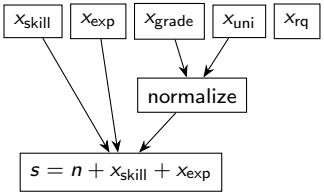
RankingSHAP - Listwise Feature Attribution Explanations for Ranking Models

Maria Heuss[1], Maarten de Rijke[1], Avishek Anand[2]

[1]University of Amsterdam, [2]TU Delft

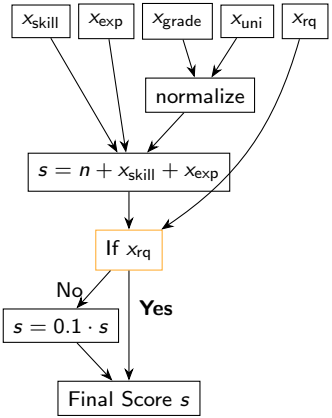
SIGIR 2025

An Example Task: Talent Search - A White Box Model



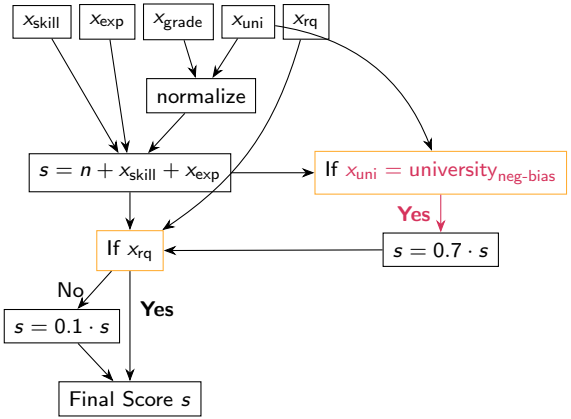
Flowchart of an unbiased white-box model

An Example Task: Talent Search - A White Box Model



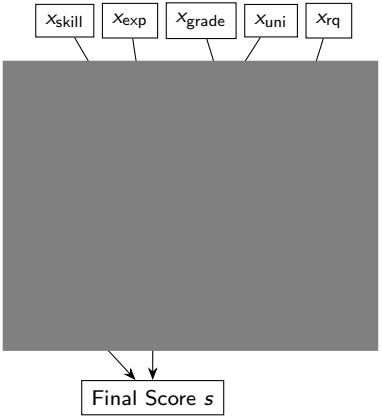
Flowchart of an unbiased white-box model

An Example Task: Talent Search - A White Box Model



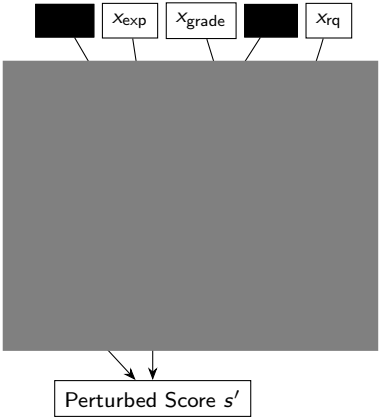
Flowchart of a biased white-box model

An Example Task: Talent Search - Explaining a Black Box Model



An opaque (black-box) model

An Example Task: Talent Search - Explaining a Black Box Model



An opaque (black-box) model

Feature Attribution Explanation

Feature Attribution

Dictionary, assigning each input feature a value representing the importance.

$$\{x_{\text{rq}} \mapsto 0.4, x_{\text{uni}} \mapsto 0.3, x_{\text{skill}} \mapsto 0.1, \\ x_{\text{exp}} \mapsto 0.1, x_{\text{grade}} \mapsto 0.1\}$$

Pointwise vs Listwise Explanations

Classical XAI generates and evaluates pointwise explanations:

- Pointwise explanation: Why does this document get a high ranking score?

Contrastivity in Explanations

Miller/Molnar: Explanations need to be contrastive.

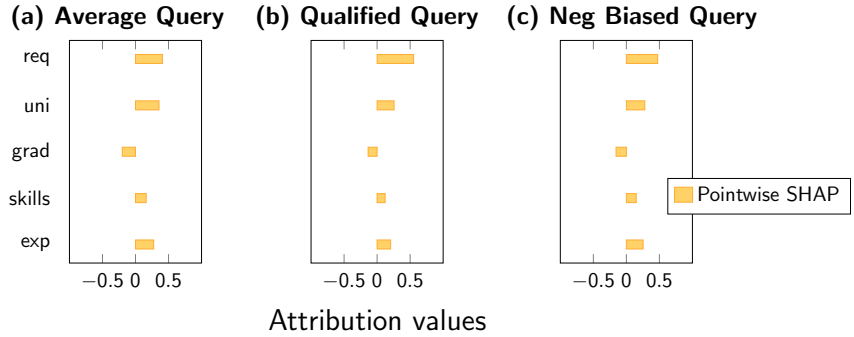


Figure: Feature attribution values for different query scenarios for a Pointwise SHAP Explainer.

