2

$C_{12} - C_2 = 5{,}000$
$C_1 - C_0 = 7{,}500$

$(5{,}000+7{,}500)/2 = \mathbf{\$6{,}250}$

$C_{12} - C_1 = 2500$
$C_2 - C_0 = 5000$

$(2500+5000)/2 = \mathbf{\$3{,}750}$

Incremento en el valor de una coalición debido al ingreso de un participante a dicha coalición

Shapley values para juego de 3 participantes

$C_{123} = 10{,}000$
$C_{0} = 0$

$C_{12} = 7{,}500$
$C_{13} = 7{,}500$
$C_{23} = 5{,}000$

$C_{1} = 5{,}000$
$C_{2} = 5{,}000$
$C_{3} = 0$

Shapley values para juego de 3 participantes

- Contribución marginal esperada P1

$$C_0 \rightarrow C_1 \rightarrow C_{12} \rightarrow C_{123}$$
$$C_0 \rightarrow C_1 \rightarrow C_{13} \rightarrow C_{132}$$
$$C_0 \rightarrow C_2 \rightarrow C_{21} \rightarrow C_{213}$$
$$C_0 \rightarrow C_2 \rightarrow C_{23} \rightarrow C_{231}$$
$$C_0 \rightarrow C_3 \rightarrow C_{31} \rightarrow C_{312}$$
$$C_0 \rightarrow C_3 \rightarrow C_{32} \rightarrow C_{321}$$

$C_{123} - C_{23} = 5{,}000$
$C_{12} - C_2 = 2{,}500$
$C_{13} - C_3 = 7{,}500$
$C_1 - C_0 = 5{,}000$

$5{,}000*(⅓) + 2{,}500*(⅙)$
$+ 7{,}500*(⅙) + 5{,}000*(⅓)$
$= \mathbf{\$5{,}000}$

Shapley values para juego de 3 participantes

$C_0 \rightarrow C_1 \rightarrow C_{12} \rightarrow C_{123}$

$C_0 \rightarrow C_1 \rightarrow C_{13} \rightarrow C_{132}$

$C_0 \rightarrow C_2 \rightarrow C_{21} \rightarrow C_{213}$

$C_0 \rightarrow C_2 \rightarrow C_{23} \rightarrow C_{231}$

$C_0 \rightarrow C_3 \rightarrow C_{31} \rightarrow C_{312}$

$C_0 \rightarrow C_3 \rightarrow C_{32} \rightarrow C_{321}$

de formas en las que se puede crear una coalición de 3 participantes

$3!$

de formas en las que P1 se puede unir a una coalición de P2 y P3

$2!*1!$

Shapley Additve exPlanations - SHAP

$$\phi_i = \sum_{S \subseteq \{1,\ldots,p\}\{i\}} \underbrace{\frac{|S|!(p-|S|-1)!}{p!}}_{\text{Weight}} \underbrace{[val(S \cup \{i\}) - val(S)]}_{\text{Marginal contribution of player i to coalition S}}$$

$|S|!$ $\Rightarrow$ *number of ways players can join S before player i*

$(p - |S| - 1)!$ $\Rightarrow$ *number of ways players can join coalition after player i joins*

$p!$ $\Rightarrow$ *number of ways to form coalition of p players*

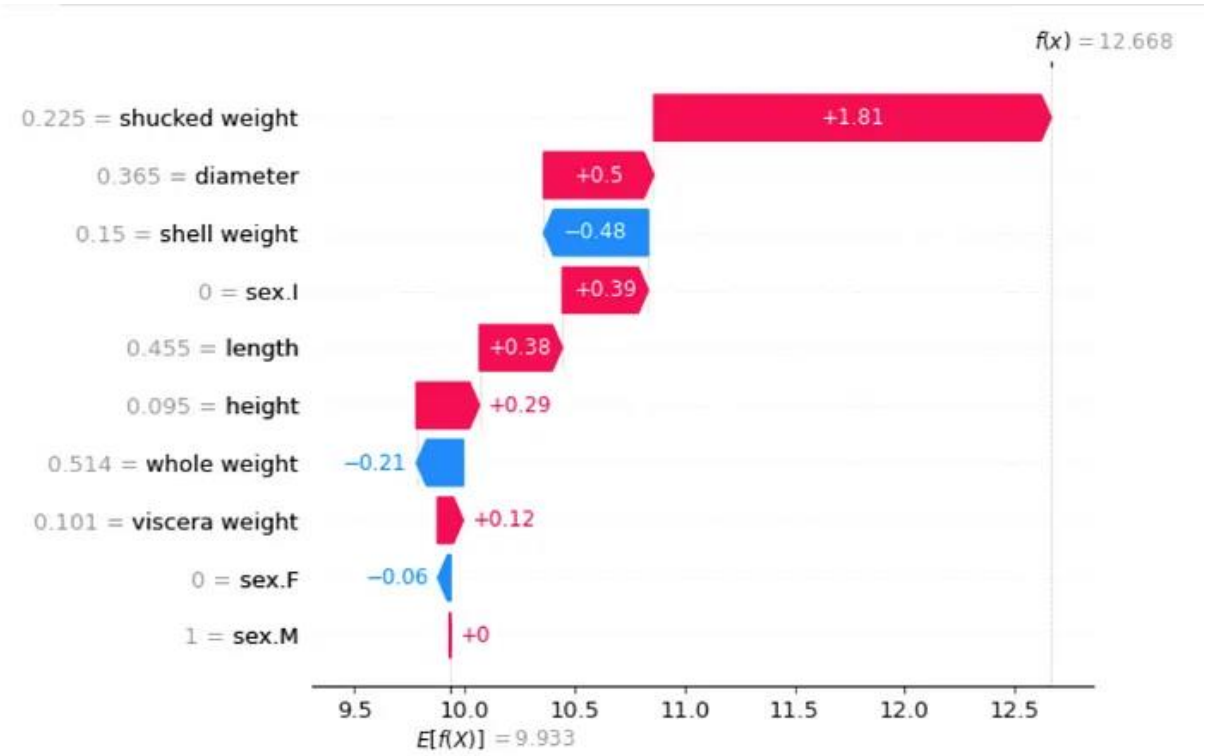
Shapley values - propiedades

Symmetry Two players are considered interchangeable if they make the same contributions to all coalitions. If two players are interchangeable then they must be given an equal share of the game's total value.

