Analyzing Political Risk Models

COMPSS 224B: Quantitative Political Risk

Iris Malone, PhD and Mark Rosenberg, PhD

May 9, 2025

Announcements

- Final Project
- Iris' Wed (14 May) OH 1-2 PT

Recap

Where We've Been:

- 7 deadly sins of political risk modeling stem from pre-set ideas about DGP and patterns
- Gauss-Markov assumptions are often violated, making parametric models problematic
- Non-parametric methods more flexible, but less interpretable
- Class imbalance distorts model evaluation

Recap

Where We've Been:

- 7 deadly sins of political risk modeling stem from pre-set ideas about DGP and patterns
- Gauss-Markov assumptions are often violated, making parametric models problematic
- Non-parametric methods more flexible, but less interpretable
- Class imbalance distorts model evaluation

New Terminology:

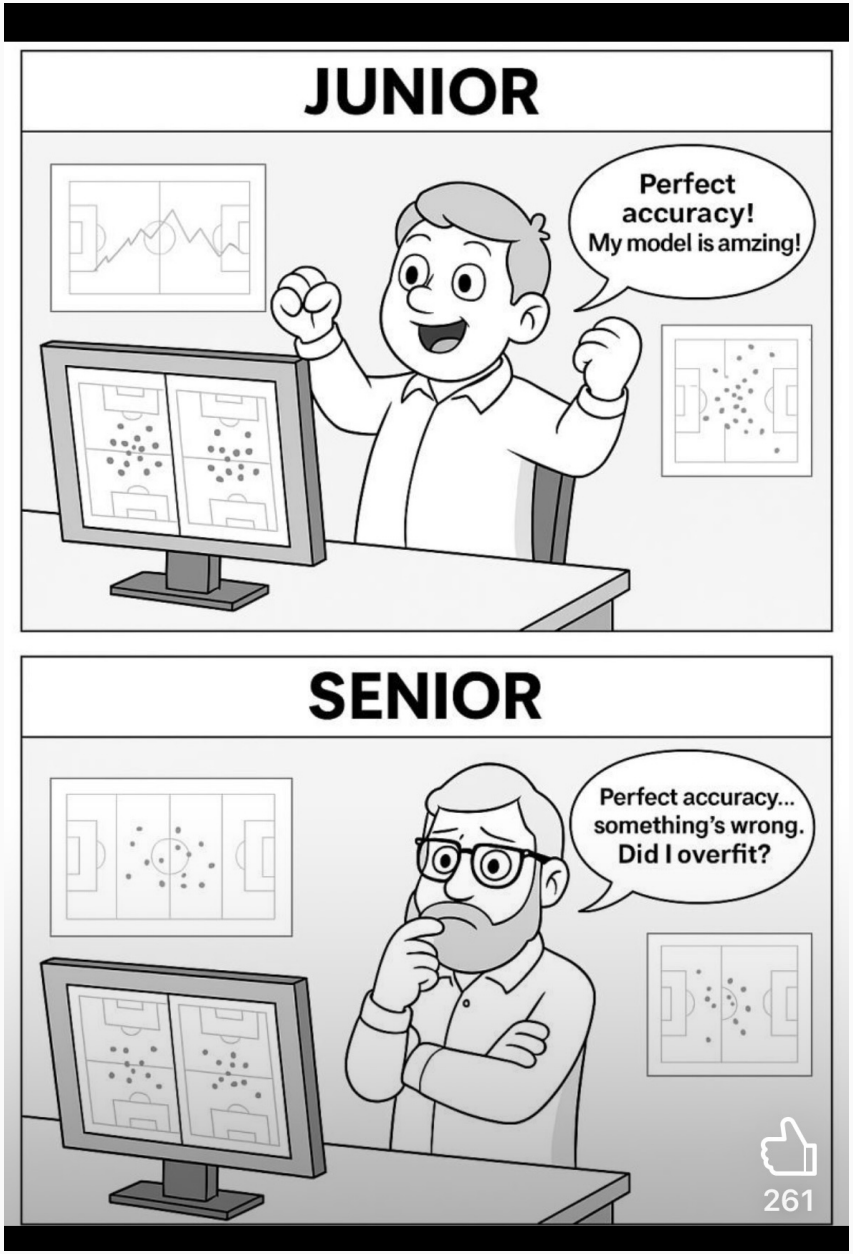
- Conditional inference tree
- No Information Rate (NIR)
- Synthetic Minority Oversampling Technique (SMOTE)

Agenda

1. Why We Need Interpretable ML
2. Types of Interpetable ML
    - By-Design
    - Post-Hoc Methods
3. Trust
4. Explainability
5. Debugging
6. Course Wrap-Up

Why We Need Interpretable ML

Motivation: ML can feel like magic...or sorcery.



Fear of AI and ML “sorcery” is growing...

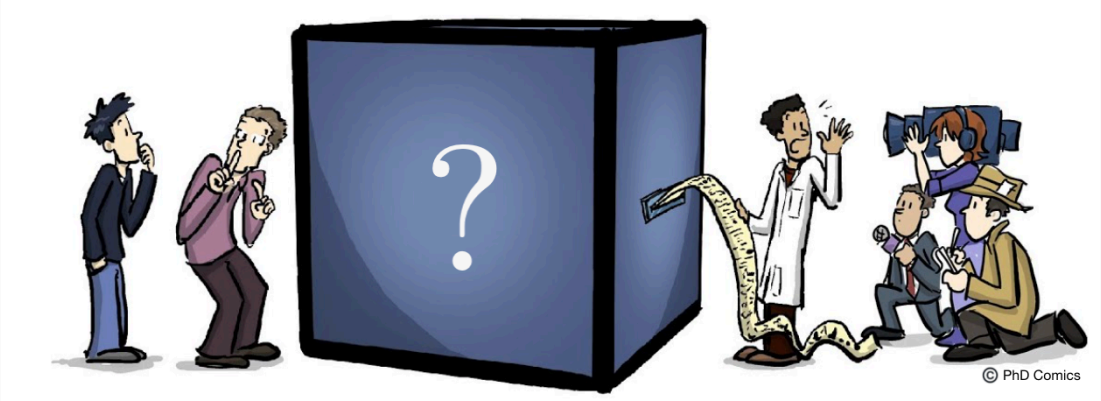


Figure 2: Machine Learning as a Black Box

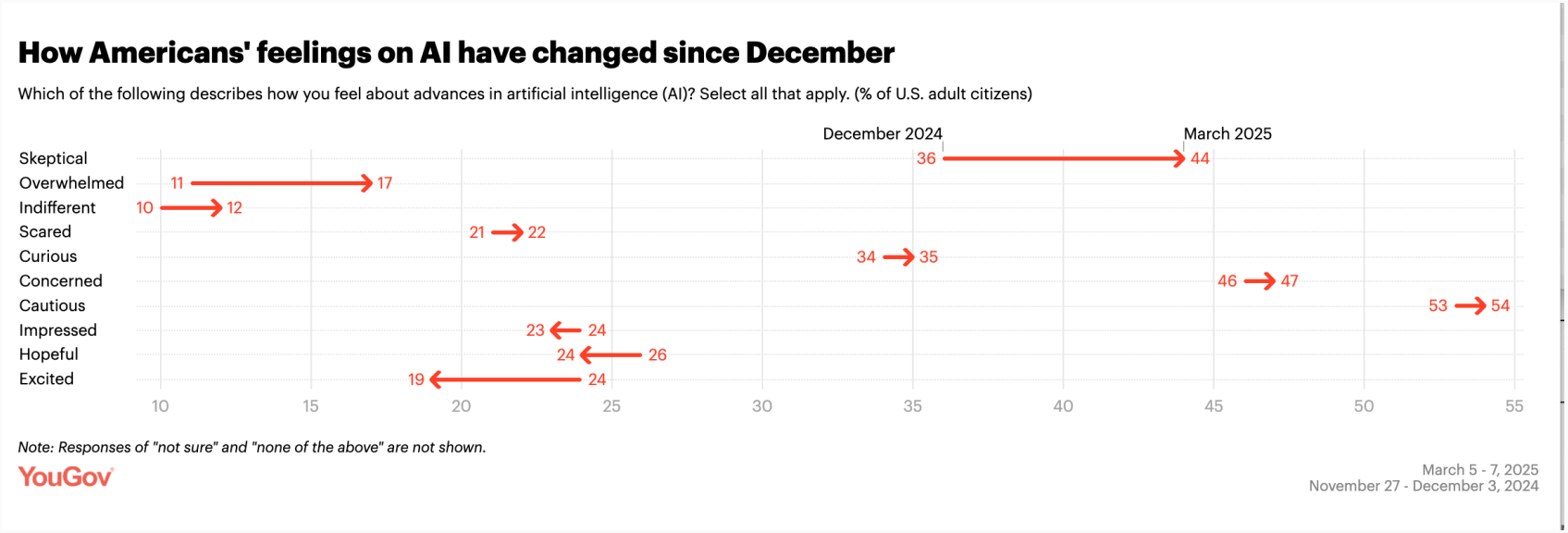


Figure 3: YouGov Survey, March 2025

C-suite seems trepidatious about it...

To some, AI seems like a dangerous abyss that demands cautionary steps. But the real risks come from inaction. What is behind this AI paralysis?

One of the most significant barriers to AI adoption in the C-suite is a fundamental lack of understanding about the technology. CEOs, especially those who didn't grow up in the tech industry, often feel overwhelmed by the complexity of technology. While they might be familiar with AI in a general sense, many do not have the deep technical knowledge required to fully grasp its potential applications or risks. As a result, they may fear making decisions that could expose the company to unforeseen challenges or lead to costly mistakes.

BUSINESS

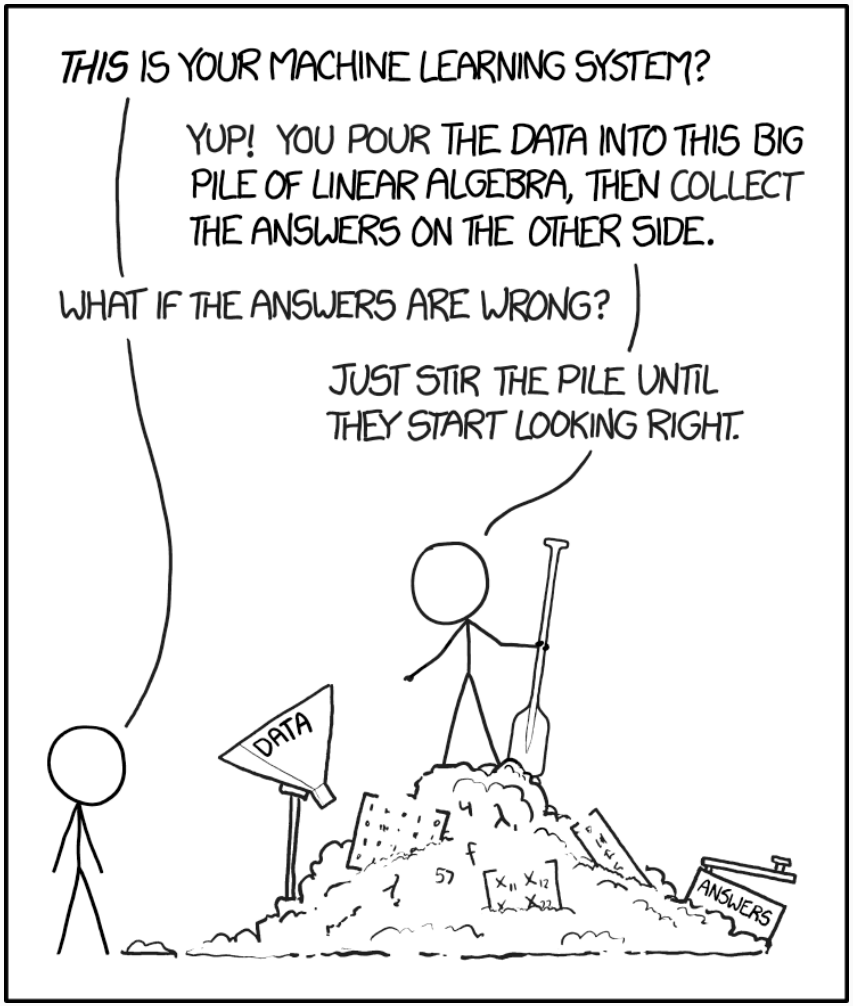
Why The C-Suite Is The Bottleneck For Implementing AI Projects

