Shapley values

Aprendizaje automático

Juan David Martínez
jdmartinev@eafit.edu.co

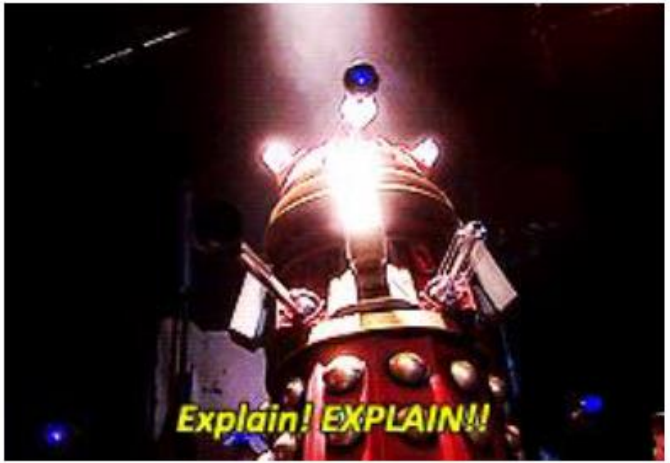
2023

UNIVERSIDAD EAFIT

Agenda

- ML interpretable
- SHapley Additive exPlanations

Estado actual de ML

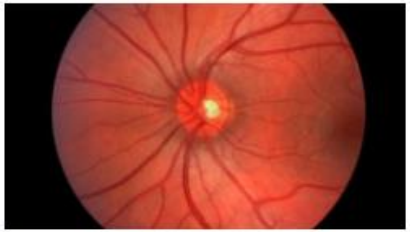


Usos

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

