Consistent Individualized Feature Attribution for Tree Ensembles

SHAP (SHapley Additive exPlanation) values

https://arxiv.org/abs/1802.03888

https://github.com/slundberg/shap

Supervised ML

	Feature 1	Feature 2	...	Feature M	Target Var.
point 1					0.1
point 2					1.2
...					...
point n					0.6

Some numbers

point n+1					???

Outline

- Motivation
- Feature importance metrics
- What are SHAP values and why are they useful?
- How are SHAP values calculated?
- Outlook

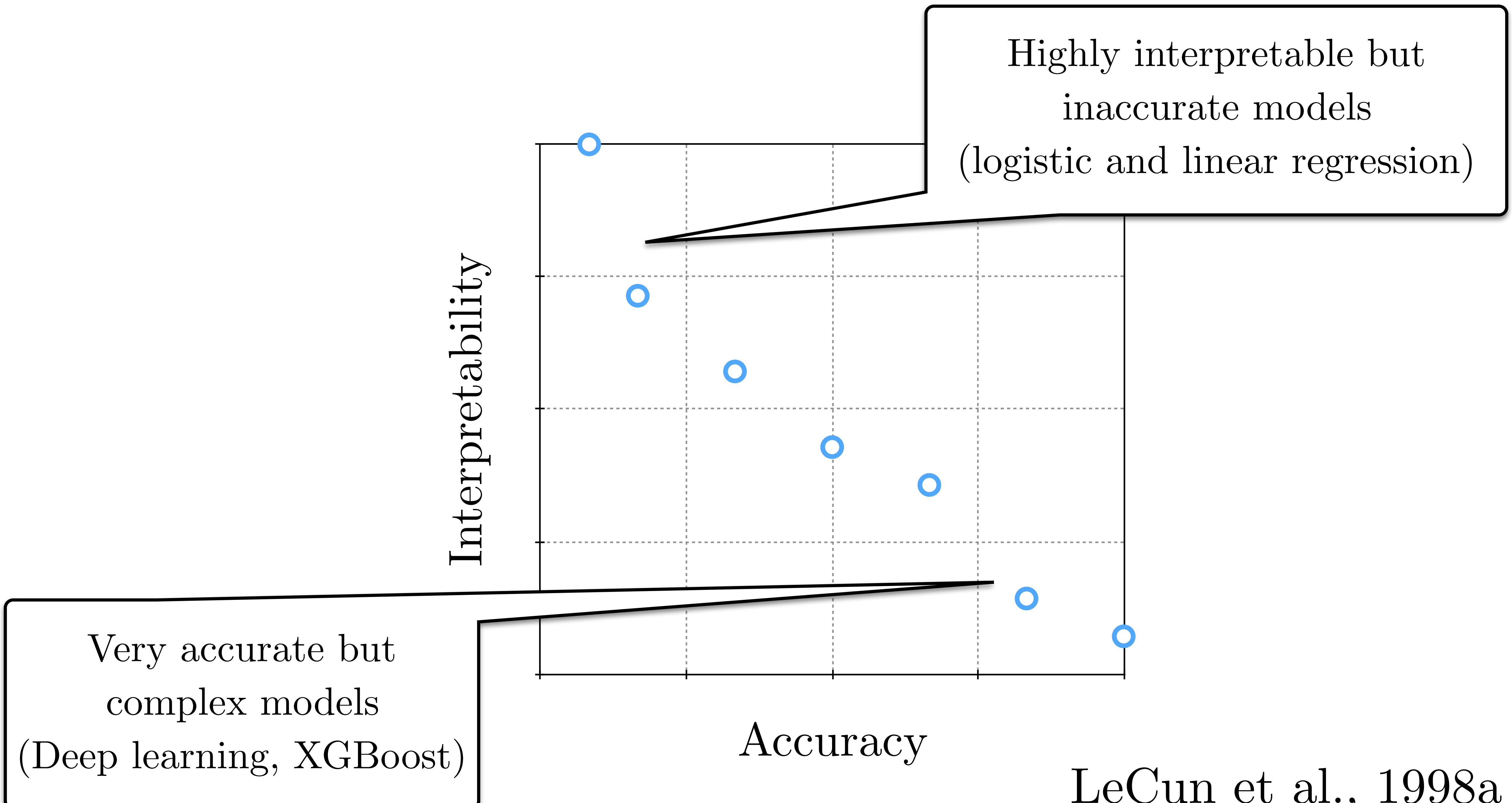
Outline

- **Motivation**
- Feature importance metrics
- What are SHAP values and why are they useful?
- How are SHAP values calculated?
- Outlook