FB By Tim Houlne, Forbes Books Author.
for Forbes Books, AUTHOR POST | Paid Program

Apr 28, 2025, 10:22am EDT

Share Save

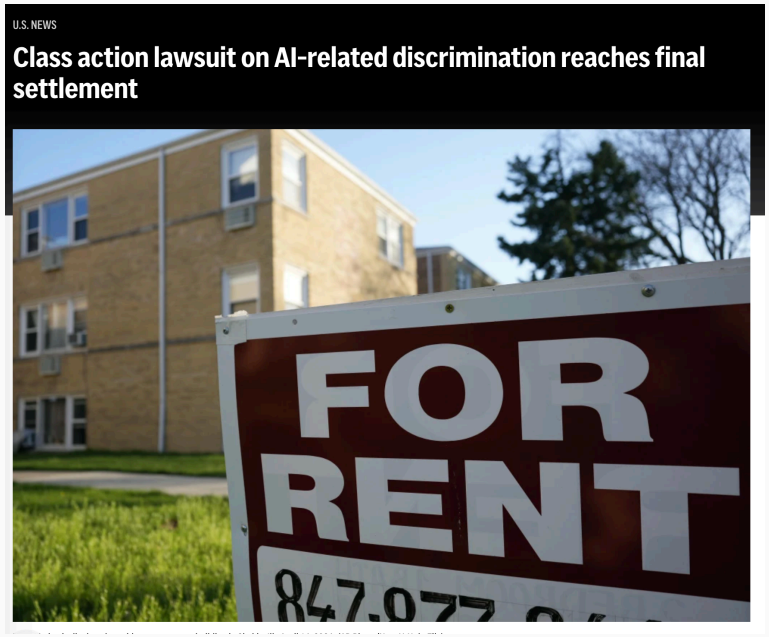
People are right to be skeptical: ML models are prone to abuse



..and can be overtly harmful

Louis et al v SafeRent: Rental application agency used discriminatory AI algorithm because trained on historical (discriminatory) data

> "**Decision-making algorithms, such as the ones at issue here, are often opaque,**" Webber said. "**Vendors who develop these algorithms are not willing to disclose all the data they consider or how the data is weighted in score modeling.** This is gravely concerning to fair housing, employment, and civil rights advocates as potentially discriminatory bias can be easily coded into automated decision-making platforms. The ability to hold such vendors accountable is essential for full enforcement of the civil rights laws."



More broadly, it's dangerous, costly, and likely unethical to not understand ML workings...

- *Louis et al v SafeRent:* Discriminatory AI rental screening tool
- *Mobley v. Workday:* Discriminatory AI applicant screening tool (race/age)
- *Williams v. Wells Fargo:* 'Digital redlining' in credit risk and refinancing ML assessments
- *Stephen Schwartz (Levidow, Levidow & Oberman):* Sanctioned after using GenAI to write brief and cited (hallucinatory) legal cases
- ...

Goals of Interpretability

Definition: "**Interpretability** is the degree to which a human can understand the cause of a decision" (Biran and Cotton 2017)

Goals of Interpretability

Definition: "**Interpretability** is the degree to which a human can understand the cause of a decision" (Biran and Cotton 2017)

1. **Trust:** Why should we trust the model?

Goals of Interpretability

Definition: "**Interpretability** is the degree to which a human can understand the cause of a decision" (Biran and Cotton 2017)

1. **Trust:** Why should we trust the model?
2. **Explainability:** How did the model get from input to output?

Goals of Interpretability

Definition: "**Interpretability** is the degree to which a human can understand the cause of a decision" (Biran and Cotton 2017)

1. **Trust:** Why should we trust the model?
2. **Explainability:** How did the model get from input to output?
3. **Causality:** Why did the model make this connection?

Goals of Interpretability

Definition: "**Interpretability** is the degree to which a human can understand the cause of a decision" (Biran and Cotton 2017)

1. **Trust:** Why should we trust the model?
2. **Explainability:** How did the model get from input to output?
3. **Causality:** Why did the model make this connection?
4. **Transferability:** Does this pattern generalize?

Goals of Interpretability

Definition: "**Interpretability** is the degree to which a human can understand the cause of a decision" (Biran and Cotton 2017)

1. **Trust:** Why should we trust the model?
2. **Explainability:** How did the model get from input to output?
3. **Causality:** Why did the model make this connection?
4. **Transferability:** Does this pattern generalize?
5. **Debugging:** Is the explanation meaningful?

Goals of Interpretability

Definition: "**Interpretability** is the degree to which a human can understand the cause of a decision" (Biran and Cotton 2017)

1. **Trust:** Why should we trust the model?
2. **Explainability:** How did the model get from input to output?
3. **Causality:** Why did the model make this connection?
4. **Transferability:** Does this pattern generalize?
5. **Debugging:** Is the explanation meaningful?
6. **Regulation:** Is the model fair? Reliable? Safe?

Good ML explanations pass the Thanskgiving table test

Ask yourself: would an aunt, uncle, grandparent understand your explanation?

Good ML explanations are ...

- Stories.
  Hook and narrative.
- Contrastive
- Concise
- Relatable
- Probable
- Generalizable

Good ML explanations are ...

- Stories.
  Hook and narrative.
- Contrastive
  Why X outcome and not Y outcome?
- Concise
- Relatable
- Probable
- Generalizable

Good ML explanations are ...

- Stories.
  Hook and narrative.
- Contrastive
  Why X outcome and not Y outcome?
- Concise
  The main driver is...
- Relatable
- Probable
- Generalizable

Good ML explanations are ...

- Stories.
  Hook and narrative.
- Contrastive
  Why X outcome and not Y outcome?
- Concise
  The main driver is...
- Relatable
  This is like...(Analogies, metaphors, historical examples)
- Probable

- Generalizable

Good ML explanations are ...

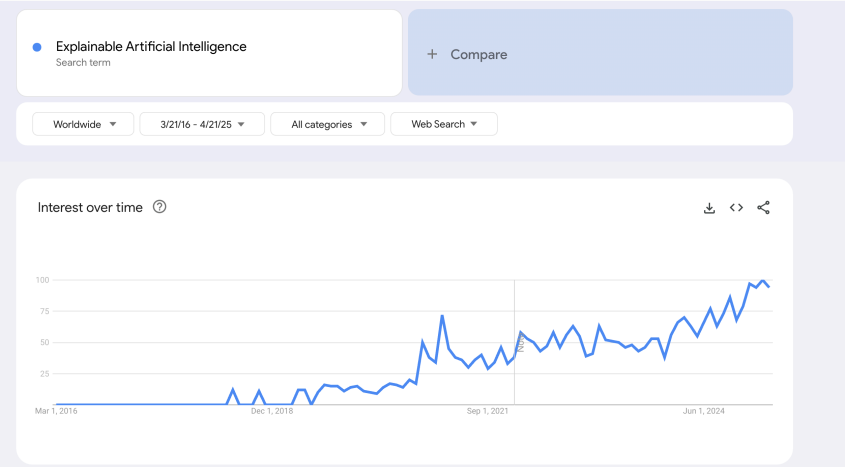
- Stories.
  Hook and narrative.
- Contrastive
  Why X outcome and not Y outcome?
- Concise
  The main driver is...
- Relatable
  This is like...(Analogies, metaphors, historical examples)
- Probable
  Consistent with prior beliefs (inevitable trade-off with trustworthiness)
- Generalizable

Good ML explanations are ...

- Stories.
  Hook and narrative.
- Contrastive
  Why X outcome and not Y outcome?
- Concise
  The main driver is...
- Relatable
  This is like...(Analogies, metaphors, historical examples)
- Probable
  Consistent with prior beliefs (inevitable trade-off with trustworthiness)
- Generalizable
  How often does the explanation apply?

Explainable AI (XAI) is increasingly part of corporate, legal, and government regulations.

Entity / Regulation	Explainability Requirement
EU General Data Protection Regulation (GDPR)	Individuals have the right to ”meaningful information about the logic involved” in AI algorithms using their data.
EU AI Act (2024)	Requires developers to document how algorithms make decisions.
JPM Model Review and Governance	Assesses the risk, explainability, and transparency of every ML model before deployment.
McKinsey QuantumBlack	Oversight by AI governance committees and an AI risk review.



