Teaching Interpretable AI and Critical Data Literacy in Civic Education

Olushina Olawale Awe, PhD
PH Ludwigsburg University of Education, Germany

Paderborn Colloquium on Artificial Intelligence and Data Science Education at School Level.
19 November 2025

UNTERSTÜTZT VON / SUPPORTED BY
Alexander von HUMBOLDT STIFTUNG

Overview

From Data Literacy to Critical Data Literacy

Traditional Data Literacy

Focuses on:

- reading graphs and tables
- basic statistics (mean, median, correlations)
- simple inference (confidence intervals, p-values)

Critical Data Literacy (CDL) in Civic Education

Adds a critical lens:

- Who collected the data, and why?
- What is missing or underrepresented?
- Who may be harmed or advantaged by a model?
- interpret statistical and algorithmic claims

Literature Landscape: From AI Governance to Civic Data Literacy

AI as a Democratic and Civic Instrument

Various scholars increasingly call for aligning **AI with democratic values, public accountability, and civic participation. Züger & Asghari (2023)** frame AI within *public interest theory*, shifting focus from corporate control to democratized, transparent governance. **Margetts (2022)** emphasizes *good governance principles*, accountability, equity, and trust as central to public AI.

Participatory, Human-Centered, and Trustworthy AI

Moon (2023) calls for participatory AI design to reduce algorithmic harms. **Robles & Mallinson (2025)** show that trust in AI depends on civic engagement and institutional responsiveness. **Allen et al. (2025)** advocate *power-sharing liberalism*, embedding civic representation in AI oversight, while **McKenna (2025)** highlights human awareness, creativity, and collective intelligence for democratic resilience.

AI Literacy, Civic Education, and Explainability

McStay (2020) warns against manipulative AI in EdTech, urging ethical scrutiny and accountability. **Weber-Stein & Engel (2025)** propose embedding **Stat Lit. into civic education**. **Siamtanidou et al. (2025)** explore gamified, affective AI to foster participatory learning. Others include: Ridgway(2022), Ridgway et al (2023), Giustini & Dastyar (2024), Yim (2024), Daher (2025), Velander et al (2024), and Veldhuis et al (2024).

New Frontier

- Interpretable AI in Civic Education $\Rightarrow$ a concrete arena where students can *see* and *question* algorithmic reasoning.
- Every "black box" must be opened, ethically, pedagogically, and technically!
- *Interpretable AI is an essential skill for democratic reasoning and civic literacy.*

Motivation:
The New Roles of Algorithms in Democratic Life

- Algorithms now influence:
  - credit scoring, welfare eligibility
  - predictive policing, risk assessment
  - education tracking, resource allocation etc
- Many of these systems are:
  - opaque ("black boxes")
  - data-hungry and high-impact
  - weakly understood by the public
- **Civic education** must therefore address:
  - How models work (mechanics)
  - How they can fail (bias, error, drift)
  - How citizens can question and contest them

What Is Interpretable AI? (Awe et al, 2025)

Working Definition

Interpretable AI refers to models and tools that allow humans to:

- understand *how* ML predictions are made
- identify *which features* drive outputs
- examine model behavior across individuals and subgroups
- communicate explanations in *plain language*

Global vs Local:

- **Global**: overall behavior (feature importance, partial dependence)
- **Local**: explanations for individual predictions (SHAP, break-down)

Why Interpretable AI for Critical Literacy?

- Black-box AI can **hide**:
  - biased training data
  - skewed error rates across groups
  - arbitrary thresholds and design choices
- Interpretable AI tools (e.g., DALEX):
  - expose model structure and feature effects
  - reveal where marginalized groups are misclassified
  - create openings for contestation and reform
- In education:
  - students learn to move from **"the model says"** to **"why does the model say this, and is it acceptable?"**

See John-Mathews (2022), Giustini & Dastyar (2024), Yim (2024), Daher (2025), Velander et al (2024), Veldhuis et al (2024) and Biecek (2018).

Interpretable AI vs. Explainable AI

Interpretable AI	Explainable AI (XAI)
Transparent by design • Model structure is directly understandable. • Reasoning can be traced from inputs to outputs.	**Black-box by design** • Internal workings are complex/opaque. • Human cannot easily follow the raw computation.
Explanations are built-in • Coefficients, rules, splits have clear meanings. • No extra tools needed to explain predictions.	**Explanations are post-hoc** • Uses tools (DALEX, SHAP, LIME, PD/ALE). • Explanations approximate the model's behaviour.
Typical models • Linear / logistic regression • Small decision trees • Simple rule lists, some GAMs	**Typical models** • Random forests, GBM, XGBoost • Deep neural networks • Kernel SVM and complex ensembles
Civic education role • Great for teaching basic statistical reasoning.	**Civic education role** • Great for showing how to "open" black-box models.