Motivation

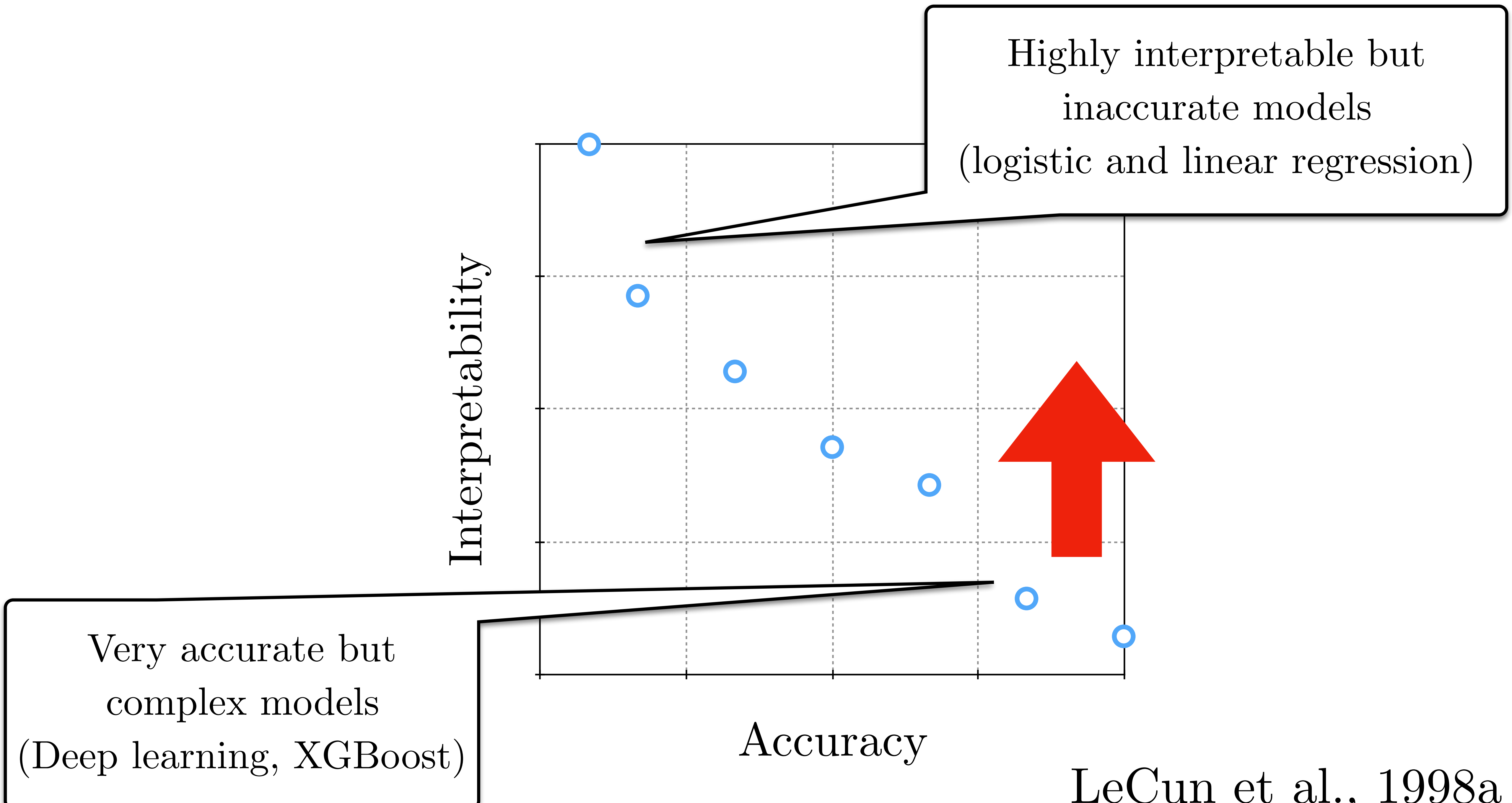
- Supervised ML models: predicted values are not enough
    - your collaborators/users need explanations to build trust and to take proper actions
    - you need explanations to debug the code, generate better features, etc.
- Some models are easier to explain/interpret than others
    - accuracy vs. interpretability compromise