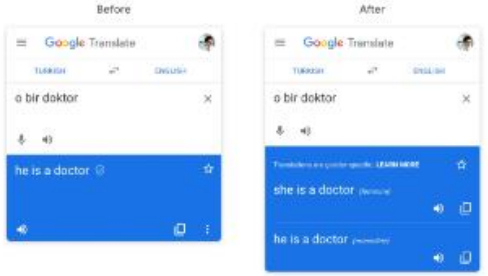
NYPost

MIT Technology Review



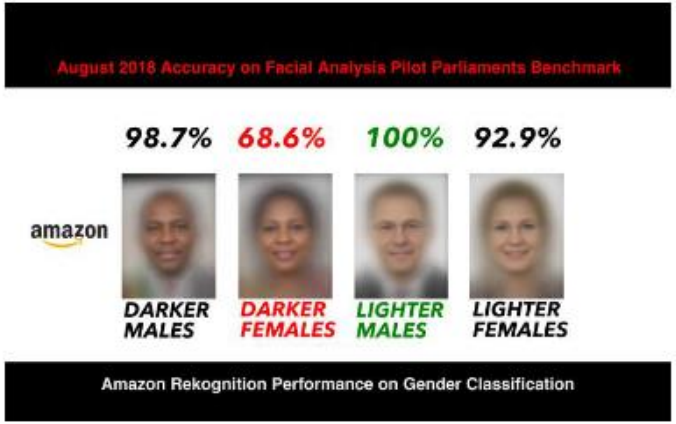
DeepMind

DeepMind

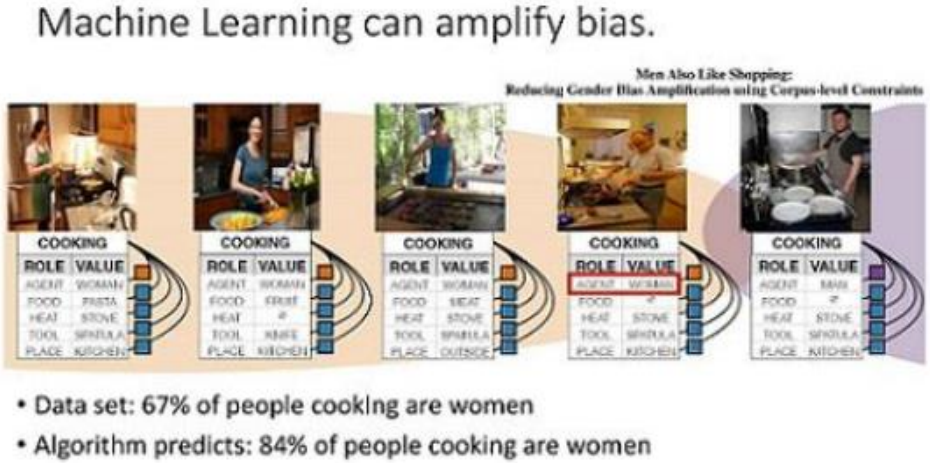


Problemas

Sesgos de los algoritmos



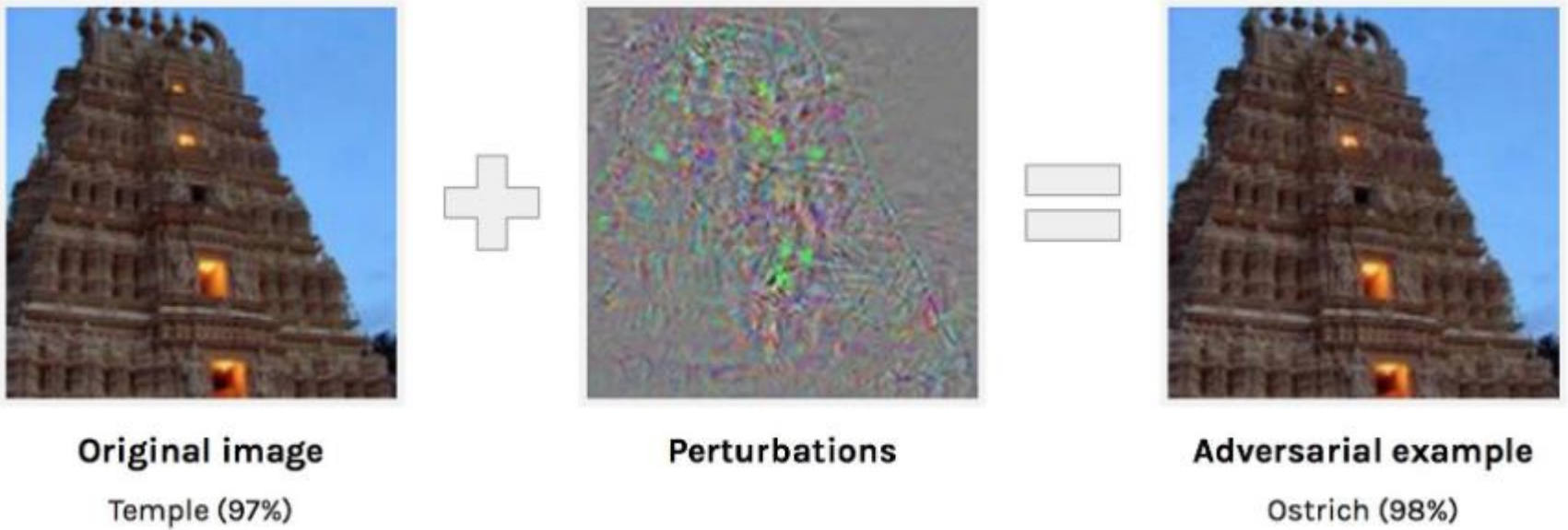
https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced



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

Problemas

Ejemplos adversarios



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

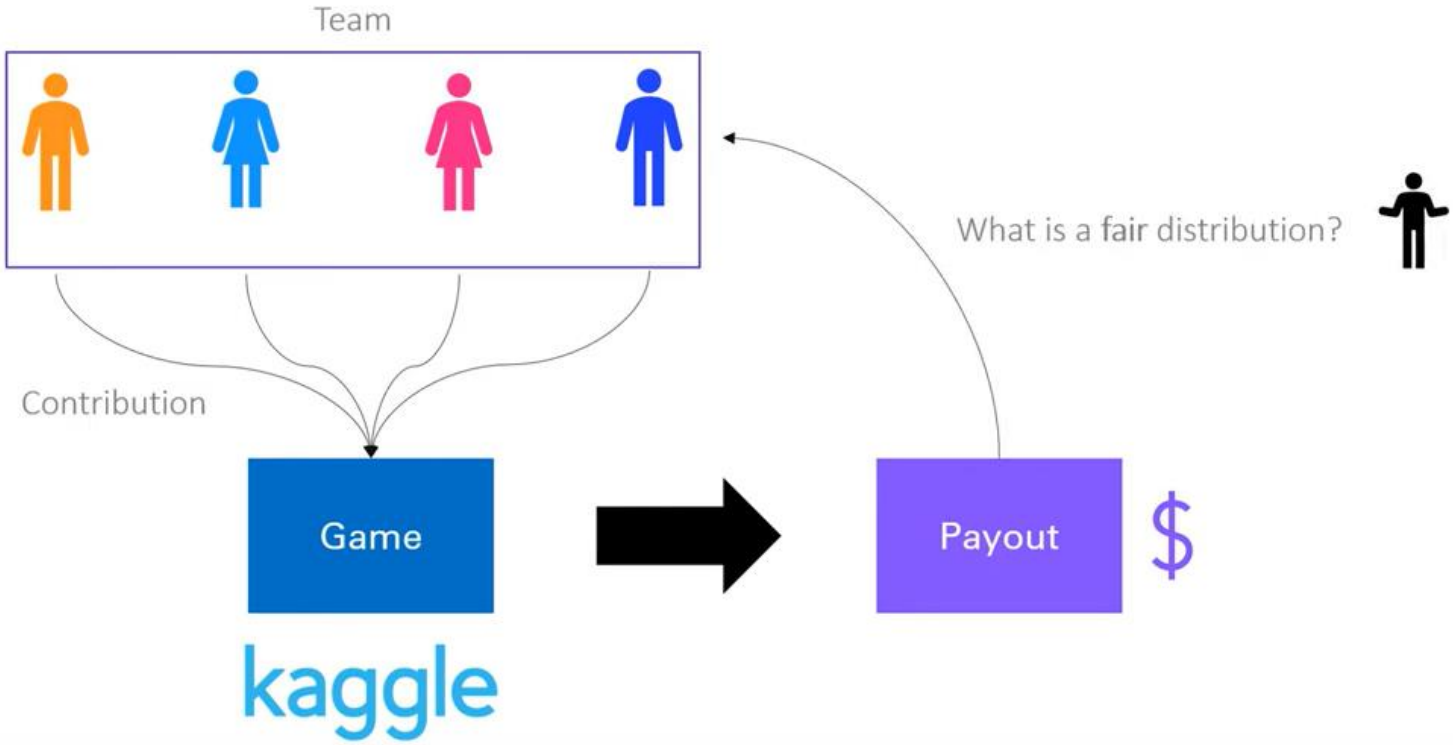