Methodology: General Regression Framework

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- y_i: observed outcome (e.g., income, access to service)
- $x_i = (x_{i1}, ..., x_{ip})$: predictors
- $f(x_i)$: functional relationship learned from data
- ε_i: random noise or unexplained variation

Each algorithm estimates $f(\cdot)$ differently while balancing accuracy, interpretability, and generalization.

1. Linear Regression

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

- β_0: intercept or the baseline value.
- β_j: coefficient measuring the change in y per unit change in x_j.

Objective:

$$L = \min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Interpretation: Clear and intuitive, excellent for policy transparency but limited in handling nonlinearity.

2. Decision Tree Regression

$$f(x) = \sum_{m=1}^{M} c_m \mathbf{1}(x \in R_m)$$

- R_m: region (leaf) of the input space created by splits.
- c_m: average outcome in that region.

Interpretation: Mimics human decision processes using "if–then" logic. Highly visual and interpretable but prone to instability.

3. Random Forest

$$f_{\text{RF}}(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$$

- $f_b(x)$: prediction from the b^{th} decision tree.
- B: number of trees.

Averages multiple decision trees built on random subsets of data and variables.
Benefit: Robust and accurate.
Limitation: Loses transparency, hard to explain "why" a decision was made.

4. Support Vector Regression (SVR)

$$f(x) = \sum_i (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

Terms:

- $K(x_i, x)$: kernel function (linear, polynomial, or radial).
- α_i, α_i^*: coefficients for support vectors.
- b: bias term.

Objective:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C \sum_i \xi_i$$

Captures nonlinear relationships in high-dimensional space; complexity often obscures interpretability.

5. Neural Network Regression

$$\hat{y} = \sigma_2(W_2\,\sigma_1(W_1 X + b_1) + b_2)$$

- W_1, W_2: weight matrices connecting layers.
- b_1, b_2: biases.
- σ_1, σ_2: activation functions introducing nonlinearity.

Loss function:

$$L = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$

Powerful but difficult to interpret, motivating the need for structured explainability.

6. Gradient Boosting Machine (GBM)

$$f_M(x) = f_{M-1}(x) + \nu h_M(x)$$

Terms:

- $h_M(x)$: weak learner correcting previous errors.
- ν: learning rate controlling update size.

Interpretation: Models are added sequentially to minimize error. Excellent accuracy but requires post hoc methods to interpret cumulative effects.

A More Intuitive Picture of GBM for Teaching Students

Key Idea

GBM builds a strong model by adding many small *weak learners* (usually shallow trees). Each new learner tries to **correct the errors** of the current model.

Analogy:

- First draft of an essay = rough approximation.
- At each revision, you fix some mistakes.
- After many revisions, the essay is much better.

GBM = iterative error-correction of predictions.

GBM Model Equation

We model a function $F(x)$ as a sum of weak learners:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \nu \, h_m(x),$$

where:

- $F_0(x)$ = initial model (e.g., mean of y)
- $h_m(x)$ = weak learner at step m (small tree)
- ν = learning rate, $0 < \nu \leq 1$
- M = number of boosting iterations

Interpretation

- $F_0(x)$ gives the **first guess**.
- Each $h_m(x)$ gives a **small correction**.
- Final prediction = first guess + all corrections.

GBM Strengths and Risks

Strengths:

- Handles nonlinear relationships and interactions.
- Works well with tabular, structured data.
- Often top-performing in practice (with tuning).

Risks / Caveats:

- Can overfit if:
  - trees are too deep
  - there are too many iterations
  - learning rate is too large
- Harder to interpret than linear models.
- If trained on biased data, can **amplify** bias.

Civic Implication

Teaching GBM without fairness and interpretability invites uncritical adoption of powerful but opaque tools.

From AI Models to Interpretability: What is DALEX?

DALEX: Descriptive Machine Learning Explanations (Biecek, 2018)
DALEX = a framework to explain any machine learning model.

- Works with regression, classification, survival, ensembles.
- Treats model as a function: $f(x)$.
- Produces:
  - Global explanations: variable importance, PDP/ALE
  - Local explanations: SHAP, break-down, what-if profiles
  - Diagnostics: residuals, performance curves
- Connects naturally with ModelStudio dashboards.

$$DALEX = Transparency + Auditability + Civic\ Accountability$$

DALEX Workflow (4 Steps)

1. **Train any ML model** (lm, rf, gbm, svm, nnet, xgboost, etc.)
2. **Wrap it with an explainer:**

```
explain(model, data=X, y=y, predict_function=...)
```

3. **Generate global explanations:**
   - Variable Importance (VI)
   - Partial Dependence (PDP), Acummulated Local Effects
4. **Generate local explanations:**
   - SHAP, Break-Down
   - What-if profiles

The Explainer Object

In R:

```
expl <- explain(
  model,
  data = X,
  y    = y,
  predict_function = predict,
  label = "model name"
)
```

Meaning:

- Treats the model as a black box.
- Stores data + predictions + metadata.
- Foundation for all DALEX explanation functions.