Motivation



LeCun et al., 1998a

Motivation



LeCun et al., 1998a

Outline

- Motivation
- **Feature importance metrics**
- What are SHAP values and why are they useful?
- How are SHAP values calculated?
- Outlook

Feature importance metrics (global)

- Gain: the average training loss reduction while using a feature

- Permutation: permute the feature values in the test set, observe change in the model's error

- Split count (for tree-based methods only): the number of times a feature was used to split on

- Cover (for tree-based methods only): same as split count but weighted by the number of points that go through the split

Feature importance metrics (local)

- LIME: Locally Independent Model-agnostic Explanations*

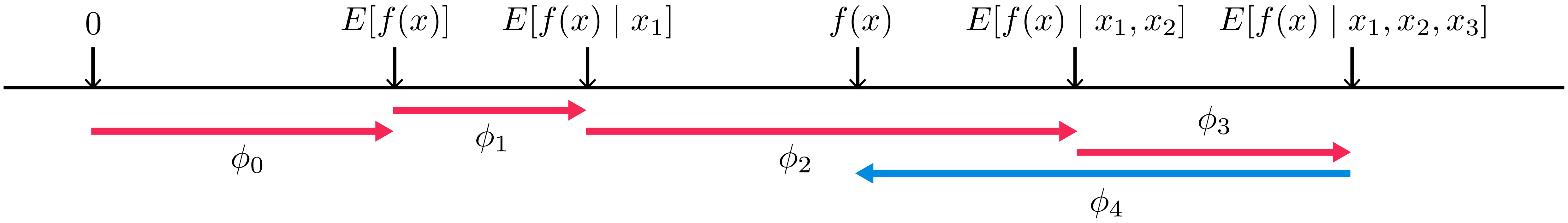
- SHAP: SHapley Additive exPlanations

* https://arxiv.org/abs/1602.04938

Outline

- Motivation
- Feature importance metrics
- What are SHAP values and why are they useful?
- How are SHAP values calculated?
- Outlook

