Towards Cotenable and Causal Shapley Feature Explanations

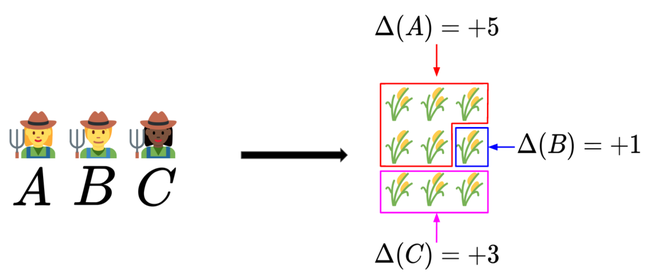
Tony Liu Lyle Ungar

liutony@seas.upenn.edu ungar@cis.upenn.edu

University of Pennsylvania

What are Shapley values?

Shapley values, from cooperative game theory, tell us **how much "credit" should each player in the game get for producing the outcome** by assigning credit based on the **marginal contribution** Δ each player makes when joining the group.



Alice, Bob, and Celine are farmers who produce 9 bushels of wheat when working together. Their Shapley values ϕ are their average Δ over all $3!$ permutations of the group order.

Shapley for feature explanation

farmers $\rightarrow$ features: X_1, X_2, X_3

wheat $\rightarrow$ model output: $f(X_1, X_2, X_3) = Y$

The importance of a feature is **how much Shapley value "credit"** it receives for producing the model output.

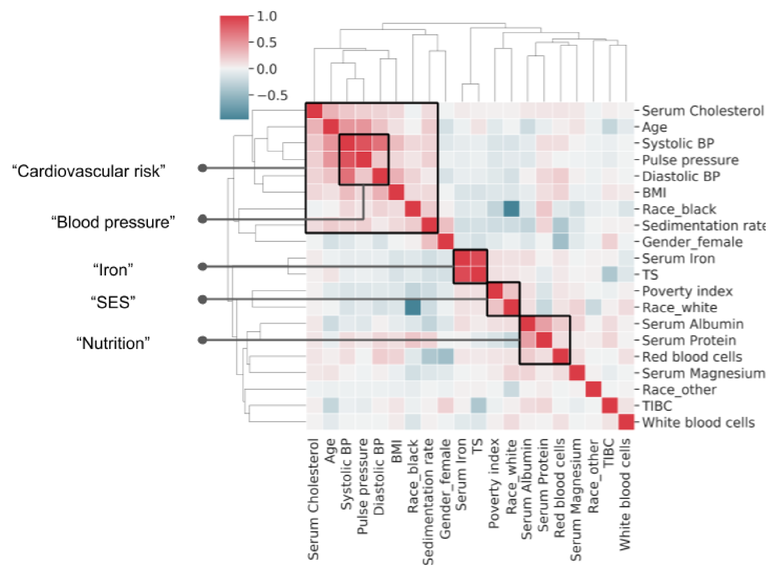
How do we "remove" features from the model to compute Shapley values? Suppose we are computing the Shapley value for group $\{X_1 = x_1\}$:

conditional Shapley: $E_{X_2,X_3|X_1}[f(X_1, X_2, X_3)|X_1 = x_1]$

interventional Shapley: $E_{X_2,X_3}[f(x_1, X_2, X_3)]$

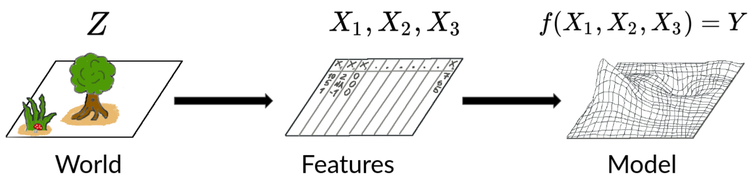
Grouping features

Grouping features increases interpretability and moves closer to satisfying **both cotenability and causality**.

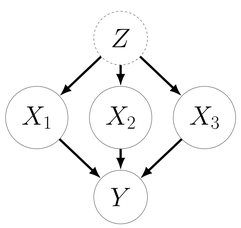


Goals for feature explanation

When interpreting machine learning models, it is important to consider the full process of producing model output [3]:



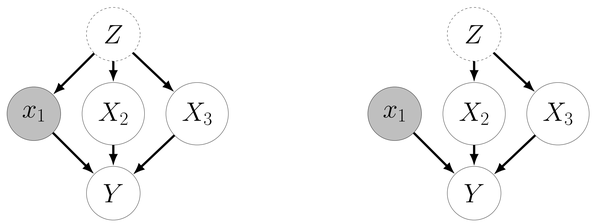
We can visualize the relationship between properties of the world Z, the measured features X_i, and the model output Y graphically:



Intepretability can then be framed as two different goals [1]:

- **Explaining the model**: understanding why the machine learning model makes a prediction.
- **Explaining the world**: understanding a real-world mechanism through the data and model output.

Cotenability and Causality



(a) **Conditional** Shapley is **cotenable** (b) **Interventional** Shapley is **causal**

- Features are often correlated because they have a **shared latent real-world cause** Z: there is a trade-off in breaking or respecting these dependencies.
- **Contenable** explanations respect correlations among features: changes in BMI must also change height or weight. **Explains the world**.
- Model-based **causal** explanations tell us how features intervene on the model: how does *setting* blood pressure to 180 affect predicted mortality? **Explains the model**.

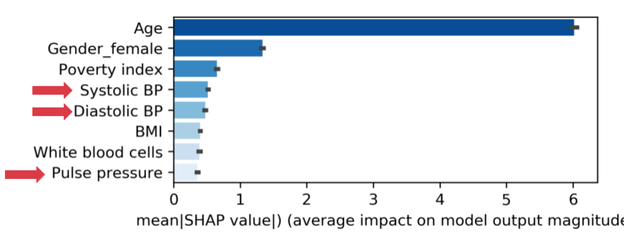
References

[1] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the Model or True to the Data? June 2020.

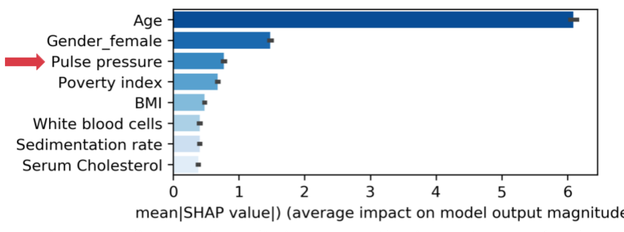
[2] Tony Liu and Lyle Ungar. Towards cotenable and causal shapley feature explanations. *AAAI Workshop on Trustworthy AI for Healthcare*, 2021.

[3] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

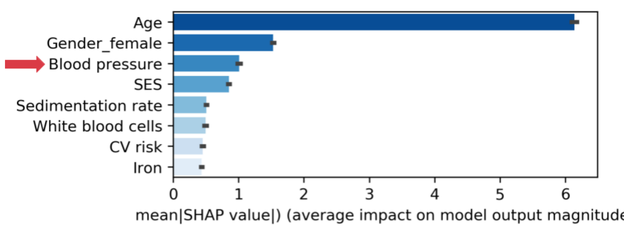
NHANES mortality case study



(a) The importance of blood pressure is spread across systolic, diastolic, and pulse pressure.



(b) Removing systolic and diastolic increases the importance of pulse pressure.



(c) Grouping all blood pressure features increases their relative importance.