DALEX Explainer

Explainer Object

DALEX builds an *explainer*:

$$\mathcal{E} = (f, X, y),$$

where:

- f = trained model (e.g., GBM),
- X = feature matrix,
- y = observed outcome.

On top of $\mathcal{E}$, DALEX provides:

- global variable importance,
- partial dependence / ALE profiles,
- local explanations (break-down, SHAP-like),
- model performance summaries.

Global Explanation: Permutation Variable Importance

For a variable X_j, DALEX defines its importance via permutation:

$$VI(X_j) = \mathbb{E}\big[L(f, X^{(j)_{\text{perm}}}, y) - L(f, X, y)\big],$$

where:

- L is a loss (e.g. RMSE, MAE),
- $X^{(j)_{\text{perm}}}$ is X with column j randomly permuted.

Intuition:

- If shuffling X_j greatly worsens performance, then X_j is important.
- If performance barely changes, X_j is less important.

Global Explanation: Partial Dependence (PD) Profiles

For a feature X_j, the partial dependence function is:

$$PD_j(z) = \frac{1}{n} \sum_{i=1}^{n} f\left(z, x_{i,-j}\right),$$

where:

- z is a value of interest for feature j,
- $x_{i,-j}$ are all other features for observation i.

Interpretation:

- $PD_j(z)$ = average prediction when X_j is set to z for everyone.
- Shows how predicted outcome changes as X_j varies, on average.

Local Explanations: Break-down and SHAP-like Decomposition

For a given observation x^*, we can decompose:

$$f(x^*) = f_0 + \sum_{j=1}^{p} \phi_j(x^*),$$

where:

- f_0 is a baseline prediction (e.g. mean),
- $\phi_j(x^*)$ is the contribution of feature j.

Break-down / SHAP-type plots show:

- how each feature pushes prediction *up* or *down*
- which features have largest positive/negative influence for that case

Pedagogical Lens

Students can see *why* a given individual or neighbourhood receives a high-risk or low-risk score in an exam.

Global Explanations: ALE (Accumulated Local Effects)

- Alternative to PDP that avoids extrapolation.
- Local derivative-based summary of feature effects.
- Safer when features are correlated.

ModelStudio Dashboard Integration

- DALEX powers the fully interactive ModelStudio dashboard.
- Compare models, inspect fairness, explore PDP/SHAP interactively.
- Great classroom tool for active learning.

Using DALEX to Teach Critical Thinking

DALEX can be used to help students:

- Question how models behave.
- Identify sources of bias or unintended harm.
- Compare linear vs nonlinear behaviors.
- Inspect fairness-relevant variables (e.g., race proxies, SES).
- Understand when models fail or generalize poorly.

Teaching Scaffold Using DALEX

Step 1 : Interpret

- Students examine VI, PDP, and SHAP values.
- Identify strongest model drivers.

Step 2 : Critique

- Discuss fairness, societal impact, sensitive variables.
- Are model assumptions realistic?

Step 3 : Redesign

- Rebuild model without problematic features.
- Compare interpretability and fairness metrics.

- DALEX is a crucial tool for teaching interpretable and explainable AI.
- Enables transparent, ethical, and critical engagement with models.
- Supports democratic participation and civic reasoning.

Interpretability = Empowerment

Training students to question, analyze, and hold AI systems accountable.

Boston Housing Data: Brief Description

Source: 1970 U.S. Census (506 tracts in Boston area).
Selected variables:

- `crim`: per-capita crime rate by town (target)
- `rm`: average number of rooms per dwelling
- `lstat`: % lower-status population
- `b`: $1000(B - 0.63)^2$ (race proxy; B = proportion "Black")
- `nox`: nitric oxides concentration
- `rad`: index of radial highway accessibility
- `chas`: dummy variable (=1 if tract bounds Charles River)

Why This Dataset?

- Contains socio-economic and racial proxies $\Rightarrow$ fairness-sensitive.
- Historically used to illustrate "crime" and "neighbourhood quality".
- Excellent context for **debunking harmful narratives** with interpretable AI.

Target and Transformation

Target: per-capita crime rate `crim`.

- Highly right-skewed: a few tracts have very high crime rates.
- This can destabilize models and exaggerate error.

Solution: Apply a log-transform:

$$y = \texttt{crim}, \quad y^{=\log(y+c),}$$

where c is a small constant if needed (e.g. $c = 0.01$).

- Reduces skewness.
- Makes residuals more homoscedastic.
- Improves interpretability of additive effects.