Shapley Additve exPlanations - SHAP

A Unified Approach to Interpreting Model Predictions

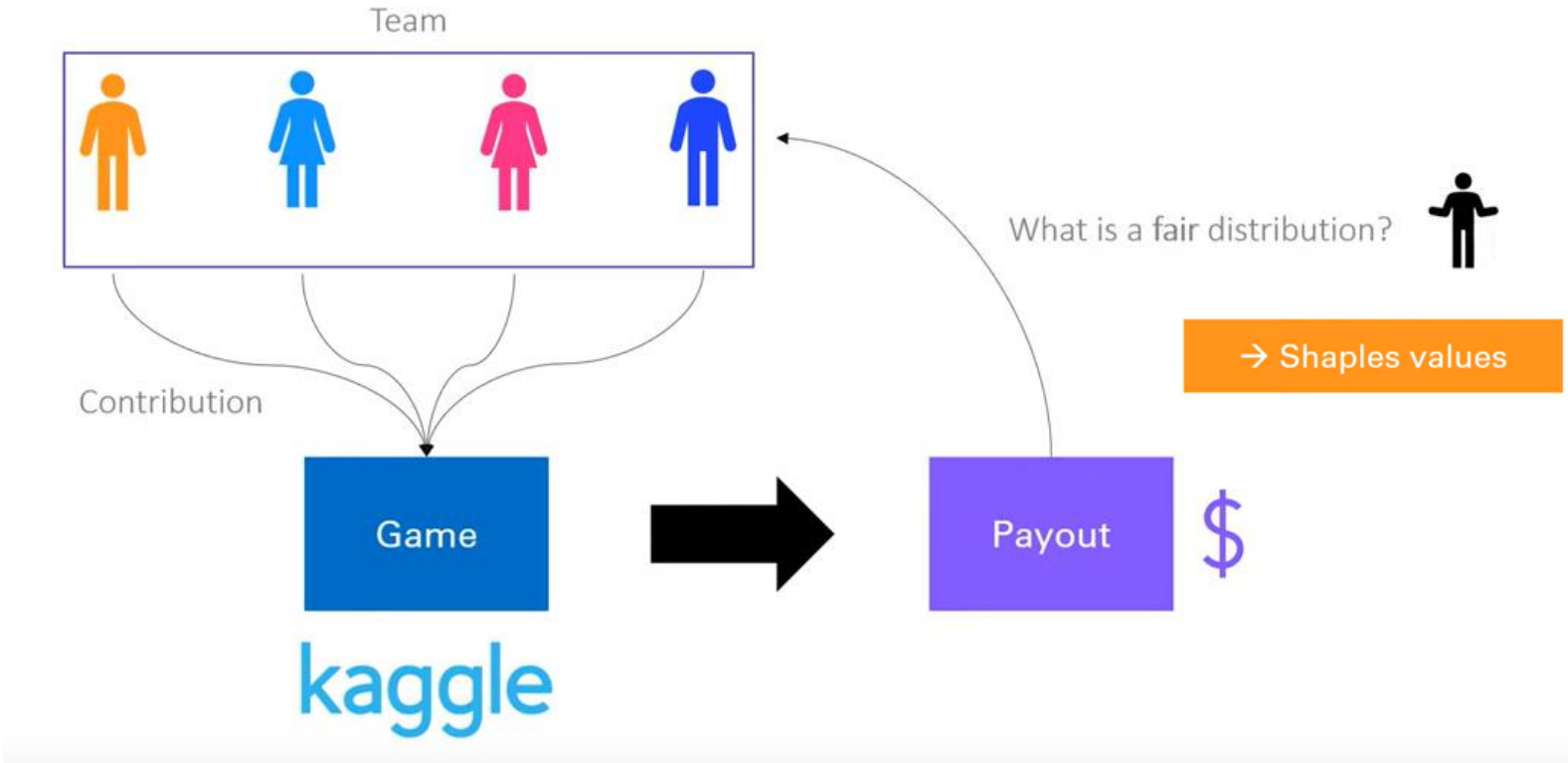
Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