Null player property If a player makes zero marginal contribution to all coalitions then they get none of the total value.

Additivity If we combine two games, then a player's overall contribution is the sum of the contributions for the two individual games. This axiom makes the assumption that any games played are independent.

Shapley values for Machine Learning



Shapley values for Machine Learning

$$val_x(S) = \int f(x_1,\ldots,x_p)\,dP_{x\notin S}$$

$f(x_1,\ldots,x_p)$ ⟶ Predicción del modelo

p ⟶ Número de características

S ⟶ Coalición de características

El valor de la coalición S es la predicción del modelo marginalizada sobre todas las características que no están en S

Shapley values for Machine Learning

Predicción del salario

$$f(x_1, x_2) = 200x_1 + 1000x_2$$

$$age \rightarrow x_1 \in [18, 60]$$

$$degree \rightarrow x_2 \in \{0, 1\}$$

Persona de 20 años con título

$$\begin{aligned} val_x(\{1, 2\}) &= f(20, 1) \\ &= 200(20) + 1000(1) \\ &= 5000 \end{aligned}$$

Shapley values for Machine Learning

$$val_x(\{1\}) = \int f(20, x_2) dP_{x_2}$$

$$= \sum_{i=0}^{1} f(20, i) P(x_2 = i)$$

$$= (200(20) + 1000(0))(0.5) + (200(20) + 1000(1))(0.5)$$

$$= 4500$$

Contribución marginal de x_2 a $S = \{1\}$

$$val_x(\{1, 2\}) - val_x(\{1\}) = 500$$

Shapley values for Machine Learning

Para calcular la contribución marginal de x_2 debemos calcular también la contribución marginal de esta característica a $S = \{\}$. Esto requiere marginalizar sobre la distribución de ambas características.

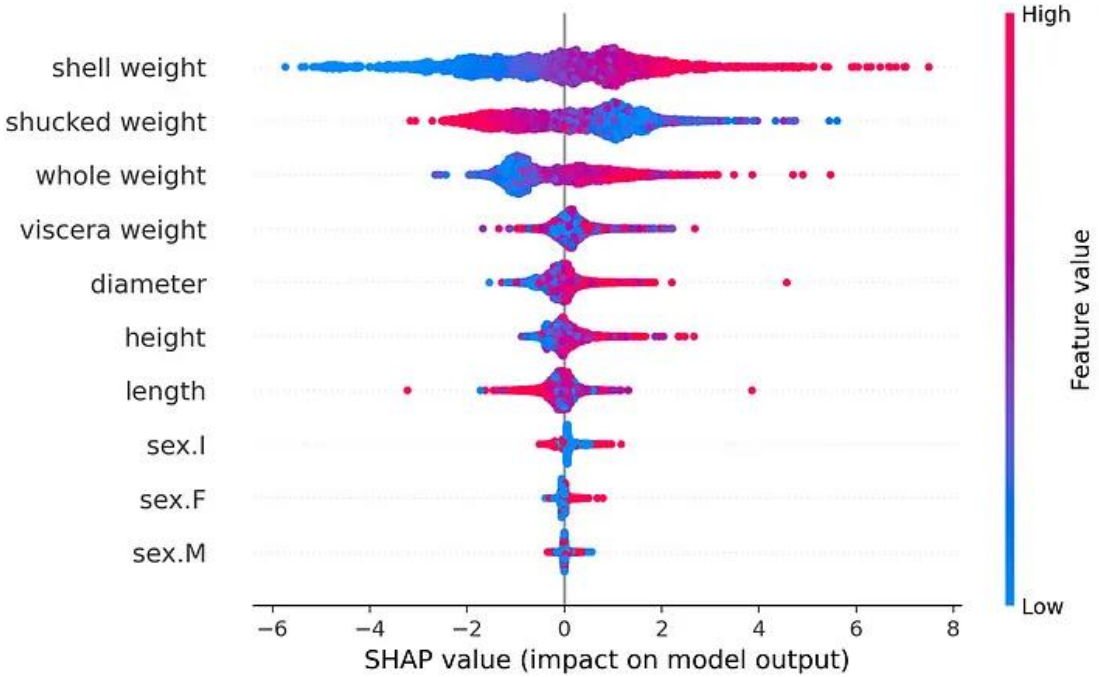
Aproximación de los Shapley values

Monte-Carlo sampling

$$\widehat{\phi}_i = \frac{1}{M} \sum_{m=1}^{M} \left(f(x_{+i}^{m}) - f(x_{-i}^{m}) \right)$$

$$f(x) = \sum_{i=1}^{p} \phi_i + E_X[f(X)]$$

SHAP values - gráficas



Referencias

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

https://towardsdatascience.com/from-shapley-to-shap-understanding-the-math-e7155414213b