Modeling Setup (Conceptual)

Models

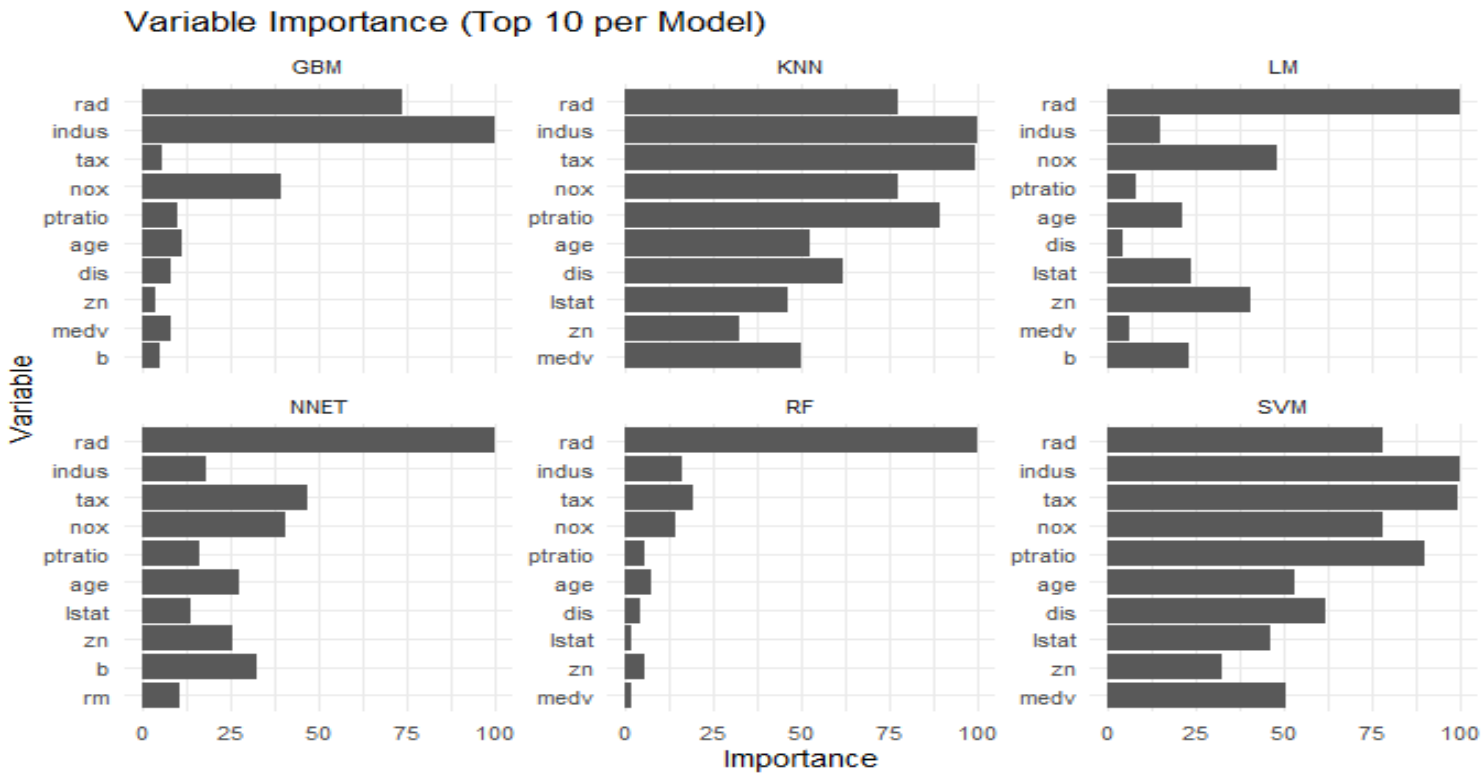
- Baseline: Linear Regression (LM)
- Tree-based: Random Forest (RF), Gradient Boosting Machine (GBM)

Workflow

1. Split into train/test.
2. Fit LM, RF, GBM to predict log(`crim`).
3. Evaluate performance on test set (RMSE, R^2, MAE).
4. Build DALEX explainers for each model.
5. Compare:
   - global variable importance
   - partial dependence (e.g., for `lstat`, `rm`)
   - local explanations for selected tracts

Variable Importance : Understanding Drivers of Crime Rate

Top Predictors:



[Variable Importance : Top 10 per Model]

Explaining the Model : DALEX Global Insights

Permutation Variable Importance:

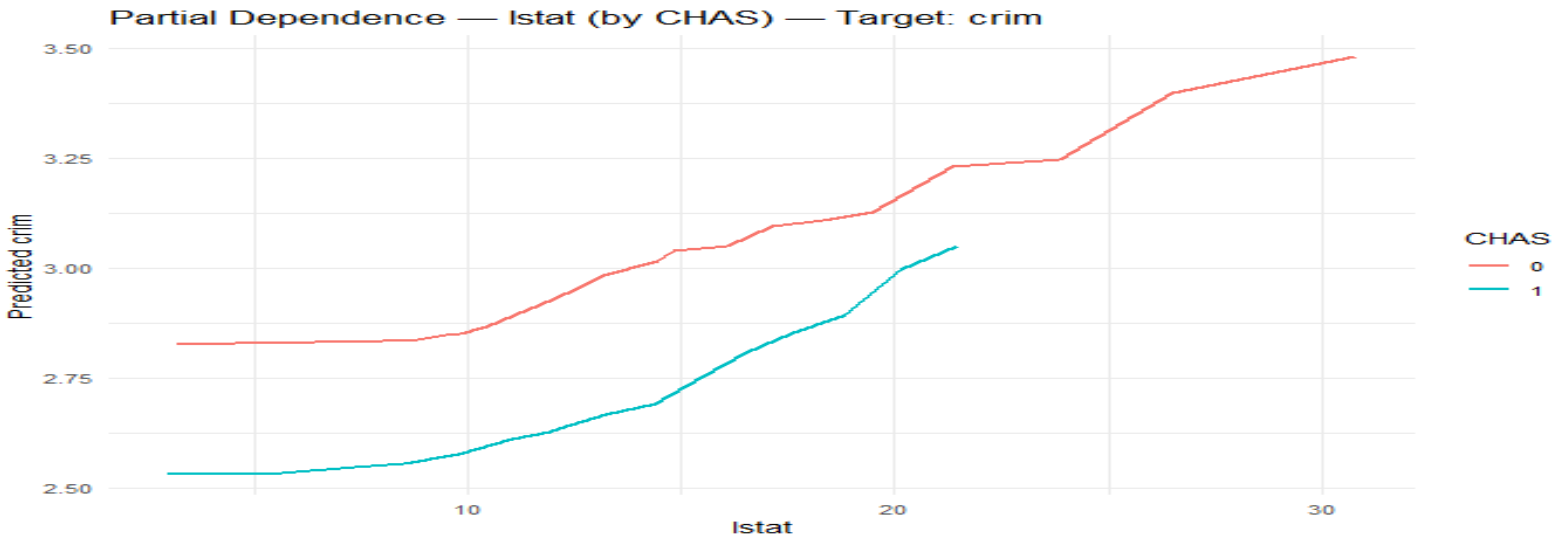
$$VI_j = \mathbb{E}[L(y, f(x^{(j)}_{\text{perm}})) - L(y, f(x))],$$

Partial Dependence:

$$PD_j(z) = \frac{1}{N} \sum_{i=1}^{N} f(z, x_{i,-j})$$

Interpretation:

- `rad` and `dis` remain top global influencers.
- Partial Dependence shows nonlinear rise in `crim` as `rad` increases.



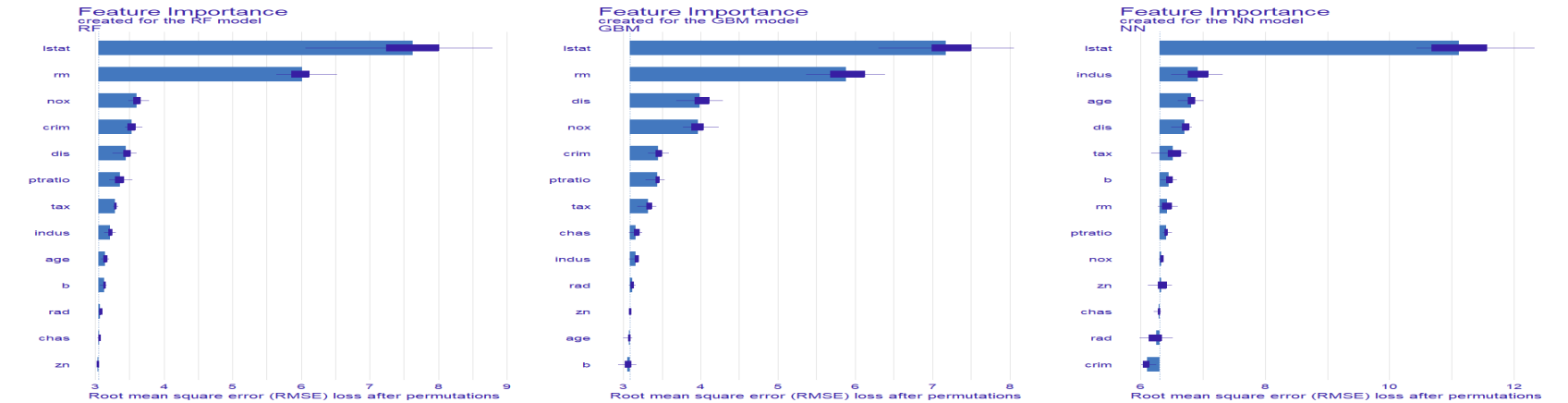
Local Explanations : SHAP and Break-Down

Mathematical Decomposition:

$$f(x_i) = f_0 + \sum_{j=1}^{p} \phi_{ij}, \quad \text{where } \phi_{ij} \text{ is the SHAP value for feature } j.$$