Types of Interpetable ML

Types of Interpretable ML

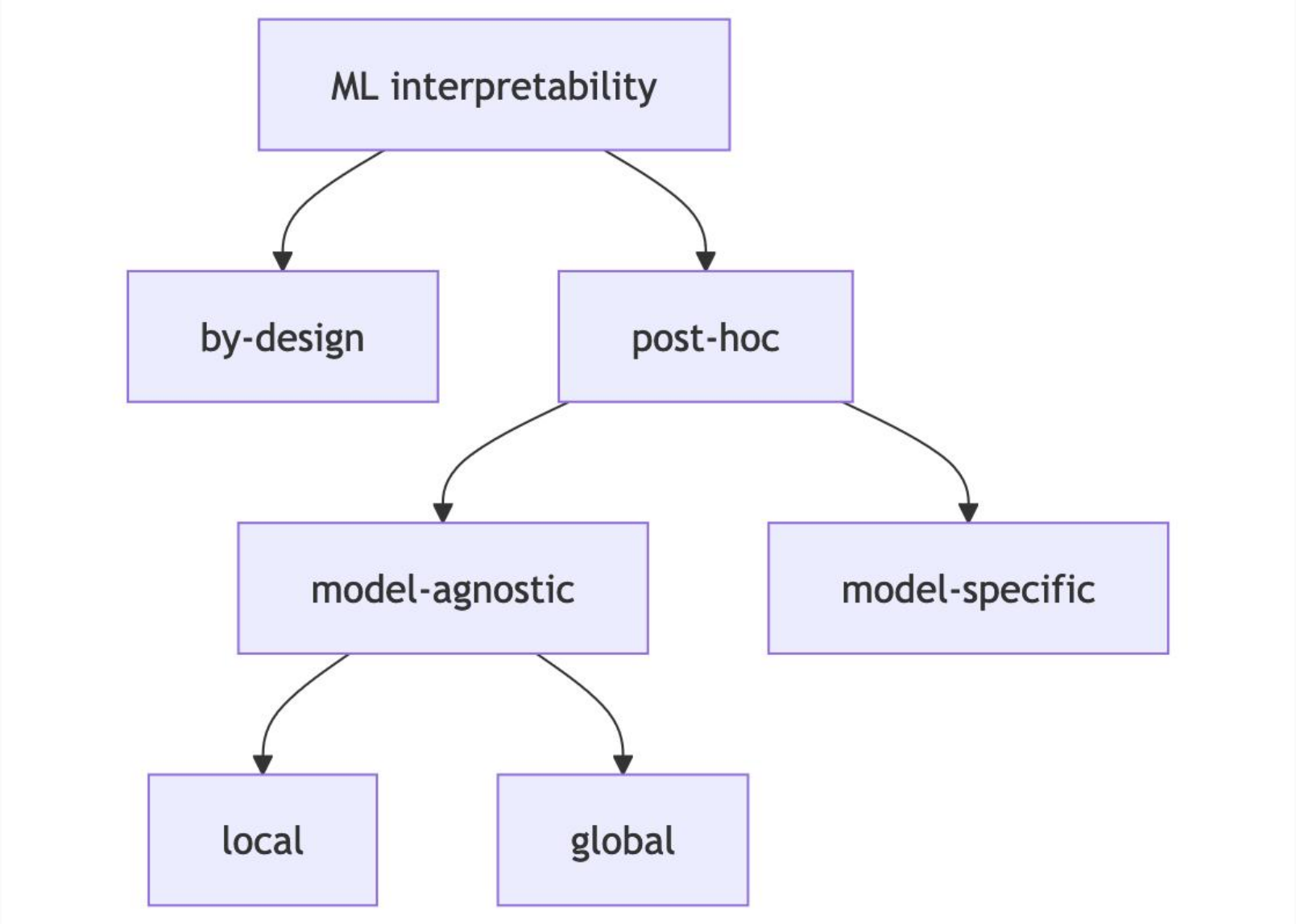


Figure 5: Source: Christopher Molnar

Types of Interpretable ML

By-Design

Post-Hoc

Types of Interpretable ML

By-Design

- Intrinsic or inherent interpretability
- What you see is what you get

Post-Hoc

Types of Interpretable ML

By-Design

- Intrinsic or inherent interpretability
- What you see is what you get

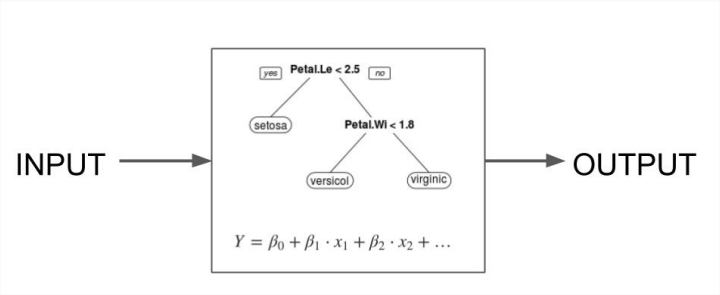


Figure 6: Source: Christopher Molnar

Post-Hoc

Types of Interpretable ML

By-Design

- Intrinsic or inherent interpretability
- What you see is what you get

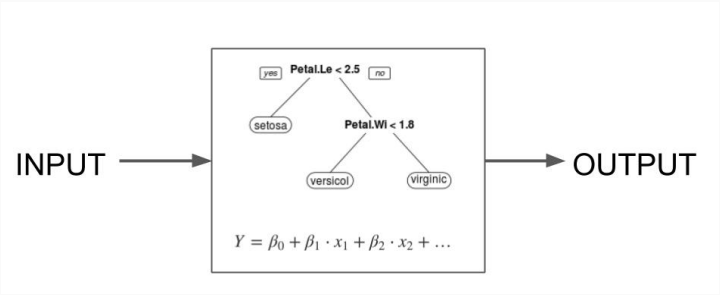


Figure 6: Source: Christopher Molnar

Post-Hoc

- Extrinsic interpretability
- Additional assembly required

Types of Interpretable ML

By-Design

- Intrinsic or inherent interpretability
- What you see is what you get

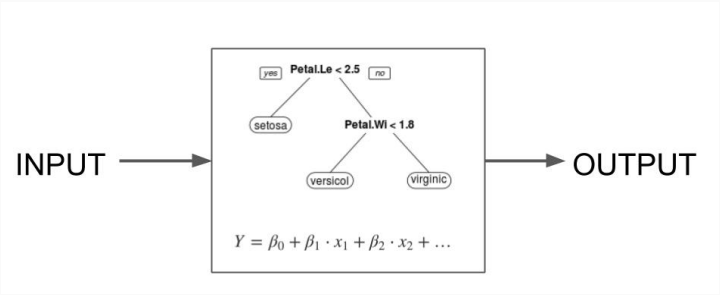


Figure 6: Source: Christopher Molnar

Post-Hoc

- Extrinsic interpretability
- Additional assembly required

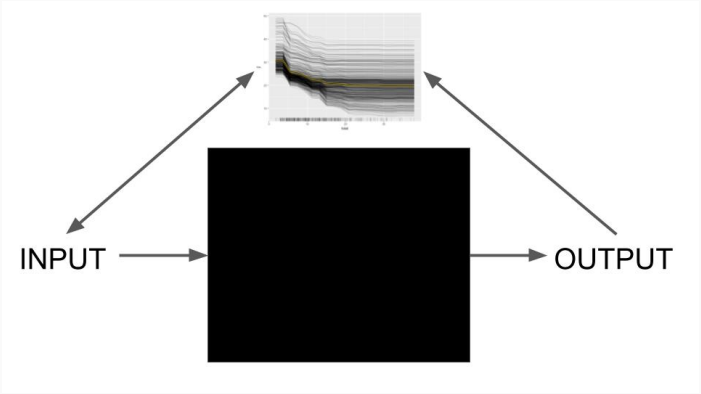


Figure 7: Source: Christopher Molnar

Types of Interpretable ML

By-Design

- Intrinsic or inherent interpretability
- What you see is what you get

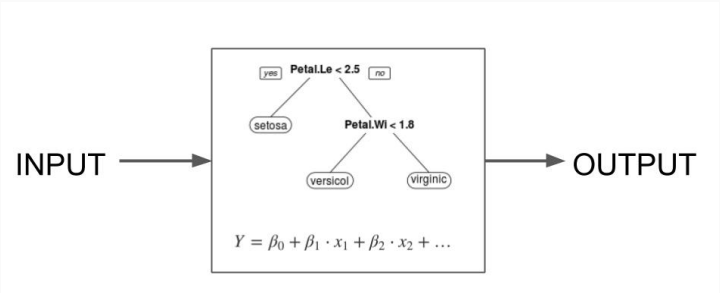


Figure 6: Source: Christopher Molnar

Post-Hoc

- Extrinsic interpretability
- Additional assembly required

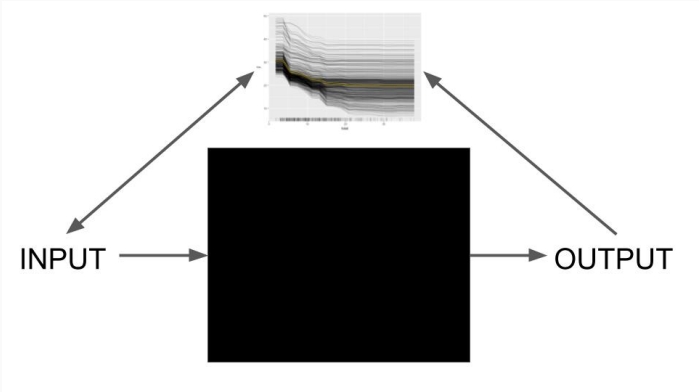


Figure 7: Source: Christopher Molnar

Warning: Lines may be blurrier than they appear, e.g. logit odds ratios or feature importance

By-design interpretability

Main idea: Models provide clear and straight-forward interpretation without any extra work.

Examples:

By-design interpretability

Main idea: Models provide clear and straight-forward interpretation without any extra work.
Examples:

- **Linear/Logit Regressions:** Models produce coefficients (and p-values) that can be directly interpreted.
- **Decision Trees:** Models create decision rules that split the data into buckets, and each leaf can be easily followed by following the splitting rules.

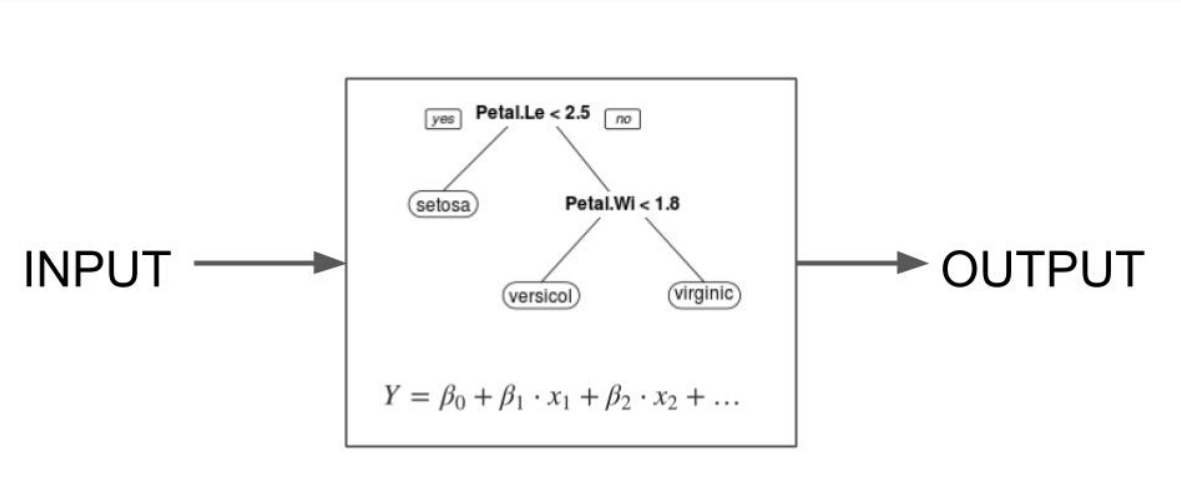


Figure 8: Source: Christopher Molnar

By-Design Interpretable Tools

Check: Can I answer the original prediction problem and explain it?

By-Design Interpretable Tools

Check: Can I answer the original prediction problem and explain it?

- Regression tables
- Coefficient (box-and-whisker) plots
- Marginal effects (interactions) plots
- Random effect plots
- Curve-fitting plots (confidence interval plots)
- Decision Tree

Motivation: Fearon and Laitin (2003)

Question: What causes civil wars?

Motivation: Fearon and Laitin (2003)

Question: What causes civil wars?