Tim Miller, "'Explanation in Artificial Intelligence: Insights from the Social Sciences'", 2019
Christophe Molnar, "*Interpreting Machine Learning Models with SHAP*", 2023

Pointwise vs Listwise Explanations

Classical XAI generates and evaluates pointwise explanations:

- Pointwise explanation: Why does this document get a high ranking score?

In IR we want to generate and evaluate pair- and listwise explanations:

- Pairwise explanations: Why is document A ranked higher than document B?
- Listwise explanations: Why are the documents ranked in this order?

Background - A conceptual idea of (Pointwise) SHAP

We want to explain the model decision of a pointwise (regression) model

$$\tilde{R} : \mathbb{R}^n \to \mathbb{R}, x \mapsto \tilde{R}(x)$$

for input $x = (x_1, \ldots, x_n)$.

Background - A conceptual idea of (Pointwise) SHAP

We want to explain the model decision of a pointwise (regression) model

$$\tilde{R} : \mathbb{R}^n \to \mathbb{R}, x \mapsto \tilde{R}(x)$$

for input $x = (x_1, \ldots, x_n)$.

- Input-output relationship: Through perturbation on the input x, determine which features are the most important for a high ranking score (output).
- Perturbation is done through masking out subsets of features ("leave away features")
- Impact of leaving away features is measured through the change in ranking score.

Background - A conceptual idea of (Pointwise) SHAP

We want to explain the model decision of a pointwise (regression) model

$$\tilde{R} : \mathbb{R}^n \to \mathbb{R}, x \mapsto \tilde{R}(x)$$

for input $x = (x_1, \ldots, x_n)$.

- Input-output relationship: Through perturbation on the input x, determine which features are the most important for a high ranking score (output).
- Perturbation is done through masking out subsets of features ("leave away features")
- Impact of leaving away features is measured through the change in ranking score.

SHAP values

$$\phi_i(x) = \sum_{S \subset \{1,\ldots n\} \setminus i} w_S \cdot \mathbb{E}_{b \sim B}[\tilde{R}(mask_{S \cup \{i\},b}(x)) - \tilde{R}(mask_{S,b}(x))],$$

Method - From SHAP to RankingSHAP

We want to explain the model decision of a listwise ranking model

$$R : \{\mathcal{D}_q\}_q \to S_d, \{x^j\}_j \mapsto \pi_q$$

for input $\mathcal{D}_q = \{x^j\}_j$.

Method - From SHAP to RankingSHAP

We want to explain the model decision of a listwise ranking model

$$R : \{\mathcal{D}_q\}_q \to S_d, \{x^j\}_j \mapsto \pi_q$$

for input $\mathcal{D}_q = \{x^j\}_j$.

- We want to determine the features $\{1, \ldots, n\}$ most important to the order of the documents.
- How do we mask the input consisting of several document feature vectors?
- How do we measure the impact of a perturbation on the model output (ranking)?

Method - Measuring the change in output: Listwise Explanation Objectives

- Reduce the model prediction to a single value, measuring the change in model output.
- Investigate a particular aspect of the model decision:
  - Overall order of documents
  - Position of one specific document in the list
  - Order of documents among different groups.

Method - Measuring the change in output: Listwise Explanation Objectives

- Reduce the model prediction to a single value, measuring the change in model output.
- Investigate a particular aspect of the model decision:
  - Overall order of documents
  - Position of one specific document in the list
  - Order of documents among different groups.

Example: Explain the listwise order of the documents with the rank similarity with the original ranking π_q:

$$g_q(\tilde{\pi}) = \tau(\pi_q, \tilde{\pi}). \tag{1}$$

We use SHAP on $g_q \circ R$.

Method - Masking

We sample the masks $mask_{t,b}$ in the same way as we would for pointwise SHAP but apply it to each document feature vector:

$$mask_{t,b}(\mathcal{D}_q) = \prod_{|\mathcal{D}_q|} mask_{t,b}(x_{q,i})$$

Method - RankingSHAP

We want to explain the model decision of a listwise ranking model

$$R : \{\mathcal{D}_q\}_q \to S_d, \{x^j\}_j \mapsto \pi_q$$

for input $\mathcal{D}_q = \{x^j\}_j$. (For example: $R = \text{ranked} \circ \prod_{|\mathcal{D}_q|} \tilde{R}$)

Method - RankingSHAP

We want to explain the model decision of a listwise ranking model

$$R : \{\mathcal{D}_q\}_q \to S_d, \{x^j\}_j \mapsto \pi_q$$

for input $\mathcal{D}_q = \{x^j\}_j$. (For example: $R = \text{ranked} \circ \prod_{|\mathcal{D}_q|} \tilde{R}$)