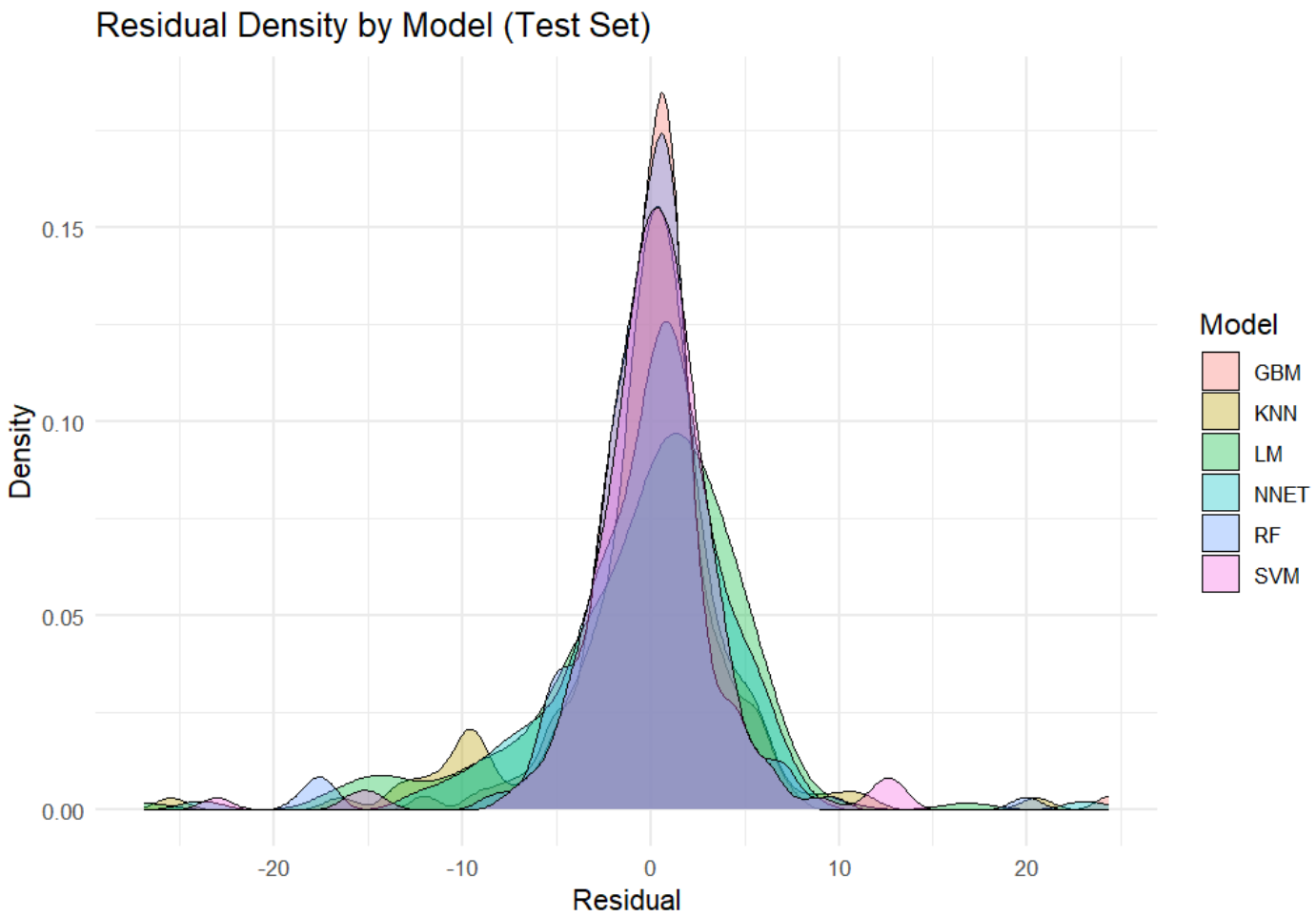
Insights:

- For a median-predicted case: `lstat`, `nox`, and `rad` increase crime estimate.
- SHAP enables transparent feature attribution : no black-box mystery.



[SHAP & Break-Down for Median-Predicted Observation]