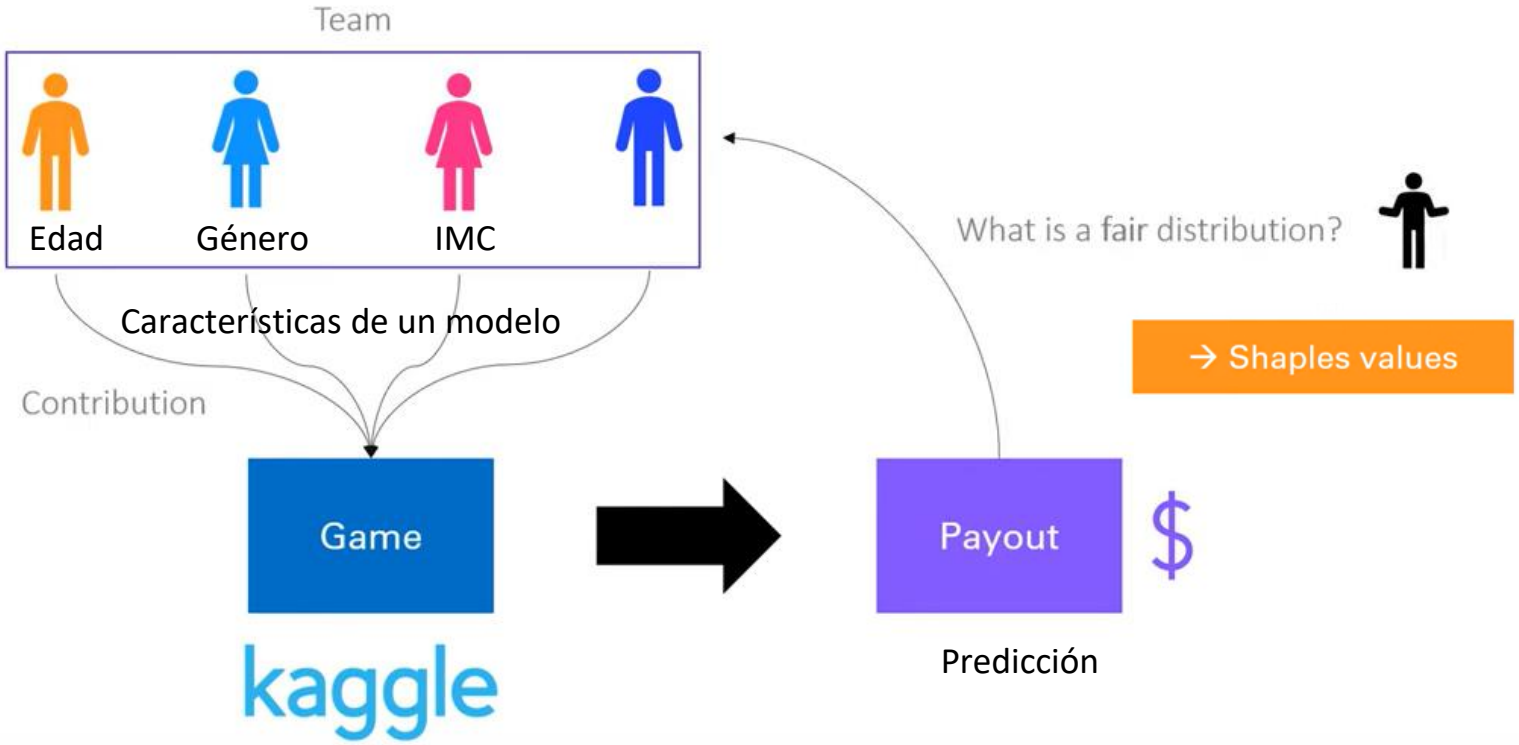
Shapley Additve exPlanations - SHAP



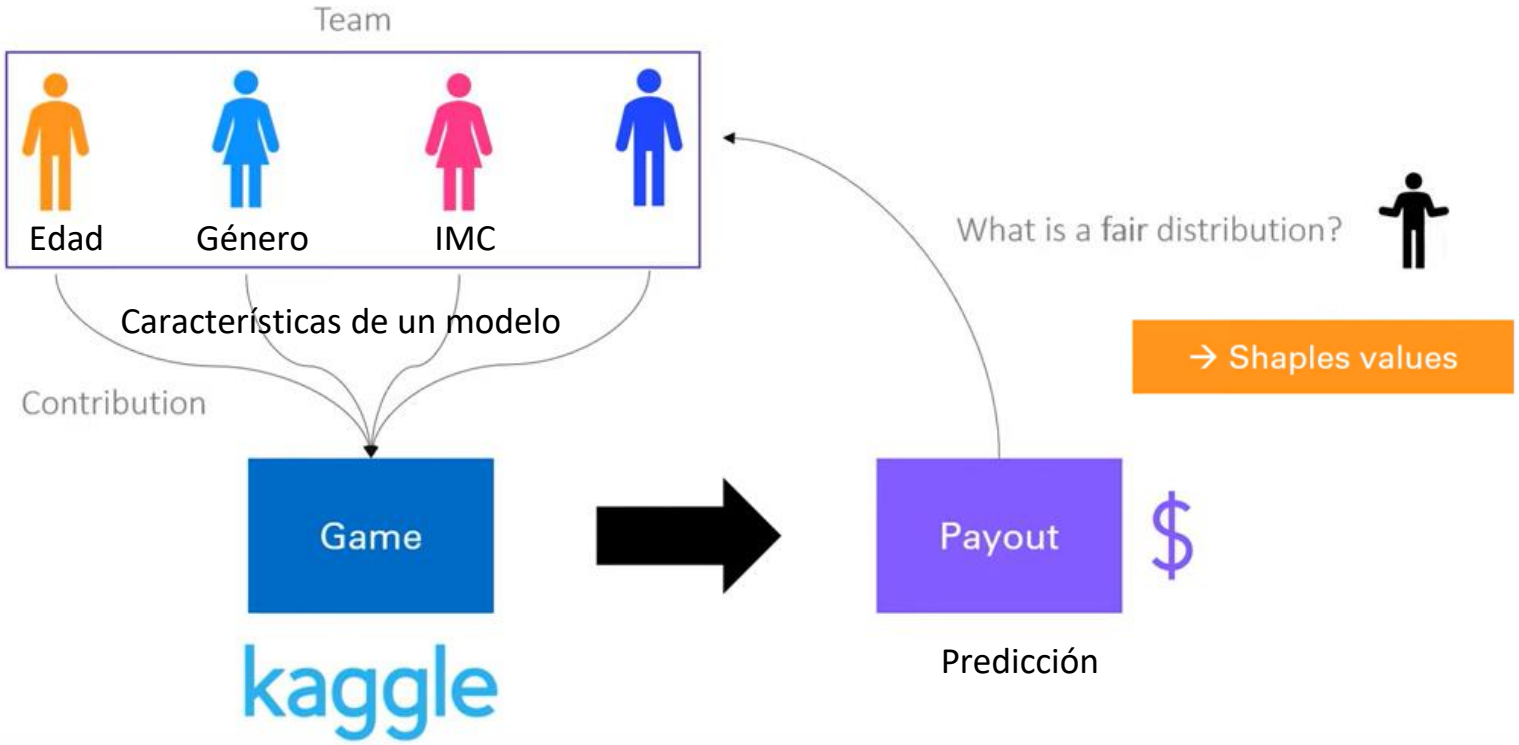
Shapley Additve exPlanations - SHAP



Shapley Additve exPlanations - SHAP

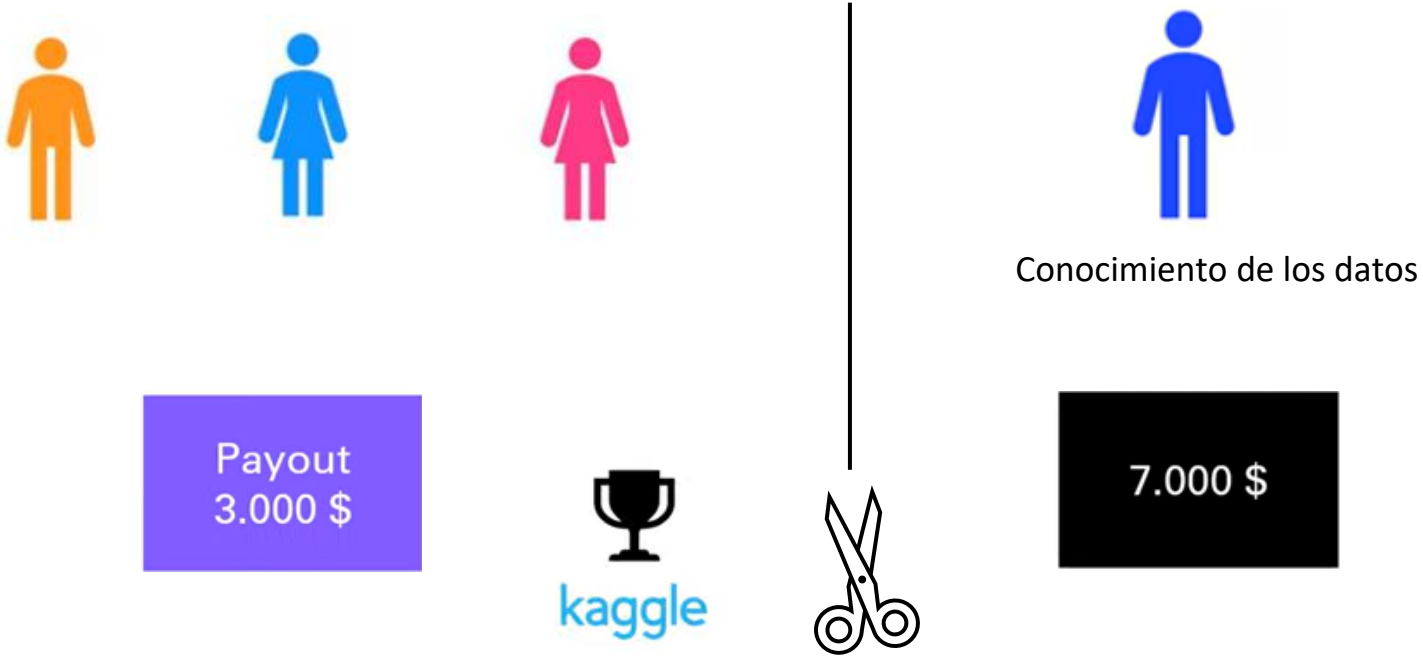


Shapley Additve exPlanations - SHAP



Shapley Additve exPlanations - SHAP

Shapley Additve exPlanations - SHAP