SHAP values



- $E[f(x)|x_s]$ - the prediction of the model if only features x_s are used

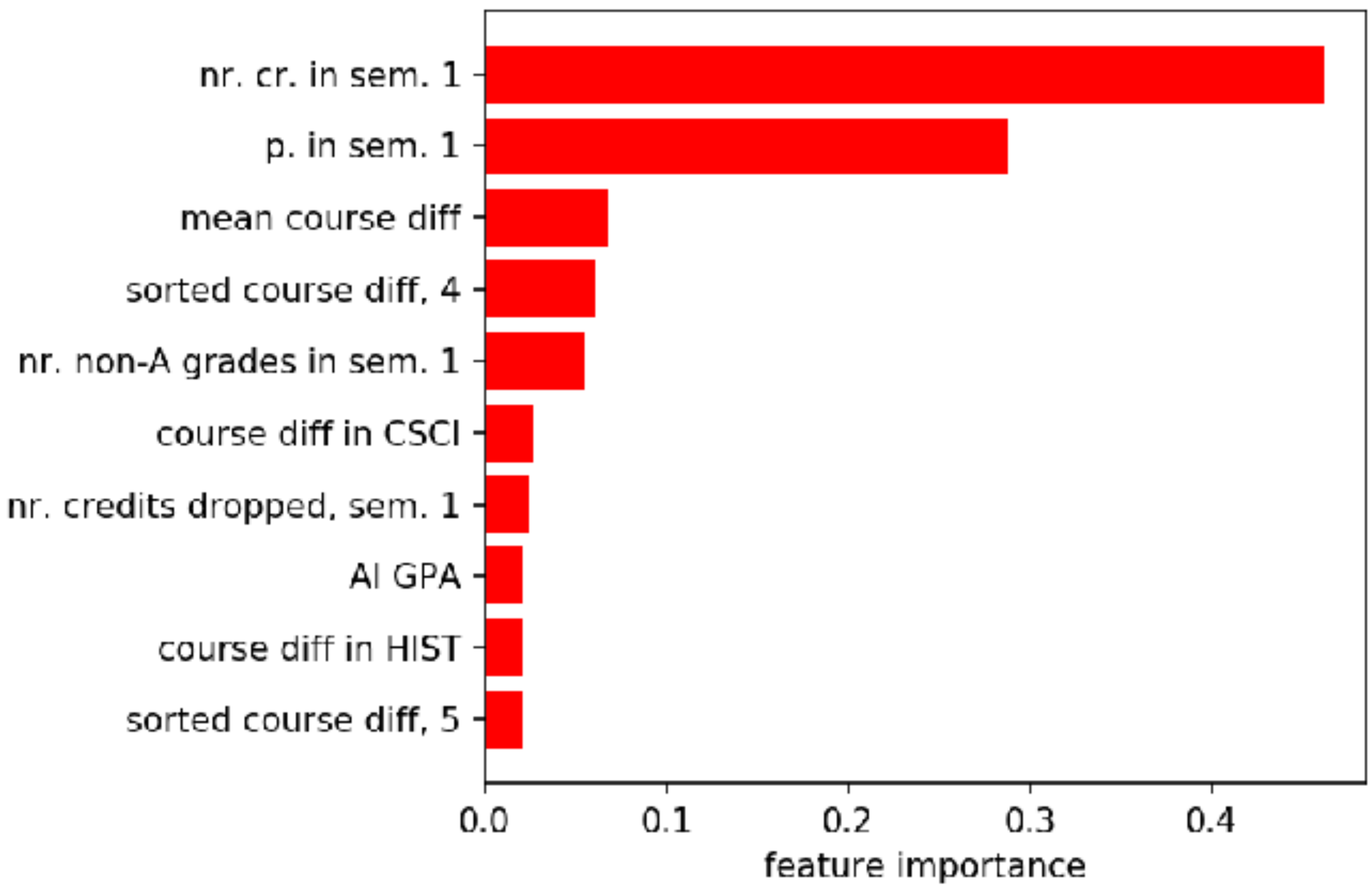
- $E[f(x)]$ - bias term, only the target variable is used to predict

- $E[f(x)|x_1,x_2,x_3,x_4] = f(x)$

- $\sum \phi_i = f(x)$

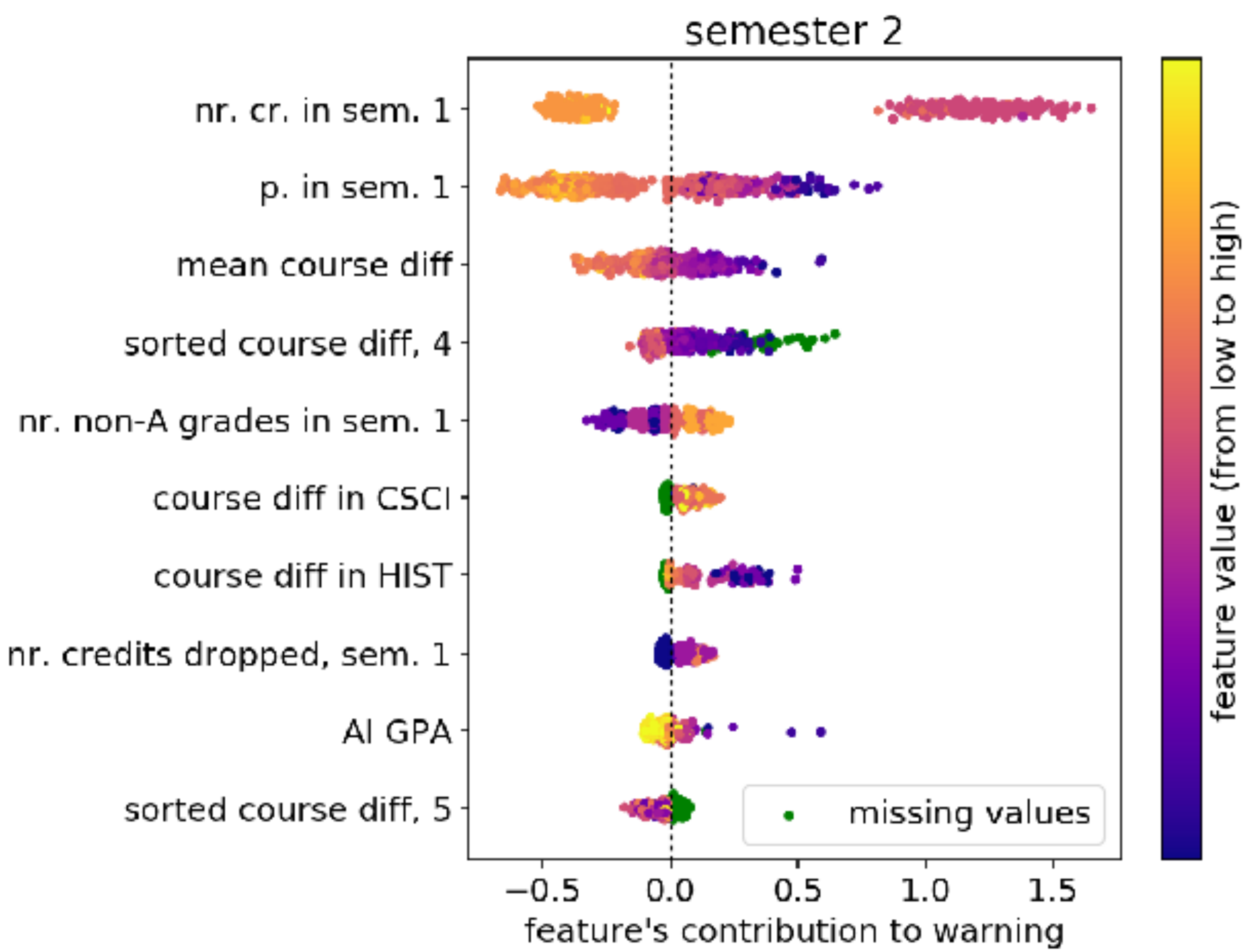
SHAP values - why are they useful?

