Explain Yourself

Expanding and Optimizing Models to Enable Fast Shapley Value Approximations

Holger Ziekow, Peter Schanbacher and Valentin Göttisheim
Furtwangen University, Germany

Presenter: Valentin Göttisheim - email: Valentin.Goettisheim@hs-furtwangen.de

Resume

Academic Background:

- PhD candidate Data Science, Université de Haute-Alsace since 2023.
- Academic staff member Data Science, Furtwangen University since 2021.

Research Focus:

- Explainable Artificial Intelligence (XAI)
- Large Language Models (LLM)
- In industrial and medical domain

Striving for Explainable AI Models

Shift to inherently explainable models for trustworthy, transparent AI.

- **Demand for Transparency:** *Transparency fosters trust, accountability, and meets regulatory demands.*
- **Complexity of Neural Networks:** *Non-linear, high-dimensional interactions complicate feature interpretation.*
- **Limitations of Post-Hoc Explanations:** *Approximate, sometimes inconsistent and difficult to interpret fully.*
- **Advantages of Inherent Explainability:** *Embedding fair feature attributions directly aligns model outputs with transparency goals.*

HOCHSCHULE FURTWANGEN UNIVERSITY HFU

Striving for Explainable AI Models

Shift to inherently explainable models for trustworthy, transparent AI.

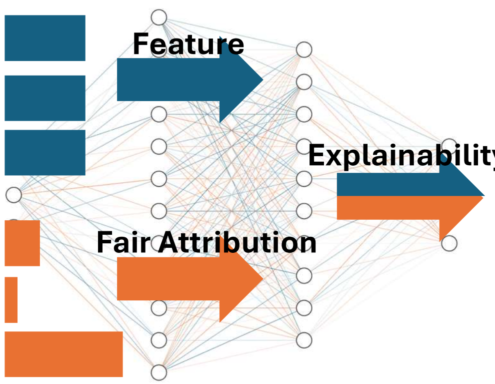
Explainability embedded within the loss function

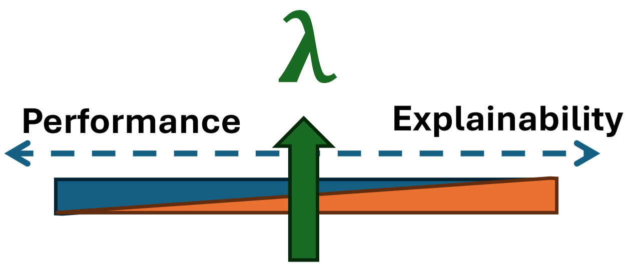
$$\text{Total Loss} = \text{Performance Loss} + \lambda \times \text{Explainability Loss}$$

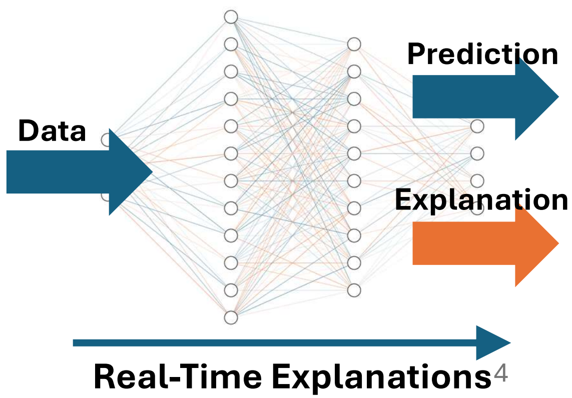
Enables the model to learn fair feature attributions during training.

Explicit trade-off between predictive performance an explainability.

Real-Time generation of predictions and Shapley values during inference.







Agenda

Explain Yourself - Expanding and Optimizing Models to Enable Fast Shapley Value Approximations

1. Shapley Value Landscape
2. Integrated Approach: Methodology
3. Results from Synthetic and Real-World Data
4. Conclusion and Future Directions

The Shapley Value Landscape

- <u>Fair Principles</u>

- Feature Attribution
- Simplifying Methods

Fair Attribution with Shapley Values [1,2,3]

Quantifies a feature's contribution to model predictions.