- Motive:
  - Ethnic/Religious Fractionalization
  - Oil/Natural Resource Rents
  - Population
  - Regime Type
- Opportunity:
  - GDP/Capita
  - Mountain Terrain
  - New State
  - Non-Continguous

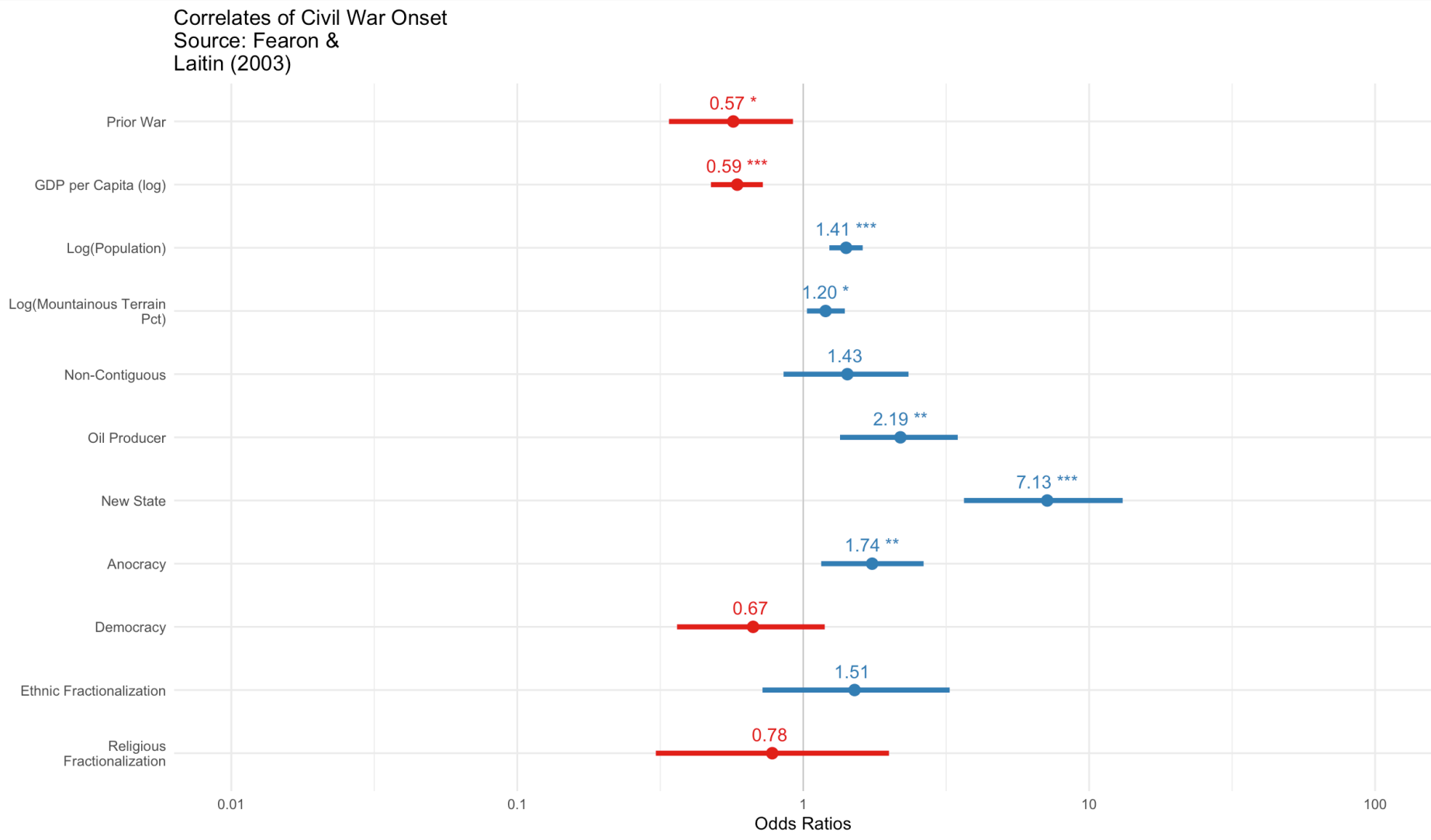
Motivation: Fearon and Laitin (2003)

Question: What causes civil wars?

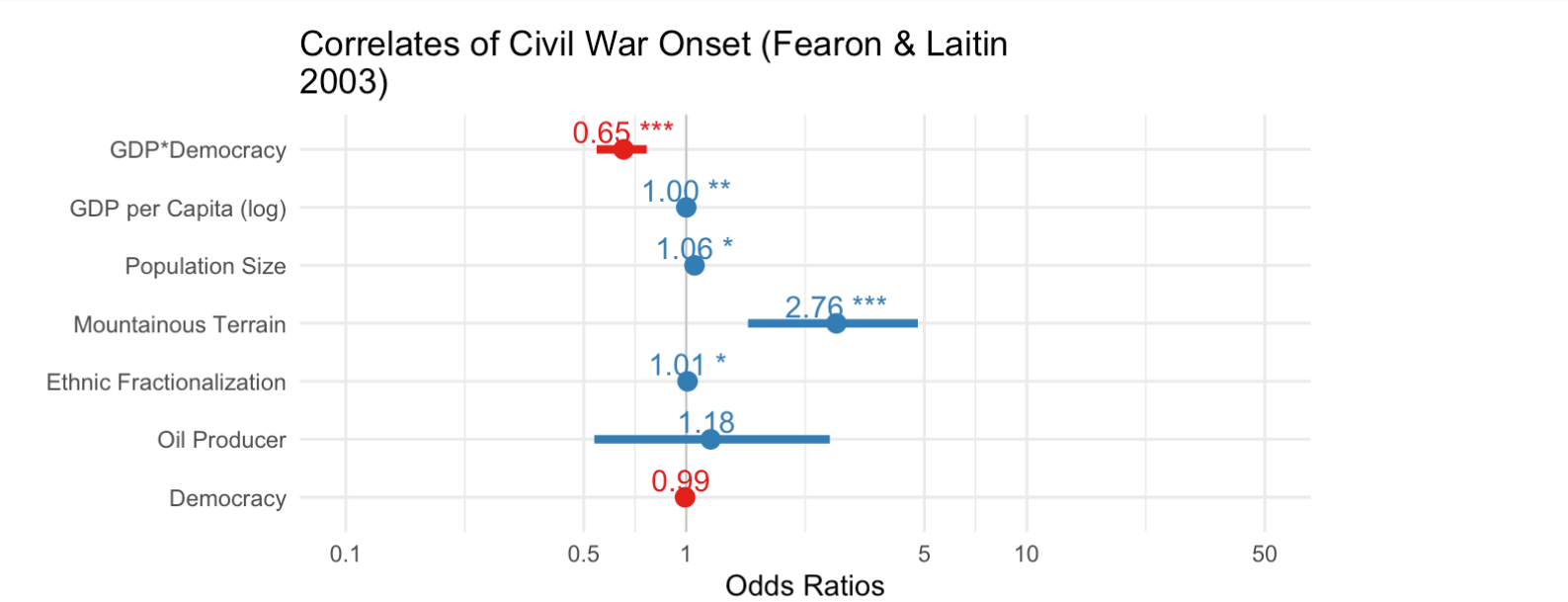
- Motive:
  - Ethnic/Religious Fractionalization
  - Oil/Natural Resource Rents
  - Population
  - Regime Type
- Opportunity:
  - GDP/Capita
  - Mountain Terrain
  - New State
  - Non-Continguous
- **Method:** Logit regression, 1945-2003 (later updated to 2012)

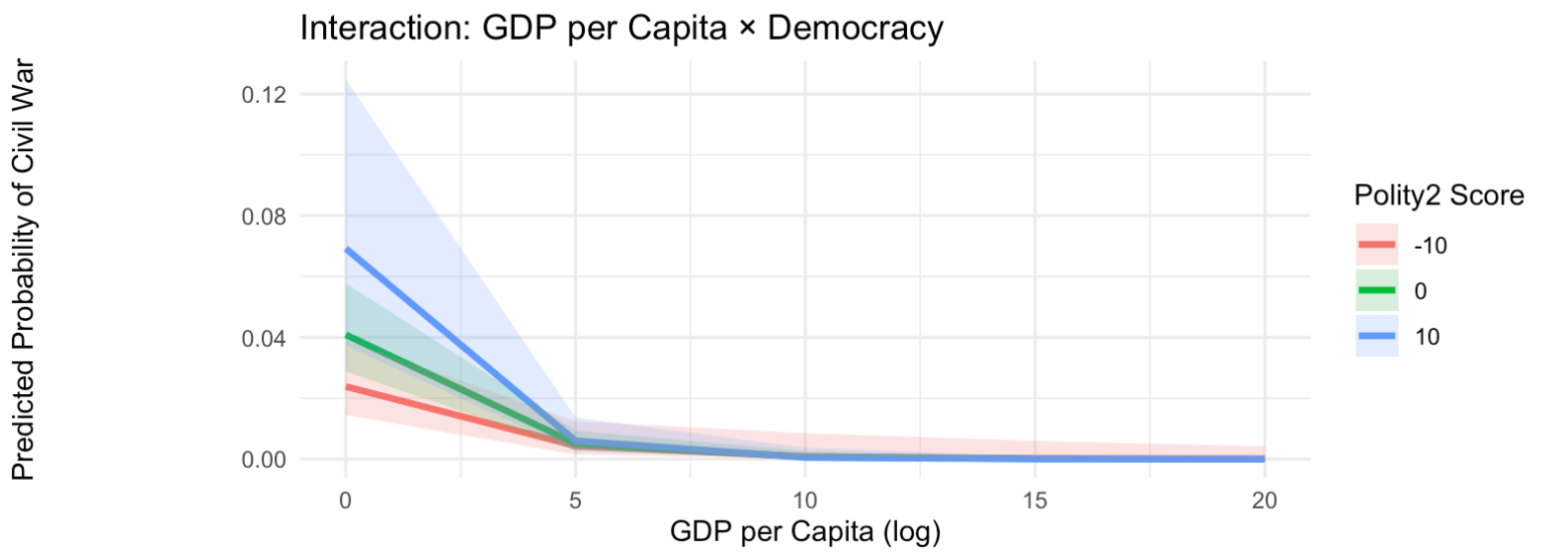
See also Hill and Jones (2014)

Coefficient Plots: This is a story about...



Marginal Effects: For a given regime type, civil wars are more likely when...

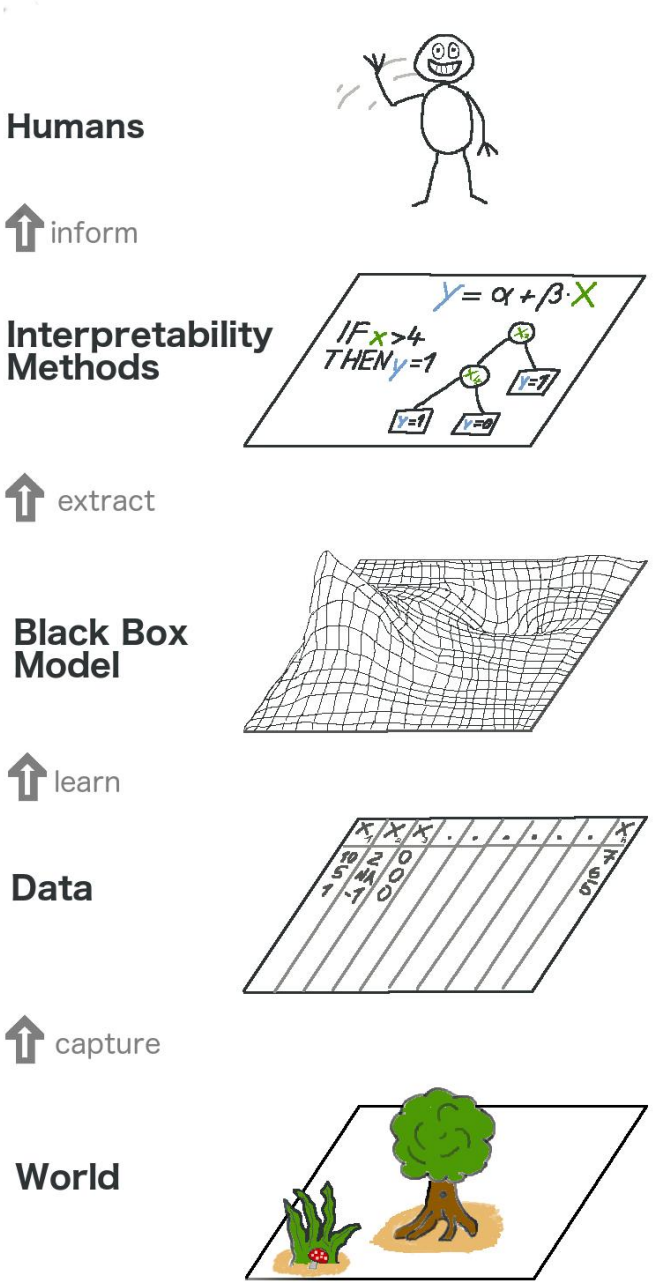




Post-hoc methods

- **Model-agnostic**
  - Ignore the black box
  - Outcome-focused
  - E.g. permutation importance, SHAP, PDP
- **Model-specific**
  - Focus on the black box
  - Pattern-focused
  - Typically algorithm-specific, e.g. neurons in a neural net or odds ratio in logit