From global feature importance plots...



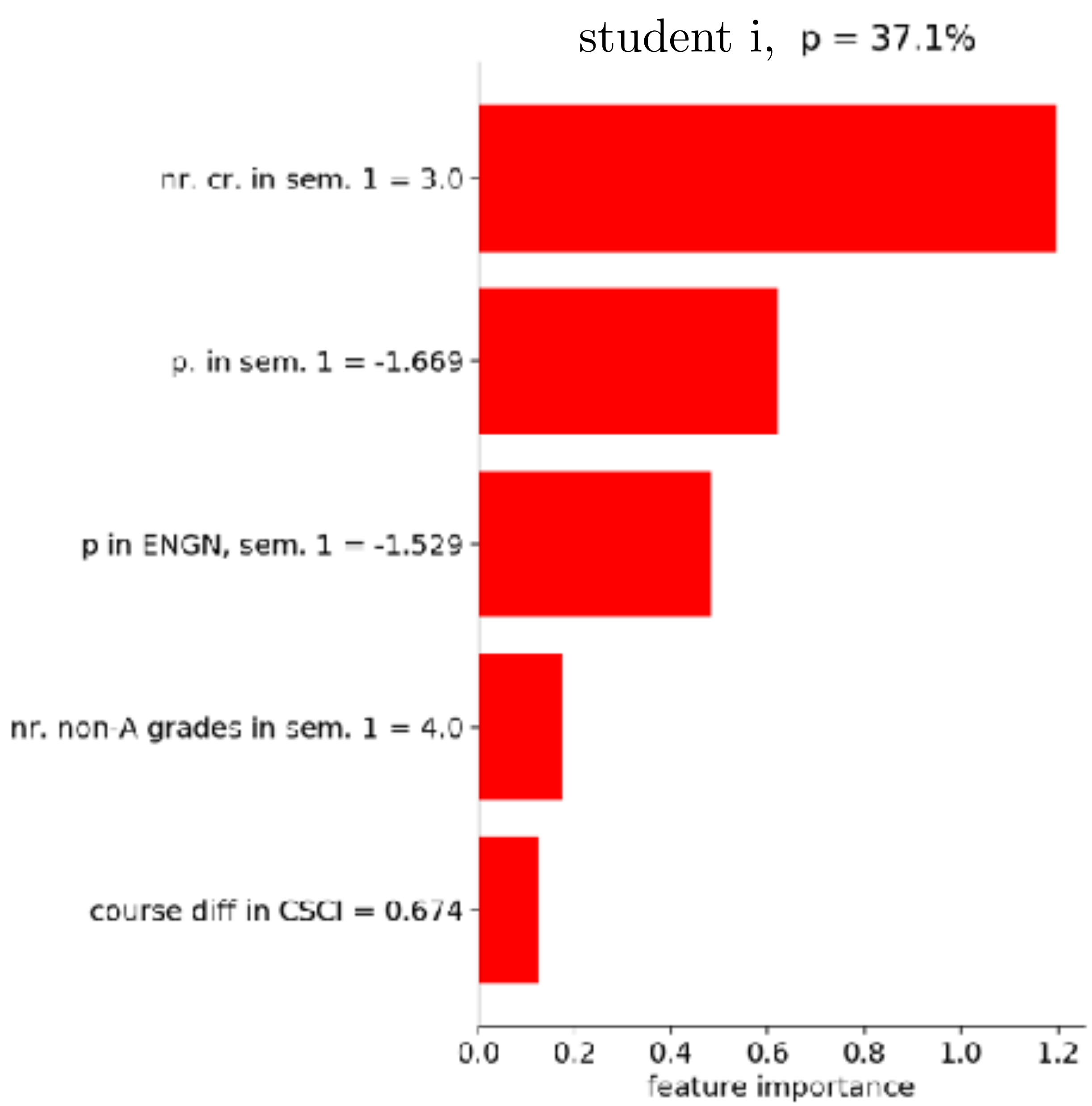
SHAP values - why are they useful?