***Fair* Principles:** Treats features as players in a coalition.

- **Efficiency:** Total prediction distributed among features
- **Symmetry:** Equal contributions receive equal values
- **Dummy:** Irrelevant features have zero attribution
- **Additivity:** Supports combining contributions

Benefit: Fair feature attributions

Challenge: Computational complexity

The Shapley Value Landscape

- Fair Principles
  - Feature Attribution
- Simplifying Methods

Feature Attribution: [3]

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

Key Terms:

- $\phi_i(f)$ is the Shapley value for feature i.
- N is the set of all features.
- S is a subset of features not containing i.
- $f(S)$ is the model's output with features in S.

The Shapley Value Landscape

- Fair Principles
- Feature Attribution
  - Simplifying Methods

Feature-Removal Approaches: [4]

- **Baseline:** Replace missing features with baseline values (e.g., mean, zero).
- **Marginal:** Evaluate subsets to compute marginal effects.
- **Conditional**: Account for feature dependencies via conditional expectations.

Efficient Computation:

- **KernelSHAP** [3]**:** Model-agnostic; broadly applicable.
- **TreeSHAP** [5]**:** Model-specific; optimized for decision trees.
- **…**

Limitations: Post-hoc methods are not aligned with training.

Integrated Shapley Values

Embedding Explainability within the loss function

$$\text{Total Loss} = \underbrace{\text{Performance Loss}} + \lambda \times \underbrace{\text{Explainability Loss}}$$

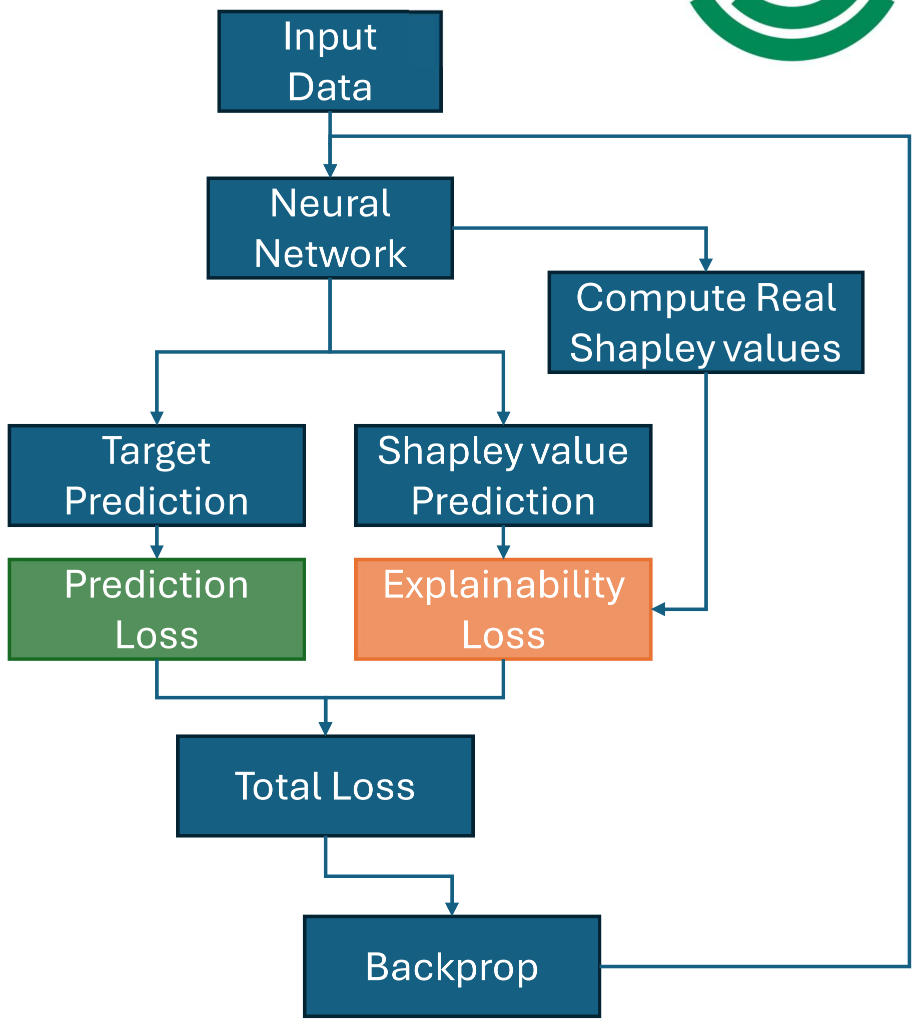
$$g = \arg\min_{g} \mathrm{E}\left[\underbrace{\left(y - g_0(x)\right)^2}_{\text{Prediction Loss}} + \lambda \sum_{i=1}^{N} \underbrace{\left(\phi_i(x) - \hat{\phi}_i(x)\right)^2}_{\text{Explainability Loss}}\right]$$