- We want to determine the features $\{1, \ldots, n\}$ most important to the order of the documents.
- We mask each document feature vector with the same mask $mask_{t,b}(\mathcal{D}_q) = \prod_{|\mathcal{D}_q|} mask_{t,b}(x_{q,i})$
- We explain a certain aspect of the ranking decision, given by listwise explanation objective g_q

Listwise Explanations: Contrastivity by Design

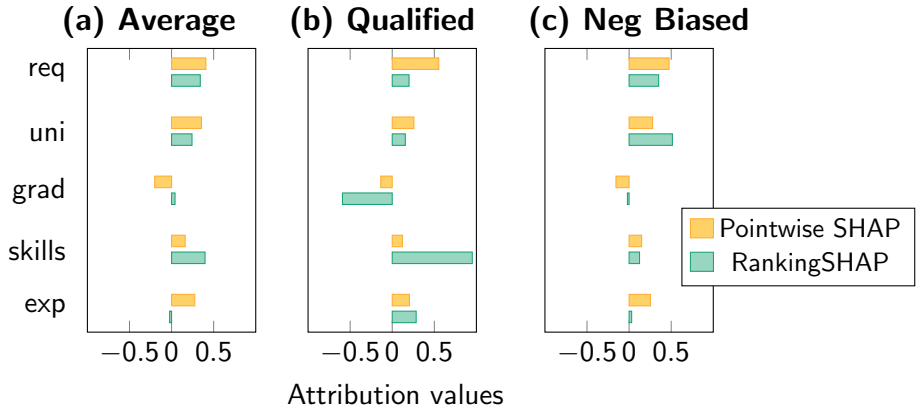


Figure: Feature attribution values for different query scenarios for a Pointwise SHAP Explainer.

Evaluation of RankingSHAP: Deletion and Preservation Check

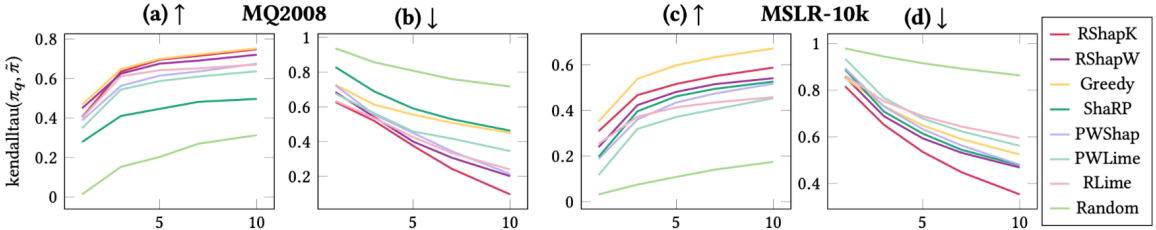


Figure: Preservation (a, c) and Deletion Check (b, d). Only top-k (k in the x-axis) features are kept/masked. Higher values indicate better explanations for the Preservation and worse for the Deletion Check.

Conclusion

- We need listwise explanations to explain listwise decisions.
- In the paper: We rigorously define listwise feature attribution.
- RankingSHAP is a flexible feature attribution approach that can explain different aspects of the ranking model decision.

- Limitation: SHAP is computationally expensive/slow and can be hard to interpret
- Limitation & Future Work: Quantitative evaluation of explanations is hard
- Future work: Use of RankingSHAP in real life use cases

References I

[1] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.

[2] Christophe Molnar. *Interpreting Machine Learning Models with SHAP*. Independently published, 2023.

Shapley Values

Passengers	Cost	Note
$\emptyset$	\$0	No taxi ride, no costs
{Alice}	\$15	Standard fare to Alice's & Bob's place
{Bob}	\$25	Bob always insists on luxury taxis
{Charlie}	\$38	Charlie lives slightly further away
{Alice, Bob}	\$25	Bob always gets his way
{Alice, Charlie}	\$41	Drop off Alice first, then Charlie
{Bob, Charlie}	\$51	Drop off luxurious Bob first, then Charlie
{Alice, Bob, Charlie}	\$51	The full fare with all three of them

Christophe Molnar, "*Interpreting Machine Learning Models with SHAP*", 2023

Shapley Values - Marginal Contributions

Addition	To Coalition	Cost Before	Cost After	Marginal Contribution
Alice	$\emptyset$	\$0	\$15	\$15
Alice	{Bob}	\$25	\$25	\$0
Alice	{Charlie}	\$38	\$41	\$3
Alice	{Bob, Charlie}	\$51	\$51	\$0

Shapley value

$$\phi_{\text{Alice}} = \sum_{S \subseteq \{\text{Bob},\text{Charlie}\}} w_S \cdot (v(S \cup \{\text{Alice}\}) - v(S))$$