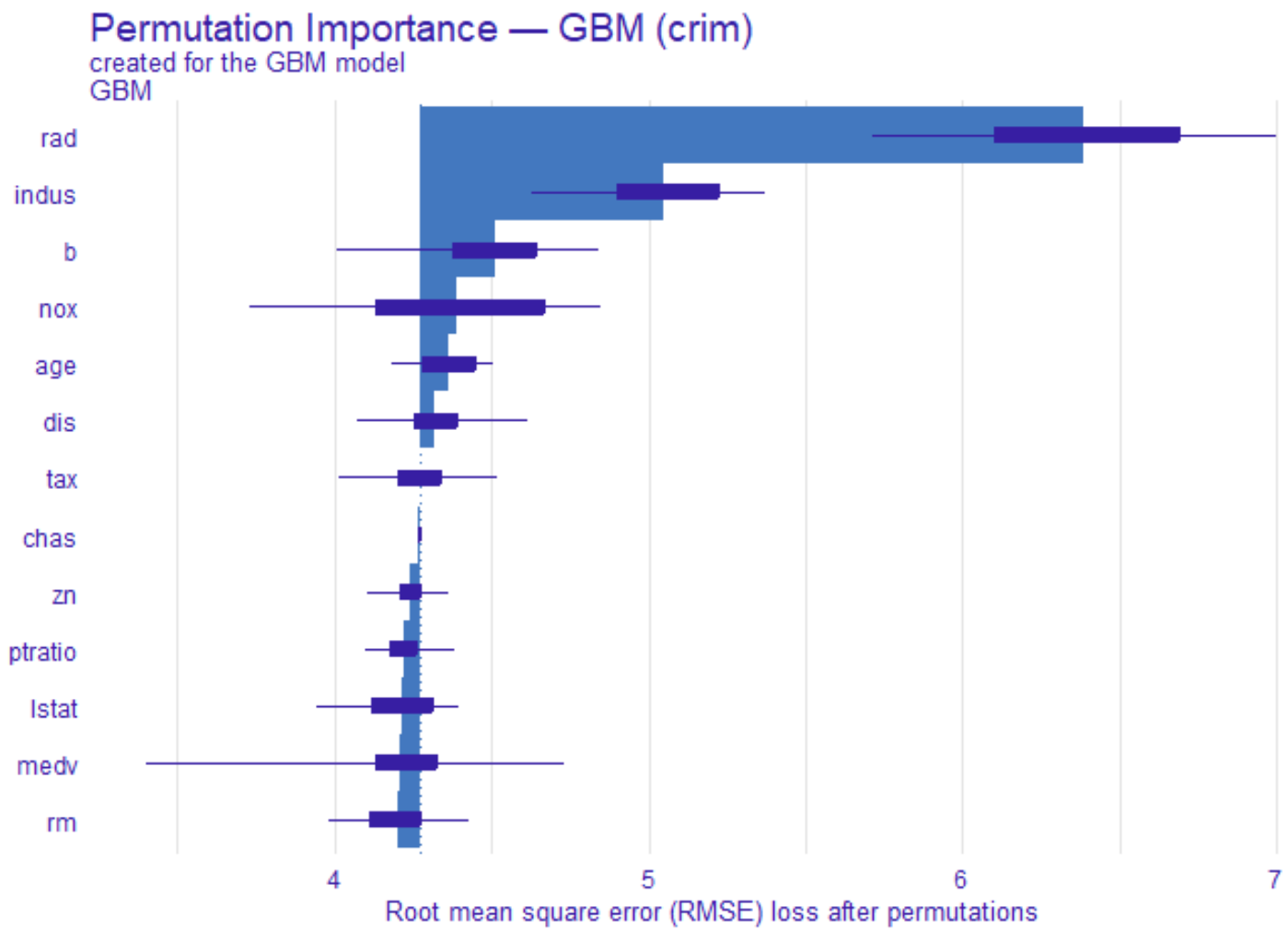
Model Performance



Interpretation (example narrative):

- GBM and RF outperform LM in RMSE and R^2.
- All models benefit from log-transform of `crim`.

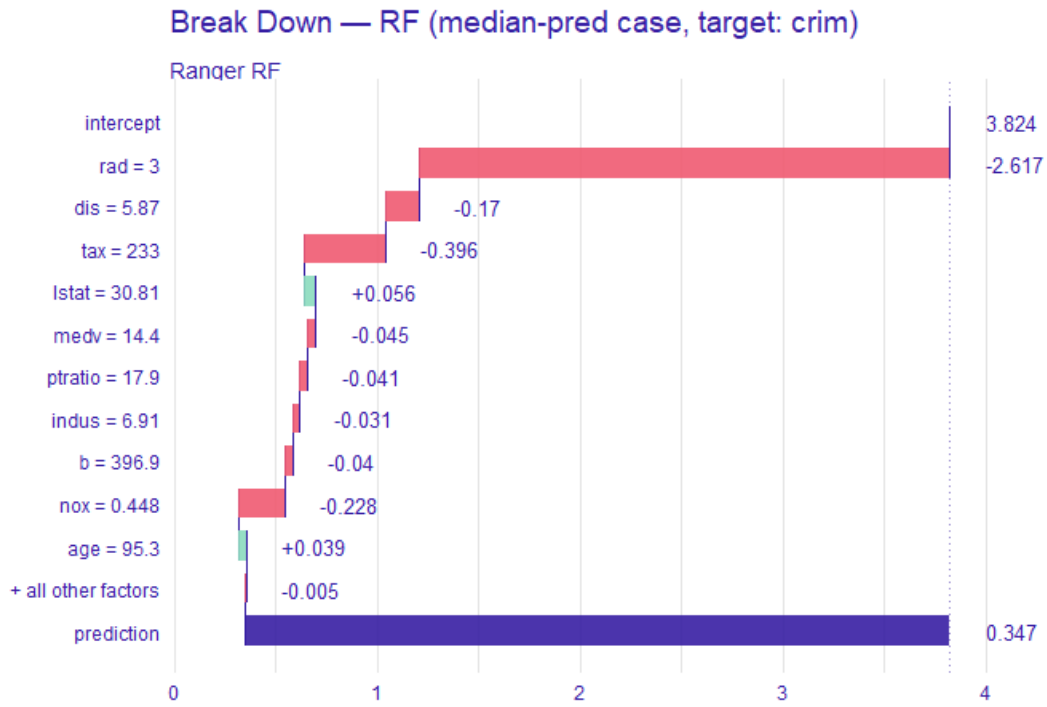
Variable Importance for GBM



Typical pattern:

- rad and `indus` often top predictor (poverty proxy).