Prediction Loss $\left(y - g_0(x)\right)$:

- Minimizes the difference between actual y and predicted $g_0(x)$ values

Explainability Loss $\left(\phi_i(x) - \hat{\phi}_i(x)\right)$:

- Minimizes the discrepancy between true Shapley values $\phi_i(x)$ and their approximations $\hat{\phi}_i(x)$



Integrated Shapley Values

Neural Network Architecture & Methodology

$$g = arg\,\min_{g} \mathrm{E}\left[\left(y - g_0(x)\right)^2 + \boldsymbol{\lambda}\sum_{i=1}^{N}\left(\phi_i(x) - \boldsymbol{\hat{\phi}_i(x)}\right)^2\right]$$

How $\boldsymbol{\hat{\phi}_i(x)}$ is Derived

Utilize KernelExplainer to compute real Shapley values $\phi_i(x)$ during training.

Learning Approximation:

- The model integrates Shapley approximations as an output
- Learning to predict $\hat{\phi}_i(x)$ by minimizing the difference between $\phi_i(x)$ and $\hat{\phi}_i(x)$

Mechanism of $\boldsymbol{\lambda}$

λ balances the emphasis between prediction accuracy and Shapley value attributions.

Low λ Values:

- Prioritizes prediction accuracy.
- Minimizes prediction error with minimal emphasis on Shapley values.

High λ Values:

- Enhances Shapley value precision and explainability.
- Increases the weight of explainability loss, improving feature attribution accuracy.

Results from Synthetic and Real-World Data

Experimental Design to Evaluate Shapley Integration

Experimental Setup

Objective: *Demonstrate the impact of embedding Shapley values on accuracy and interpretability by varying the λ parameter.*

Synthetic Dataset:

- Created to observe trade-offs between prediction accuracy and explainability.
- Includes controlled linear and complex feature relationships.

Real-World Dataset (Wine Quality):

- Source: Wine-quality-red dataset from OpenML.
- Features like ‘sulphates’, ‘alcohol’, and ‘total sulfur dioxide’.

Architecture and Training

Input Layer: 3 nodes (one for each feature).

Hidden Layers:

- Layer 1: 16 neurons, Leaky ReLU activation.
- Layer 2: 8 neurons, Leaky ReLU activation.

Output Layer:

- 1 Target Prediction y
- 3 Shapley Value Approximations $(\phi^1(x), \phi^2(x), \phi^3(x))$

Training Configurations:

- **λ = 0:** Accuracy-focused, with minimal emphasis on Shapley values.
- **λ = 1:** Balanced, with equal weight on prediction accuracy and explainability.
- **λ = 1000:** Shapley-focused, prioritizing interpretability over prediction accuracy.

Experiment I.