Model-Agnostic Methods



Goals and Methods

Goal	Question	Toolkit
Trust	Does the model solve the prediction problem?	MSE, RMSE, MAE Confusion Matrix ROC/AUC Separation Plot

Goals and Methods

Goal	Question	Toolkit
Trust	Does the model solve the prediction problem?	MSE, RMSE, MAE Confusion Matrix ROC/AUC Separation Plot
Explainability	What is driving the model?	Permutation Importance SHAP values Accumulated Local Effects (ALE)
	Why did the model predict Y_i?	Partial Dependency Plot (PDP) Hypothesis testing, K-L, meta-learners

Goals and Methods

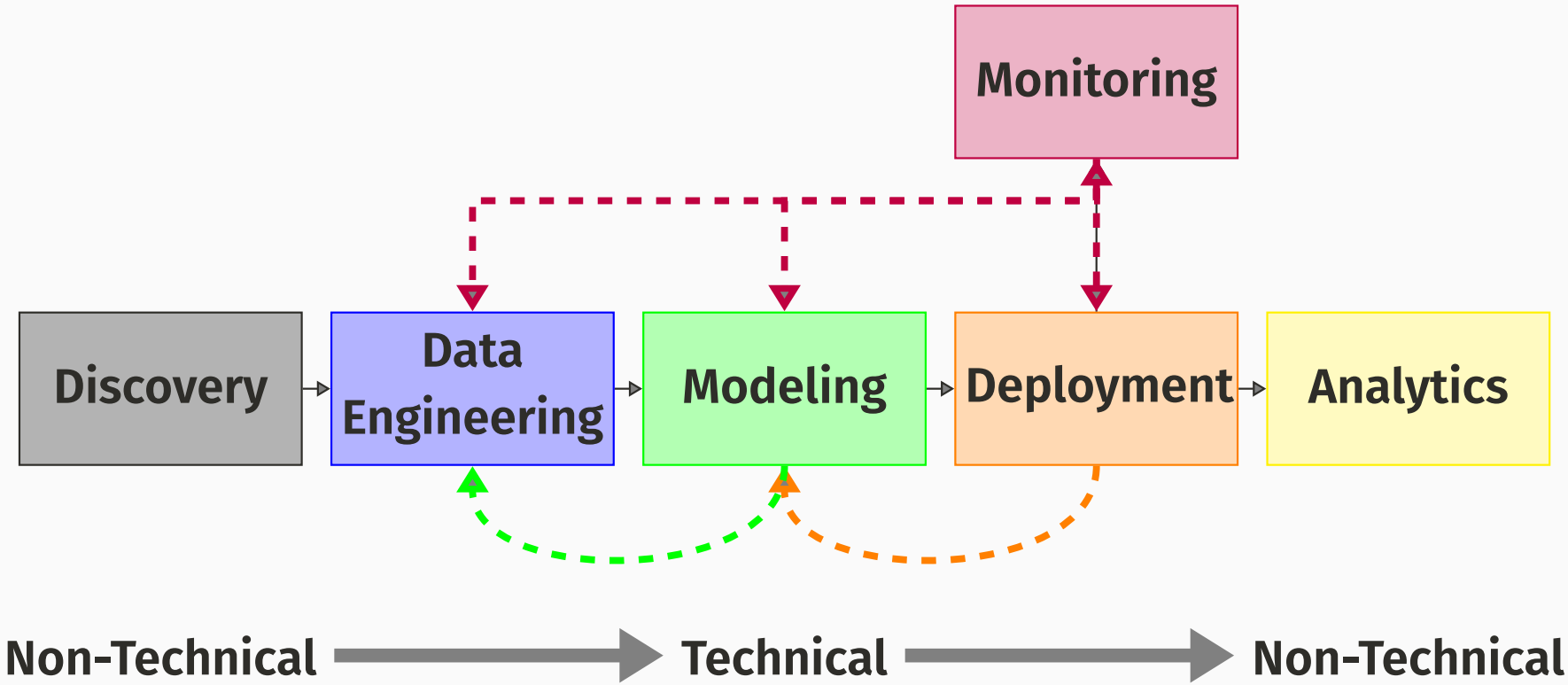
Goal	Question	Toolkit
Trust	Does the model solve the prediction problem?	MSE, RMSE, MAE Confusion Matrix ROC/AUC Separation Plot
Explainability	What is driving the model?	Permutation Importance SHAP values Accumulated Local Effects (ALE)
	Why did the model predict Y_i?	Partial Dependency Plot (PDP) Hypothesis testing, K-L, meta-learners
Debugging	Is this explanation useful (actionable?)	Adversarial Examples Ceteris Paribus Profiles Cross-Validation Error Analysis Data Leakage Checks

Goals and Methods

Goal	Question	Toolkit
Trust	Does the model solve the prediction problem?	MSE, RMSE, MAE Confusion Matrix ROC/AUC Separation Plot
Explainability	What is driving the model?	Permutation Importance SHAP values Accumulated Local Effects (ALE)
	Why did the model predict Y_i?	Partial Dependency Plot (PDP) Hypothesis testing, K-L, meta-learners
Debugging	Is this explanation useful (actionable?)	Adversarial Examples Ceteris Paribus Profiles Cross-Validation Error Analysis Data Leakage Checks
Transferability	What if something changes...?	Counterfactuals Stress Tests Monte Carlo Simulations Synthetic Data*

Recall: MLOps Context

Analytics depends on being able to explain and interpret the ML works. If analyst can't answer these questions, then engineering or modeling likely needs to be re-evaluated.



XAI in practice

Goal	Question	Sample Use Case
Trust	Does the model solve the prediction problem?	
Explainability	What is driving the model? *Inference:* Why did the model predict Y_i?	
Debugging	Is this explanation useful? Is it actionable?	
Transferability	What is $\hat{y}_{(i)}$ if something changes...?	

h

XAI in practice

Goal	Question	Sample Use Case
Trust	Does the model solve the prediction problem?	Do Fearon and Laitin predict civil wars?
Explainability	What is driving the model? *Inference:* Why did the model predict Y_i?	
Debugging	Is this explanation useful? Is it actionable?	
Transferability	What is $\hat{y}_{(i)}$ if something changes...?	

h

XAI in practice

Goal	Question	Sample Use Case
Trust	Does the model solve the prediction problem?	Do Fearon and Laitin predict civil wars?
Explainability	What is driving the model? *Inference:* Why did the model predict Y_i?	What predicts civil wars?
Debugging	Is this explanation useful? Is it actionable?	
Transferability	What is $\hat{y}_{(i)}$ if something changes...?	

h

XAI in practice

Goal	Question	Sample Use Case
Trust	Does the model solve the prediction problem?	Do Fearon and Laitin predict civil wars?
Explainability	What is driving the model? *Inference:* Why did the model predict Y_i?	What predicts civil wars? How does GDP affect civil war onset?
Debugging	Is this explanation useful? Is it actionable?	
Transferability	What is $\hat{y}_{(i)}$ if something changes...?	

h

XAI in practice

Goal	Question	Sample Use Case
Trust	Does the model solve the prediction problem?	Do Fearon and Laitin predict civil wars?
Explainability	What is driving the model? *Inference:* Why did the model predict Y_i?	What predicts civil wars? How does GDP affect civil war onset?
Debugging	Is this explanation useful? Is it actionable?	Is the GDP-conflict link robust? What can users do with this insight?
Transferability	What is $\hat{y}_{(i)}$ if something changes...?	

h

XAI in practice

Goal	Question	Sample Use Case
Trust	Does the model solve the prediction problem?	Do Fearon and Laitin predict civil wars?
Explainability	What is driving the model? *Inference:* Why did the model predict Y_i?	What predicts civil wars? How does GDP affect civil war onset?
Debugging	Is this explanation useful? Is it actionable?	Is the GDP-conflict link robust? What can users do with this insight?
Transferability	What is $\hat{y}_{(i)}$ if something changes...?	What if... -Syria democratized in 2010? -U.S. faces an economic downturn?

h

Trust

Trust: Does the model solve the prediction problem?

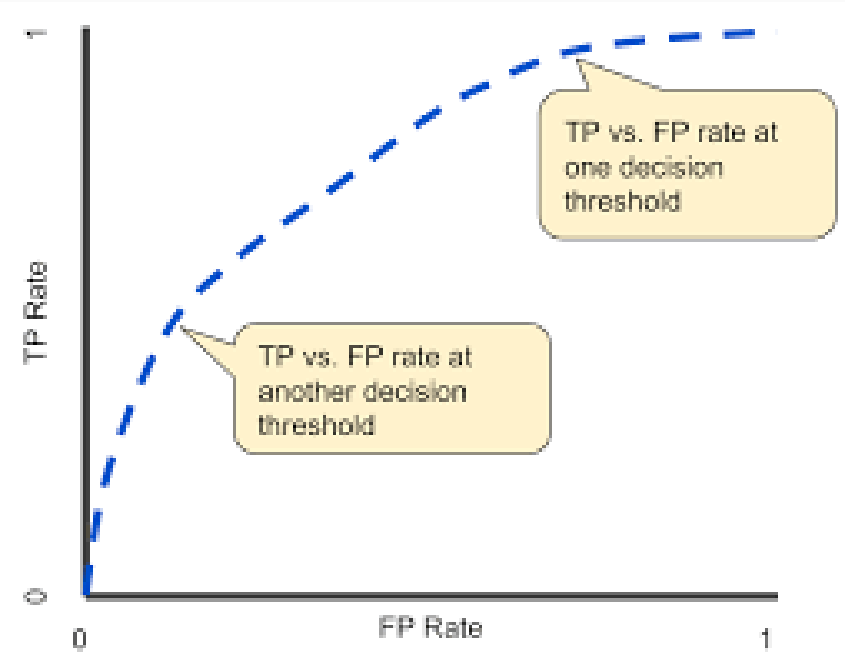
Assessment: Check the predictive power of the model

- Metrics: MSE, RMSE, MAE, etc
- Confusion matrix (sensitivity, specificity, precision-recall, f1, kappa)
- ROC/AUC
- Separation plot

Receiver Operating Curve (ROC)

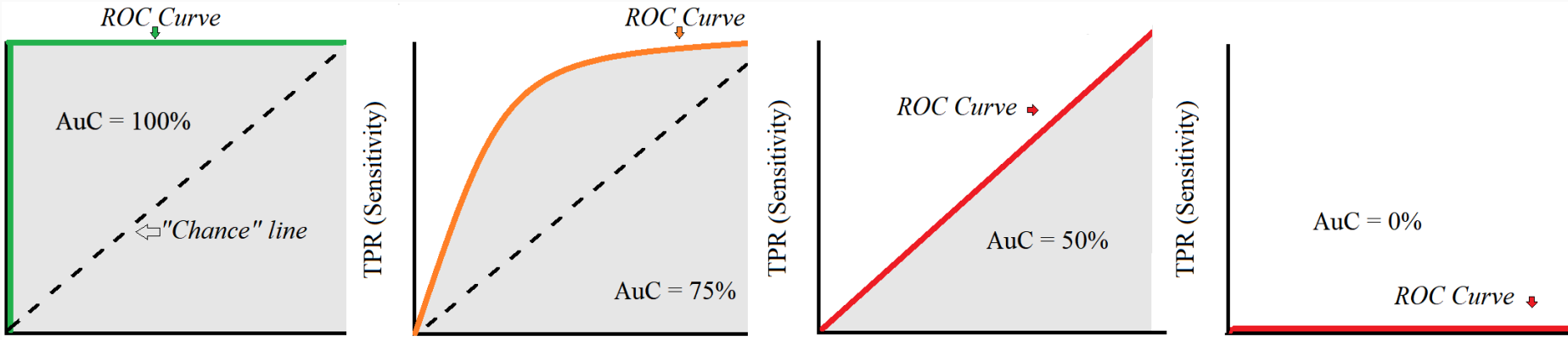
Main Idea:

- Evaluate performance of a classification model at *all* classification thresholds.
- Plot model performance of true positive rate vs false positive rate
- Boundary line tells us relative model performance



Area Under the Curve

Integral of ROC → Area Under the Curve (AUC)



Perfect Model has AUC = 1

Separation Plot

A **separation plot** is a popular visual tool of a model's predictive power

- Tells us extent to which model's predicted probability maps onto actual outcome
- Easy to visualize FP vs TP (like ROC)
- Easy to visualize sparsity of data and class distribution

See Greenhill et al. (2011) "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models" for more.