Break Down and Shap Plots



Break Down Plots Plots (example narrative):

- As `rad` increases, predicted crime rises non-linearly.
- As `dis` increases, predicted crime falls.
- This reveals structural socio-economic patterns and individual contributions.

Reflective Discussions

Teaching Discussion:

- Which features push the prediction upward?
- Which features mitigate predicted crime?
- How might media or policy interpret such outputs?
- What risks arise if such a model guides policing or investment?

Overall Scaffolded Learning Pathway

Stage 1: Conceptual Foundations

- Introduce data literacy and critical questions.
- Discuss case narratives: crime, migration, inequality.
- Use simple models (means, correlations, linear regression).

Stage 2: Teach Machine Learning

- Introduce trees, random forests, GBM via analogies.
- Emphasize residuals, "learning from mistakes," gradient.

Stage 3: Interpretable AI + Civic Inquiry

- Use DALEX to visualize feature effects and local explanations.
- Guide students to critique fairness and policy implications.

Example Lesson Flow (90 Minutes)

Part 1 (20 min):

- Intro: How algorithms influence civic life.
- Discuss a media headline about crime or risk prediction.

Part 2 (30 min):

- Fit LM and GBM on Boston Housing (pre-coded).
- Show performance metrics and residuals.

Part 3 (30 min):

- Use DALEX plots for GBM.
- Group discussion: fairness, bias, alternative modeling.

Part 4 (10 min):

- Student reflection: short written answer or poll:
- "Would you trust this model to guide resource allocation? Why or why not?"

Critical Questions for Students

- Who collected this data, and for what purpose?
- Which variables are missing that might change the story?
- Are there groups systematically misrepresented or overrepresented?
- How does the model treat marginalized communities?
- What are the risks if such a model is:
  - deployed without interpretability?
  - used to justify punitive policies?
  - used without community consent?

Goal

Move students from "Is the model accurate?" to "Is the model *appropriate*, *fair*, and *accountable*?"

Possible Assessment Tasks

Individual or Group Assignments:

- **Model audit:** Students analyze DALEX outputs and write a short audit of a GBM model (variable importance, subgroup performance).
- **Narrative critique:** Compare a media narrative (e.g., "crime and migration") with model-based evidence; write a rebuttal.
- **Model redesign:** Rerun analysis excluding sensitive variables; compare performance and interpretability.
- **Policy brief:** Two-page memo to a city council explaining model limitations and recommendations.

From Models to Democracy

- Interpretable AI tools like DALEX turn abstract models into discussable objects.
- Critical data literacy enables:
  - informed critique of algorithmic decision-making
  - public engagement in AI governance debates
  - solidarity with communities affected by algorithmic harm
- Teaching interpretable AI in civic education:
  - links statistics to ethics and politics
  - prepares students for algorithmic citizenship

Key Takeaways

- **Interpretable AI** is not just a technical concern; it is a civic necessity.
- **Gradient Boosting Machines (GBM)** provide a realistic example of powerful, high-performing models that require explanation.
- **DALEX** offers:
  - global views (importance, PD/ALE),
  - local explanations (break-down, SHAP-like),
  - accessible visualizations for learners.
- **Critical data literacy** demands:
  - questioning data origins and purposes,
  - examining bias and power,
  - using models to unsettle harmful narratives, not reinforce them.

Selected References & Touchstones

- Boyd & Crawford (2012). Critical questions for big data.
- Erickson, T., Engel, J. (2023). What goes before the CART? Introducing classification trees with Arbor and CODAP. Teaching Statistics, 45, S104-S113.
- Engel, J., Martignon, L. (2024). Data science for informed citizen: Learning at the intersection of data literacy, statistics and social justice. Revista Internacional de Pesquisa em Educação Matemática, 14(3), 1-13.
- Noble (2018). *Algorithms of Oppression*.
- Awe, O. O., Love, K., Vance, E. (2024). Fostering data science and statistics education in Africa via online team-based learning. In IASE Conference Proceedings Series.
- Biecek, P. (2018). DALEX: Explainers for complex predictive models in R. Journal of Machine Learning Research , 19 (84), 1-5.
- Giustini, D., Dastyar, V. (2024). Critical AI Literacy for Interpreting in the Age of AI. Interpreting and society, 4(2), 196-213.

Thank you!

Questions?

Contact: olushina.awe@ph-ludwigsburg.de,
olawaleawe@gmail.com

Collaborations are warmly welcome.

Figure: StatLudPol Team at PH Ludwigsburg