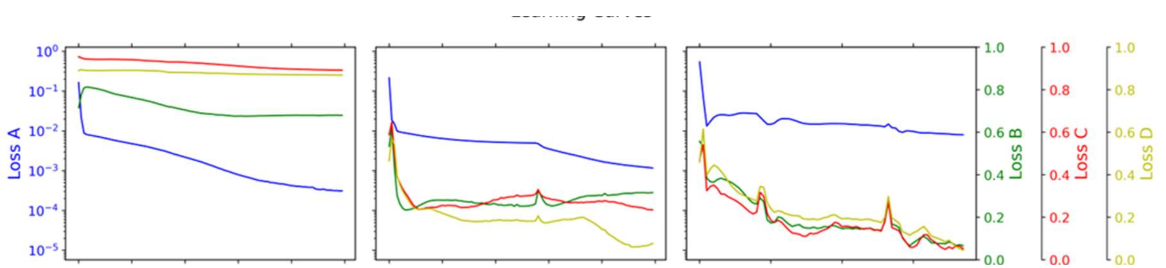


Figure: *Learning Curves with MSE for λ ∈{0,1,100} models (left to right) for the outcome of interest y (blue) and the corresponding Shapley values (green, red, orange).*

Synthetic Dataset

Design:

- Target $y = 2 \cdot x_1 + \frac{1}{2}\epsilon$ where x_1 is the main feature with noise ϵ.
- x_2: Independent, however, non-linear transformation of x_1 and y.
- x_3: Independent, uniformly distributed.

Findings:

- Higer λ:
  - Correct attributions, improving explainability.
  - Reduces MSE for Shapley values.
- Trade-off in prediction accuracy vs. interpretability

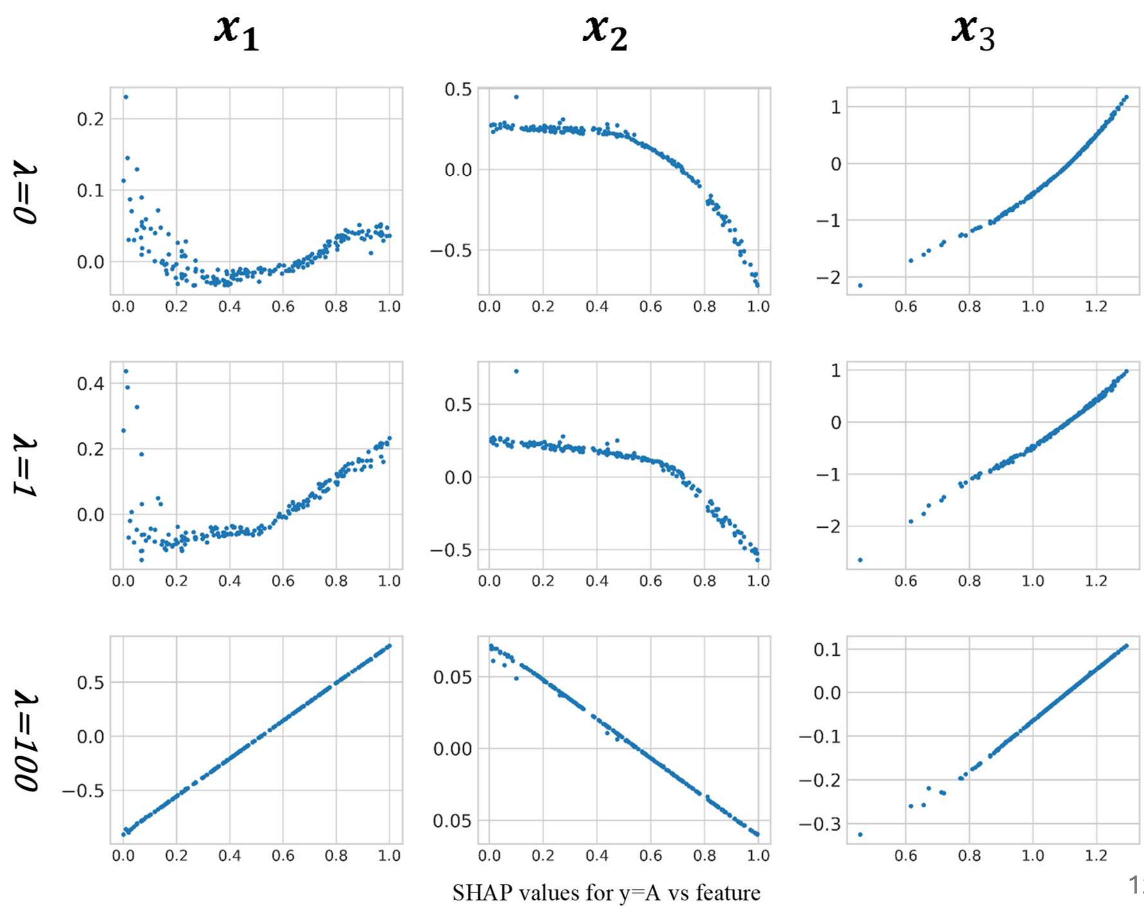


Figure*: Shapley values of features (left to right) of models λ ∈0,1,100 (top to bottom).*

Experiment II.