Shapley Additve exPlanations - SHAP

Shapley Additve exPlanations - SHAP

Todas las combinaciones posibles: Contribución marginal de cada participante

Shapley Additve exPlanations - SHAP

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

z':Subconjunto de características

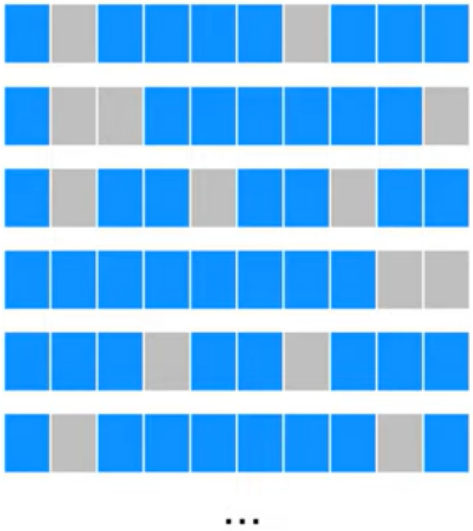
$z'\backslash$i: Subconjunto de características sin la característica i

x'	Age = 56	Gender = F	Body Mass Index = 30	Heart disease = yes	...
z'	Age = 56	Gender = F	Body Mass Index = 30	Heart disease = yes	...

Shapley Additve exPlanations - SHAP

Shapley Additve exPlanations - SHAP

https://github.com/deepfindr/xai-series/blob/master/03_shap.py



...

$2^{10} = 1024$

2^n = total number of subsets of a set