...to SHAP summary plots...



SHAP values - why are they useful?

... and local feature importance plots.



Outline

- Motivation
- Feature importance metrics
- What are SHAP values and why are they useful?
- **How are SHAP values calculated?**
- Outlook

How is it calculated?

- Cooperative game theory
- A set of m players in a coalition generate a surplus.
- Some players contribute more to the coalition than others (different bargaining powers).
- How important is each player to the coalition?
- How should the surplus be divided amongst the players?

How is it calculated?

- Cooperative game theory applied to feature attribution
- A set of m features in a model generate a prediction.
- Some features contribute more to the model than others (different predictive powers).
- How important is each feature to the model?
- How should the prediction be divided amongst the features?

How is it calculated?

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S) \right],$$

- i - the feature whose contribution we want to calculate
- M - the number of features
- S - a set of features excluding i
- |S| - the number of features in S
- $f_x(S)$ - the expected value of the prediction with features S

How is it calculated?

$$\phi_i = \boxed{\sum_{S \subseteq M \setminus \{i\}}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S) \right],$$

- Loop through all possible ways a set of S features can be selected from the M features excluding i
- $[f_x(S \cup \{i\}) - f_x(S)]$ is the contribution of feature i to the model with features S
- Weight this appropriately

How is it calculated?

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \boxed{[f_x(S \cup \{i\}) - f_x(S)]}.$$

- Loop through all possible ways a set of S features can be selected from the M features excluding i

- $[f_x(S \cup \{i\}) - f_x(S)]$ is the contribution of feature i to the model with features S

- Weight this appropriately

How is it calculated?

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \boxed{\frac{|S|!(M - |S| - 1)!}{M!}} [f_x(S \cup \{i\}) - f_x(S)],$$

- Loop through all possible ways a set of S features can be selected from the M features excluding i

- $[f_x(S \cup \{i\}) - f_x(S)]$ is the contribution of feature i to the model with features S

- Weight this appropriately

How is it calculated?

$$\Phi_i = \frac{1}{\text{nr features}} \sum_{\text{features excluding i}} \frac{\text{contribution of i to model}}{\text{nr models with same number of features}}$$

How is it calculated?

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S) \right],$$

- How to calculate $f_x(S)$?
- How to ignore the effect of features not in S?

¯_(ツ)_/¯

Outline

- Motivation
- Feature importance metrics
- What are SHAP values and why are they useful?
- How are SHAP values calculated?
- Outlook

Outlook

- SHAP interaction index
    - pairwise interactions
    - how important are features i and j together?
- Supervised Clustering
    - run clustering on the SHAP values
    - it naturally converts all input features to the same unit

Summary

- Use SHAP values in supervised ML (especially with tree-based models)
- It's a great tool!