Motivating Example: Predict War and Peace

Have observations $\{A, B, C, D, E, F\}$

	Predicted War	Predicted Peace
Actual War	{C, E}	{F}
Actual Peace	{A, D}	{B}

Motivating Example: Predict War and Peace

TABLE 1 Sample Data

Country	Actual Outcome (y)	Fitted Value ($\hat{p}$)
A	0	0.774
B	0	0.364
C	1	0.997
D	0	0.728
E	1	0.961
F	1	0.422

Brier Score

TABLE 3 Calculation of Brier Scores

Country	Actual Outcome (y)	Fitted Value ($\hat{p}$)	Brier Score ($\hat{p} - y)^2$
A	0	0.774	0.599
B	0	0.364	0.132
C	1	0.997	0.000
D	0	0.728	0.530
E	1	0.961	0.002
F	1	0.422	0.334

Separation Plot

Sort by $\hat{p}$ and color code them: red if actual event happened and tan if no event happened

TABLE 4 Rearrangement (and Coloring) of the Data Presented in Table 1 for Use in the Separation Plot

Country	Fitted Value ($\hat{p}$)	Actual Outcome (y)
B	0.364	0
F	0.422	1
D	0.728	0
A	0.774	0
E	0.961	1
C	0.997	1

Separation Plot

Take 5 observations sorted by $\hat{p}$ and color code by whether event actually happened

FIGURE 2 Separation Plot Representing the Data Presented in Table 1

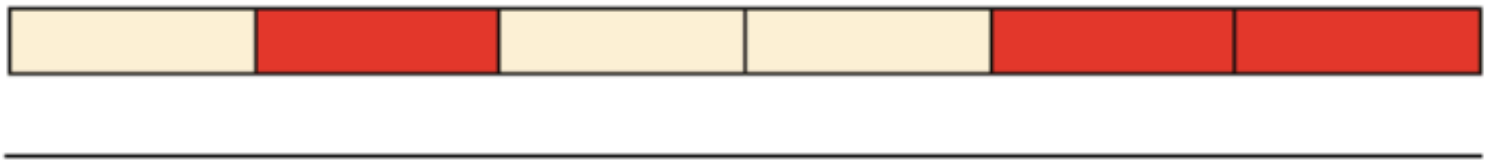
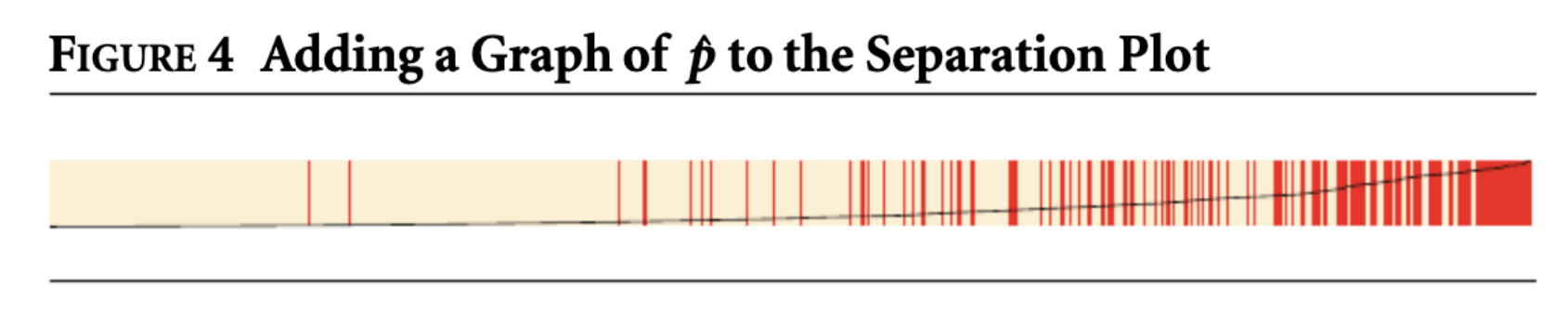


FIGURE 3 Separation Plot for a Larger Data Set

Separation Plot



Expand to larger set of occurrences and add black line for $\hat{p}$.
Sort x-axis by $\hat{p}$ to compare events versus predicted probabilities

Separation Plot

FIGURE 5 Adding the Expected Number of Events

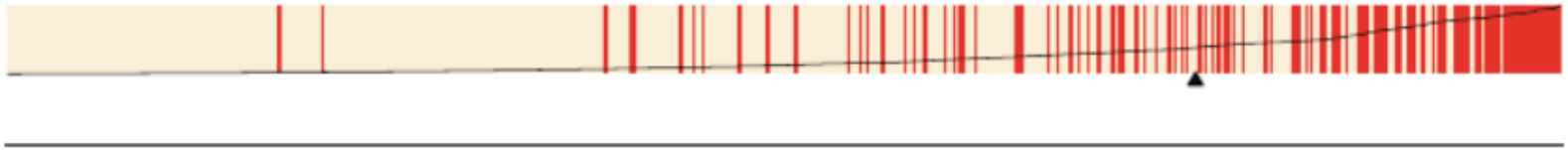


FIGURE 6 A "Perfect" Model for the Same Data Used in Figure 5

If model was perfect, we'd see complete event color-coded separation

Trust: Do Fearon and Laitin predict civil wars?

Muchlinski et al. (2016) "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data":

> Prediction is a contentious issue in the discipline of political science. Political scientists are generally taught not to be especially concerned with model fit, but to emphasize the estimation of causal parameters instead....One consequence, which we demonstrate below, is that **most statistical models of civil war onset have exceedingly weak predictive power.**

Example: Fearon and Laitin (2003) ROC

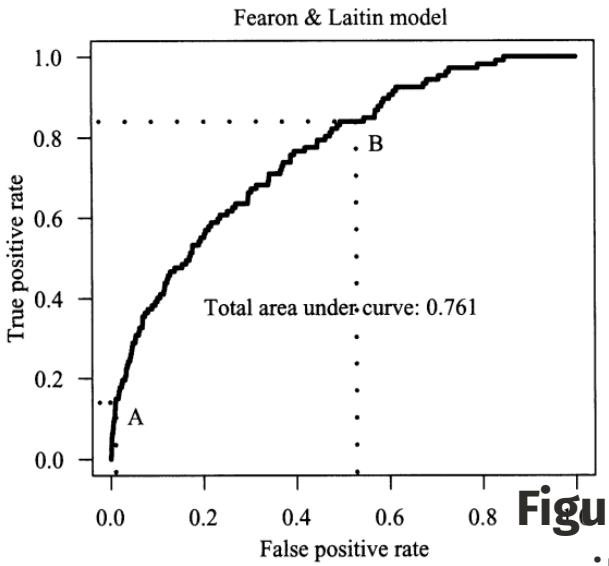


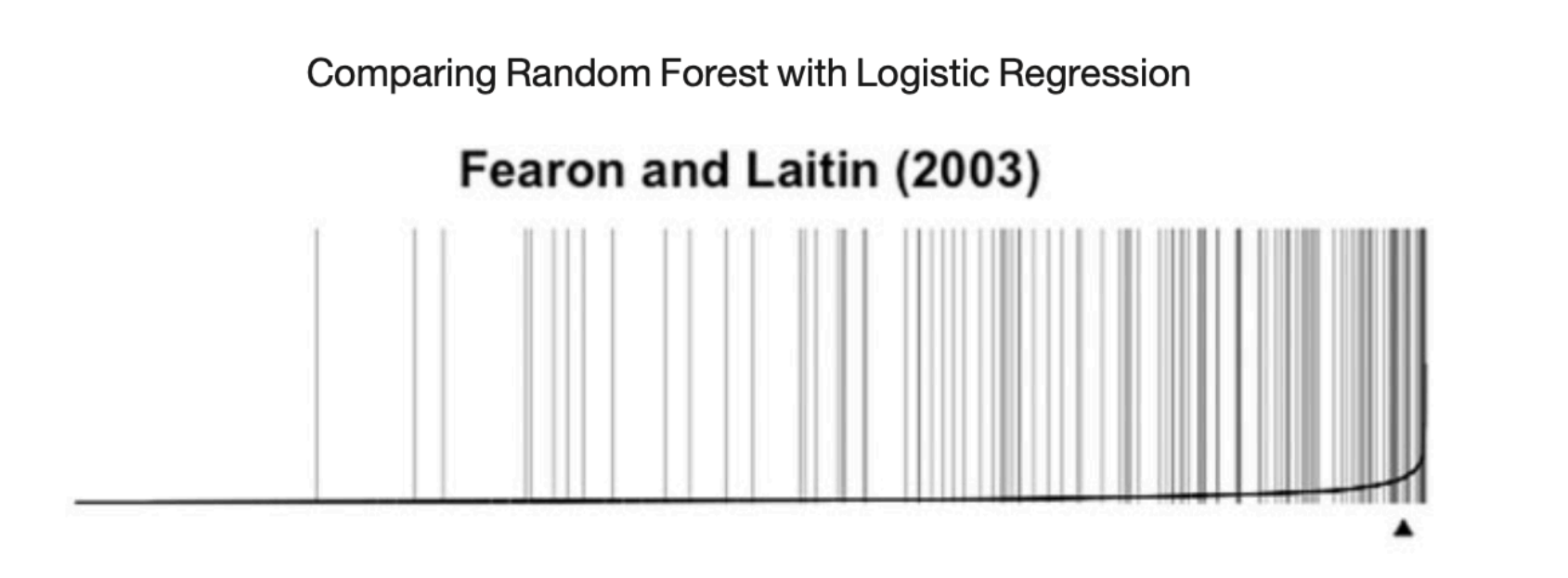
Figure 1. ROC plots

Table III. Number of correctly predicted onsets and false positives at varying cut-points

	Fearon & Laitin model	
Threshold	*Correctly predicted*	*False positives*
0.5	0/107	0
0.3	1/107	3
0.1	15/107	66

Figure 9: Source: Ward, Greenhill, and Bakke (2010): "The perils of policy by p-value: Predicting civil conflicts"

Fearon and Laitin (2003) Separation Plot



Explainability

Explainability: What is driving Y and why?

Feature Contribution:

- Permutation Importance
- SHAP

Feature Influence:

- SHAP
- Partial dependency plots (PDP)
- Accumulated Local Effects (ALE)
- Local Interpretable Model-Agnostic Explanations (LIME)
- Individual Conditional Expectation (ICE)

Inference:

- Multiple testing (e.g., F-test, Kullback-Leiber)
- Cross-validation analysis

Explainability: What is driving Y?