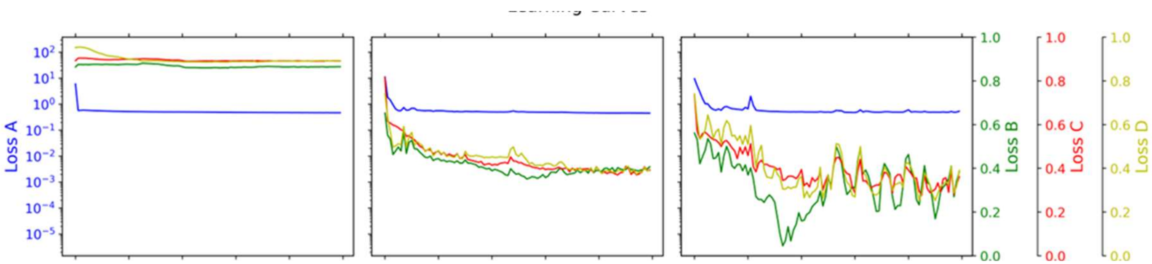


Figure: *Learning Curves with MSE for λ ∈{0,1,1000} models (left to right) for the outcome of interest y (blue) and the corresponding Shapley values (green, red, orange).*

Wine Quality Dataset

Data Source:

- Wine-quality-red dataset from OpenML [6].

Features Used:

- 'sulphates,' 'alcohol,' and 'total sulfur dioxide' chosen by explorative analysis

Findings:

- Partial dependency plots reveal more stable Shapley values at higher λ

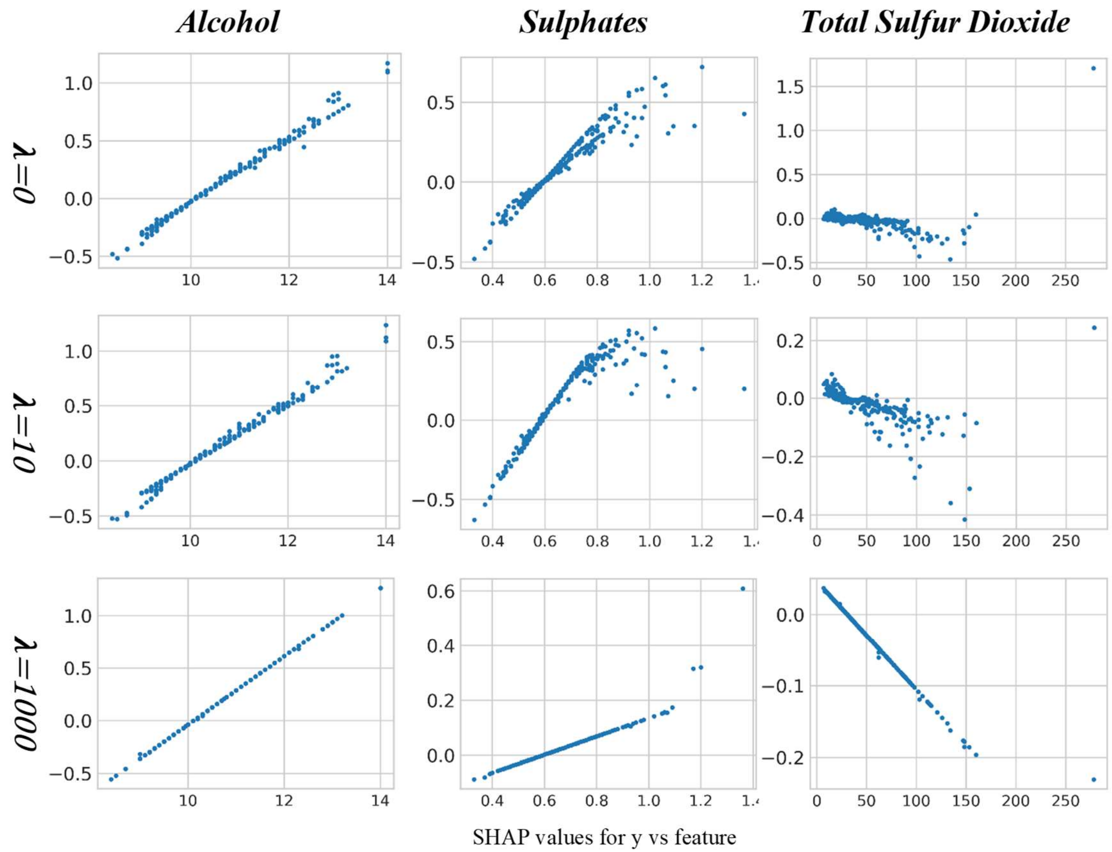


Figure: *Shapley values of features (left to right) of models λ ∈0,1,1000 (top to bottom)*

Conclusion

- Embedding Shapley values aligns feature attributions to fair principles during training.
- Adjusting λ enables a flexible trade-off between accuracy and interpretability.
- Both synthetic and real-world experiments show that increasing λ enhances explainability.

Future Research Directions

- Change approach to align feature attributions to improve scalability.
- Extending experiments to complex architectures and broader data sets could expand application potential.
- Evaluating performance-interpretability trade-off.

Questions?

Valentin Göttisheim – Valentin.Gottisheim@hs-furtwangen.de

References

1. L. Shapley, “Notes on the n-Person Game -- II: The Value of an n-Person Game.”, Santa Monica, Calif.: RAND Corporation, 1951.

2. L. Merrick and A. Taly, “The Explanation Game: Explaining Machine Learning Models Using Shapley Values” *Machine Learning and Knowledge Extraction*, Springer International Publishing, CD-MAKE 2020, pp. 17-38, Dublin, Ireland, August 25–28, 2020.

3. S. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions”, *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

4. H. Chen, I. Covert, and S. Lundberg, “Algorithms to estimate Shapley value feature attributions”, *Nature Machine Intelligence 5*, pp. 590–601, 2023

5. S. Lundberg, G. Erion, H. Chen, et al., “From local explanations to global understanding with explainable AI for trees”, *Nature Machine Intelligence 2*, pp. 56–67, 2020.

6. OpenML, “Red Wine Quality Dataset. Dataset”, [Online] https://openml.org/search?type=data&status=active&id=40691, [retrieved: 10, 2024].

Backup: House Price Prediction Example

Model Prediction: $440,000

1. **Square Footage (120m²):** $120,000
2. **Location (Valencia):** $200,000
3. **Bedrooms (4):** $120,000
4. **Proximity to Park (100m):** $0

Assumptions of Fair Attribution:

1. **Efficiency:** All features
2. **Symmetry:** Square Footage and Location
3. **Dummy:** Proximity to Park

...

Efficiency: Total contribution equals the model's prediction.

$$\phi_{Square\ Footage} + \phi_{Location} + \phi_{Bedrooms} + \phi_{Proximity\ to\ Park} = 120{,}000 + 200{,}000 + 120{,}000 + 0 = 440{,}000$$

Symmetry: Features with equal contributions receive equal attribution.

$$f(Location \cup Square\ Footage) - f(Location) = 120{,}000$$

$$f(Location \cup Bedrooms) - f(Location) = 120{,}000$$

$$\phi_{Square\ Footage} = \phi_{Bedrooms} = 120{,}000$$

Dummy: Features with no impact receive zero attribution.

$$f(S \cup Proximity\ to\ Park) - f(S) = 0 \quad \text{for all subsets } S$$

$$\phi_{Proximity\ to\ Park} = 0$$