Problem: We want to identify the most relevant (and influential) predictors in the model. How?

Explainability: What is driving Y?

Problem: We want to identify the most relevant (and influential) predictors in the model. How?

Solution: **Permutation Importance**

- Create new feature set X' that randomly shuffles one feature's values, holding all others at 'true' (observed) values
- Re-run the model on X' and and calculate how much the model's performance metric (e.g., accuracy, RMSE, AUC) drops
- Repeat this across multiple resamples
- Average the results to get variable importance estimates.
- The larger the drop in performance (worse the model is), the more important the feature.

Example of Variable Importance Plot

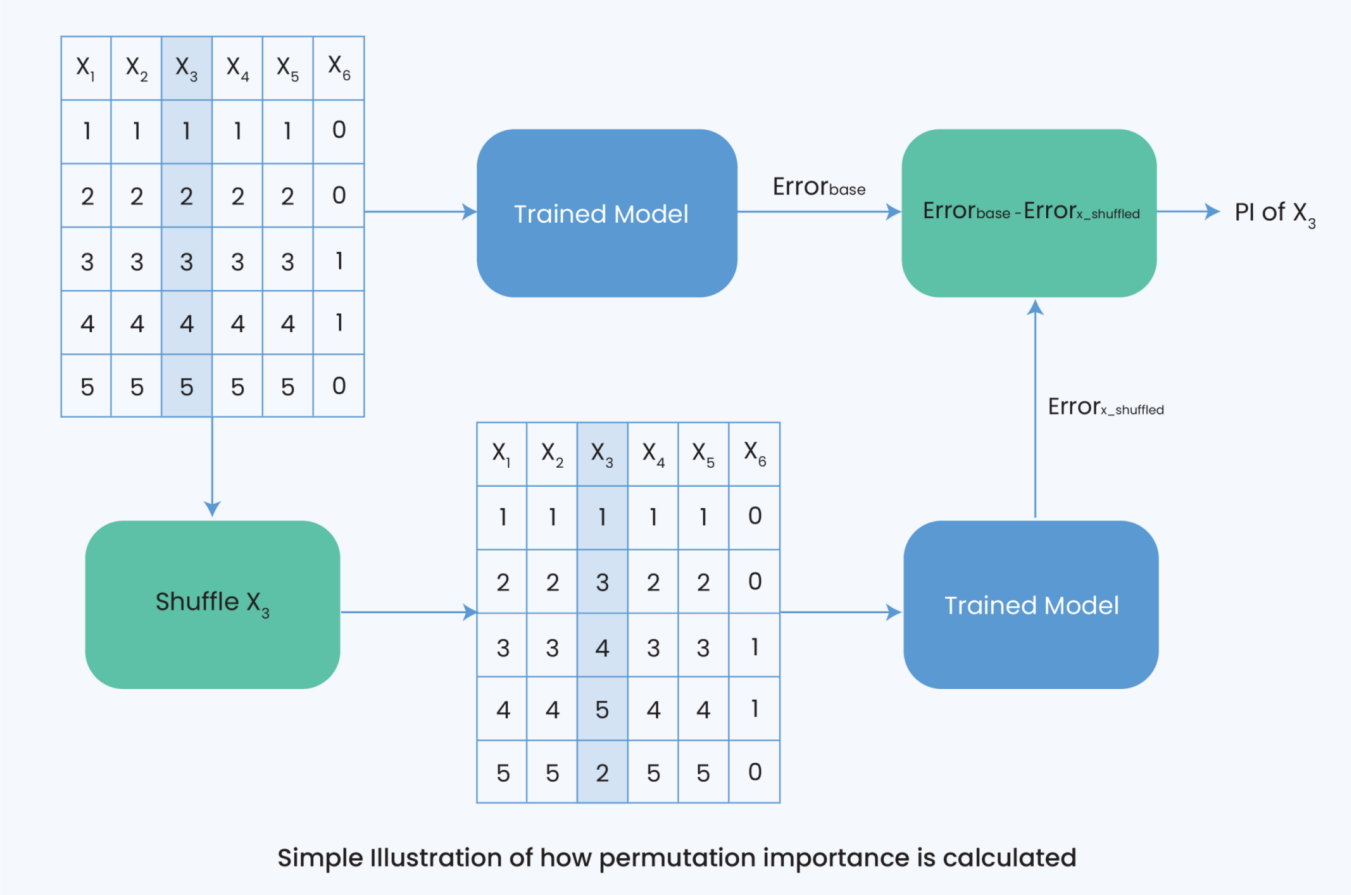


Figure 10: Higher permutation importance score ≈ more significant predictor

Limit to Permutation Methods

- Global, not local explanation
- Magnitude (relative importance), but no direction
- Assumes independence so correlated features biased
- Shuffling can bias estimates by creating values not seen in the training data (Hooker, Mentch, and Zhou 2021)

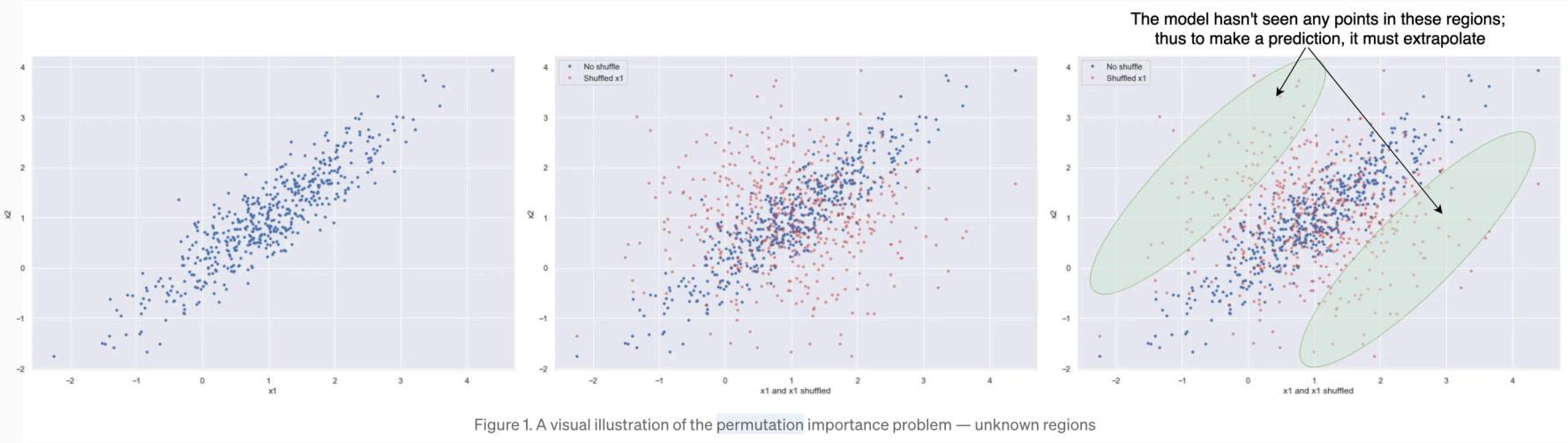


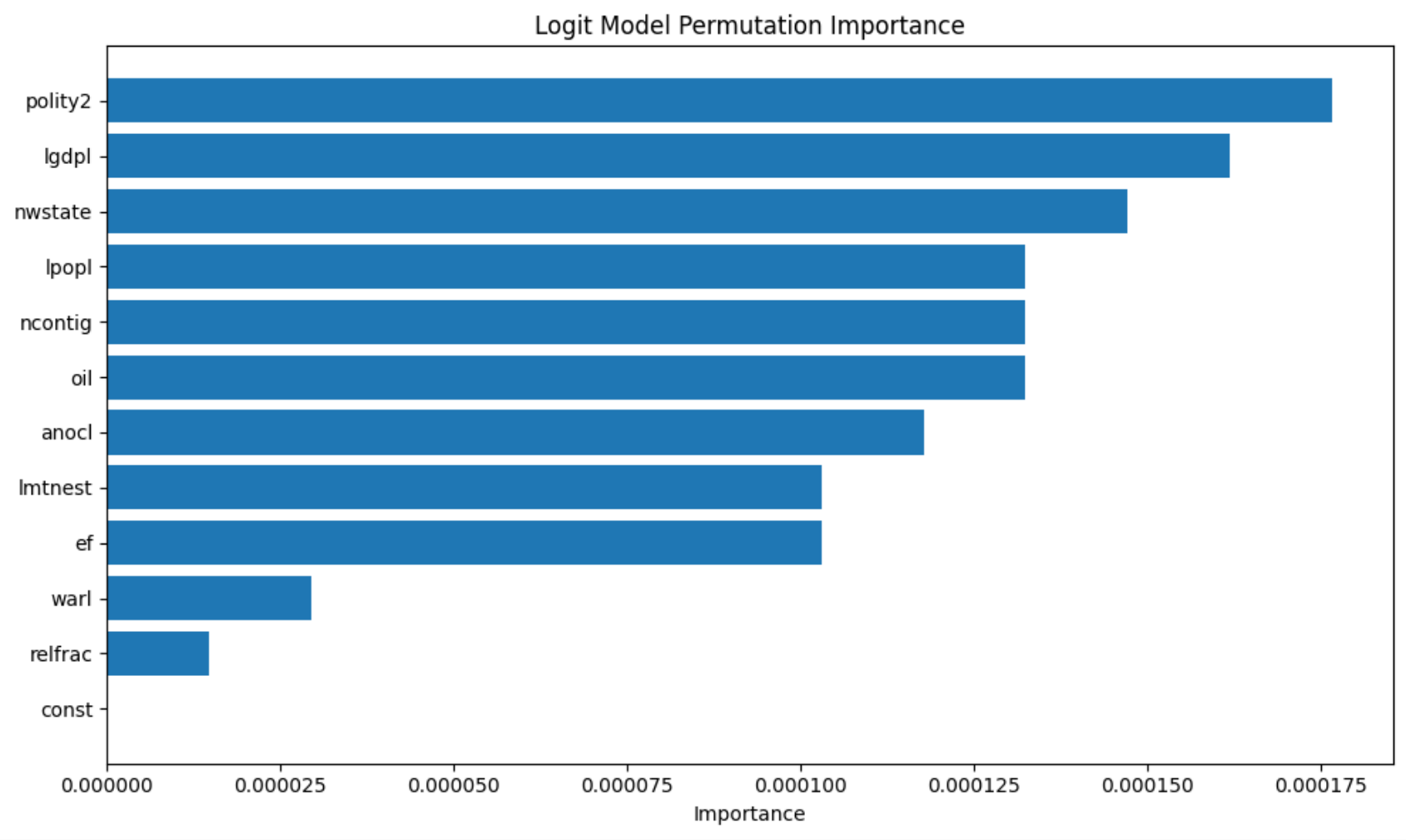
Figure 1. A visual illustration of the permutation importance problem — unknown regions

Figure 11: Source: Denis Vorotyntsev

Alternative Methods

- **Conditional Variable Importance:** conditional shuffling of feature to constrain values
- **Dropped Variable Importance:** drop feature, retrain model, compare scores
- **Tree-based Variable Importance:** calculate importance based on change to Gini impurity given split

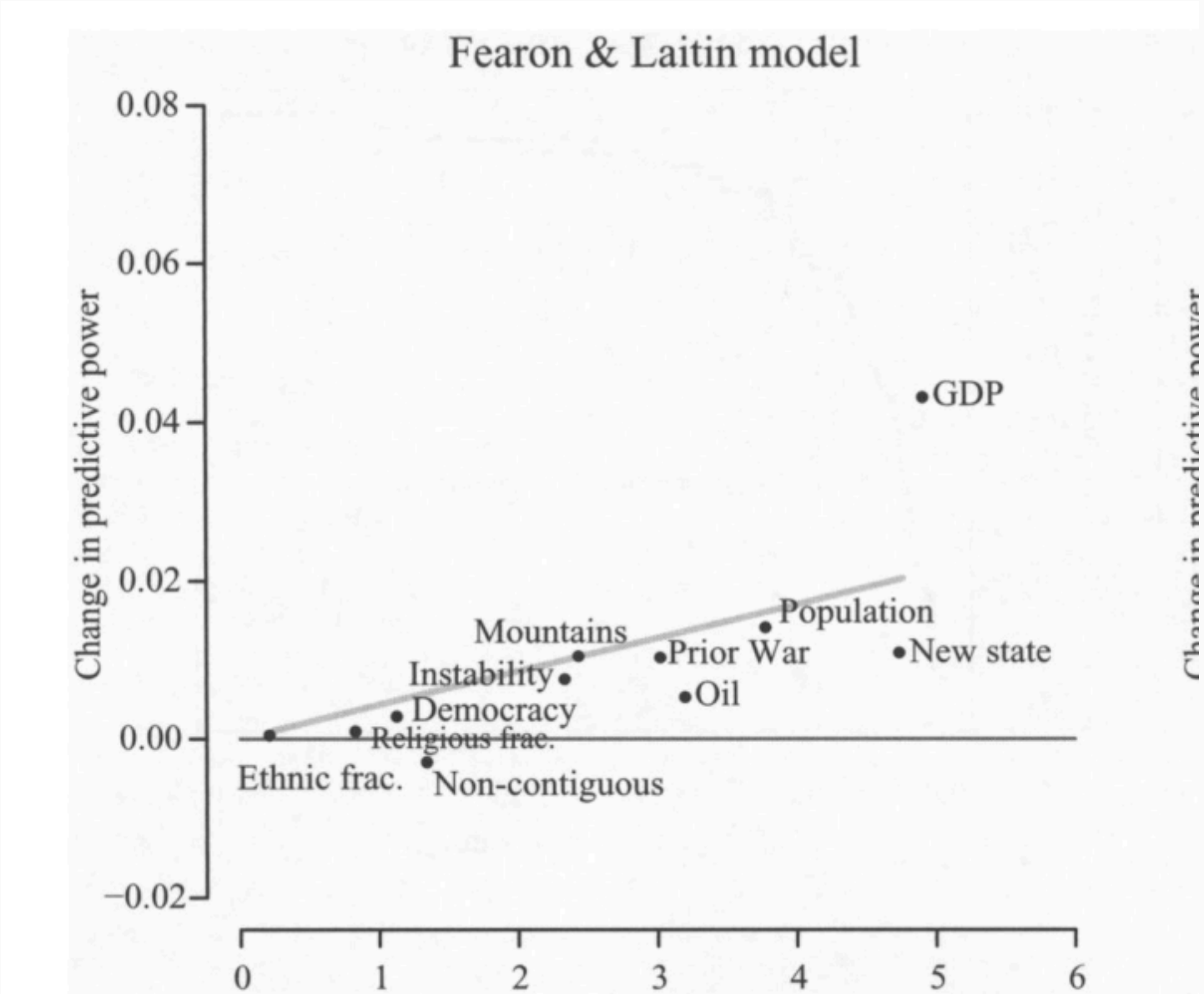
Fearon and Laitin (2003) permutation importance plot



X-axis is change in accuracy

Permutation importance vs statistical significance

Recall: Statistical significance does not mean predictive significance



SHAP (SHapley Additive exPlanations)

Goal: Quantify the contribution of each feature to a specific prediction, fairly and additively.

- Uses game theory concept of **Shapley values**
  - Treat each feature as a "player" in a cooperative game
  - Each feature contributes ϕ to the final prediction.
  - Calculates contribution as the **average marginal effect** across all possible feature combinations (coalitions).
- Assumptions
  - Fairness: $\hat{y}$ explained by all features in the model
  - Additive: $\hat{y}_i$ explained by base value and individual values
- Output: Contribution scores for feature i at a specific (local) observation.

Calculating Feature Contribution

Recall: We can calculate feature contribution ϕ in linear models for individual feature i based on $\phi_i = X_i\beta$.

Main Idea: Even if we don't have a β, SHAP can use similar intuition to calculate the actual contribution minus average effect under different configurations.

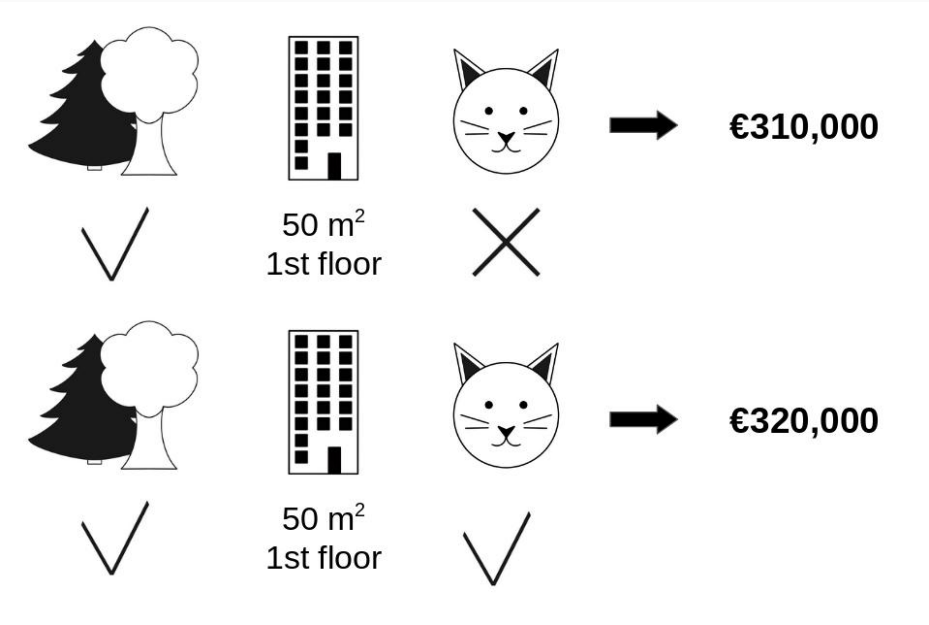


Figure 12: Housing price as function of park nearby, square footage, and cats allowed. What is ϕ of cat? (Source: Molnar)

SHAP Predictions based on feature decomposition

Math: $\hat{f}$ = baseline (average) value + $\sum$(SHAP values)

$$\hat{f}(\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

$$\hat{f}(\mathbf{x}) - \mathbb{E}[\hat{f}(\mathbf{x})] = \sum_{i=1}^{M} \phi_i$$

- $\hat{f}(\mathbf{x})$: prediction for instance $\mathbf{x}$
- $\phi_0 = \mathbb{E}_{\mathbf{x}}[\hat{f}(\mathbf{x})]$: baseline value (pooled, benchmark, etc)
- ϕ_i: contribution of feature i to the deviation from the baseline (the SHAP value)

SHAP Visualizations

- **Beeswarm Plot**: Global summary of feature importance and direction; each dot is a SHAP value for one instance-feature pair, color-coded by feature value.

SHAP Visualizations

- **Beeswarm Plot**: Global summary of feature importance and direction; each dot is a SHAP value for one instance-feature pair, color-coded by feature value.
- **Waterfall Plot**: Local summary; Bar chart version of the force plot; shows additive contributions of each feature to a single prediction from the baseline.

SHAP Visualizations

- **Beeswarm Plot**: Global summary of feature importance and direction; each dot is a SHAP value for one instance-feature pair, color-coded by feature value.
- **Waterfall Plot**: Local summary; Bar chart version of the force plot; shows additive contributions of each feature to a single prediction from the baseline.
- **Force Plot**: Local summary (tug-of-war); condensed waterfall plot; Shows how different feature effects pushes the prediction from the baseline.

Beeswarm Plot

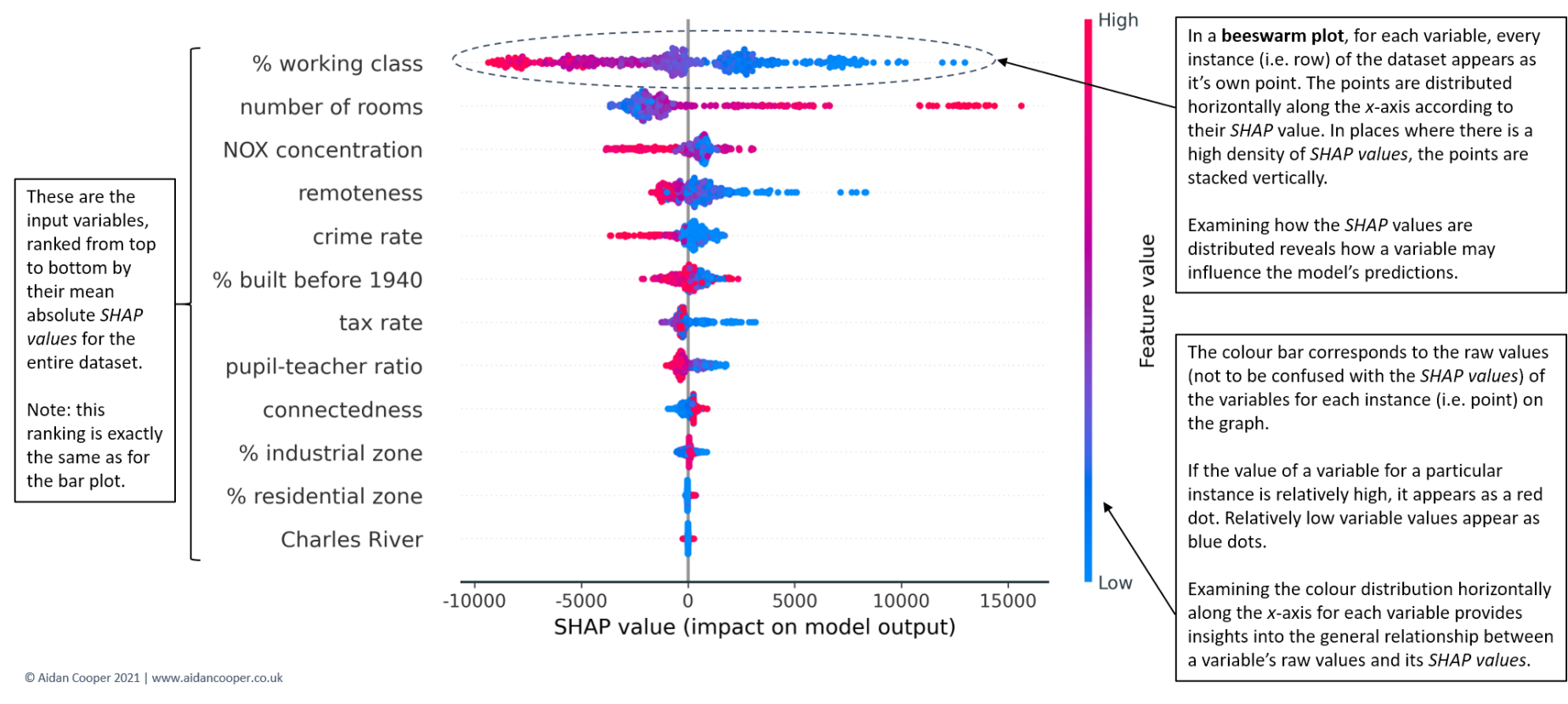


Figure 13: Source: Aidan Cooper

Waterfall Plot

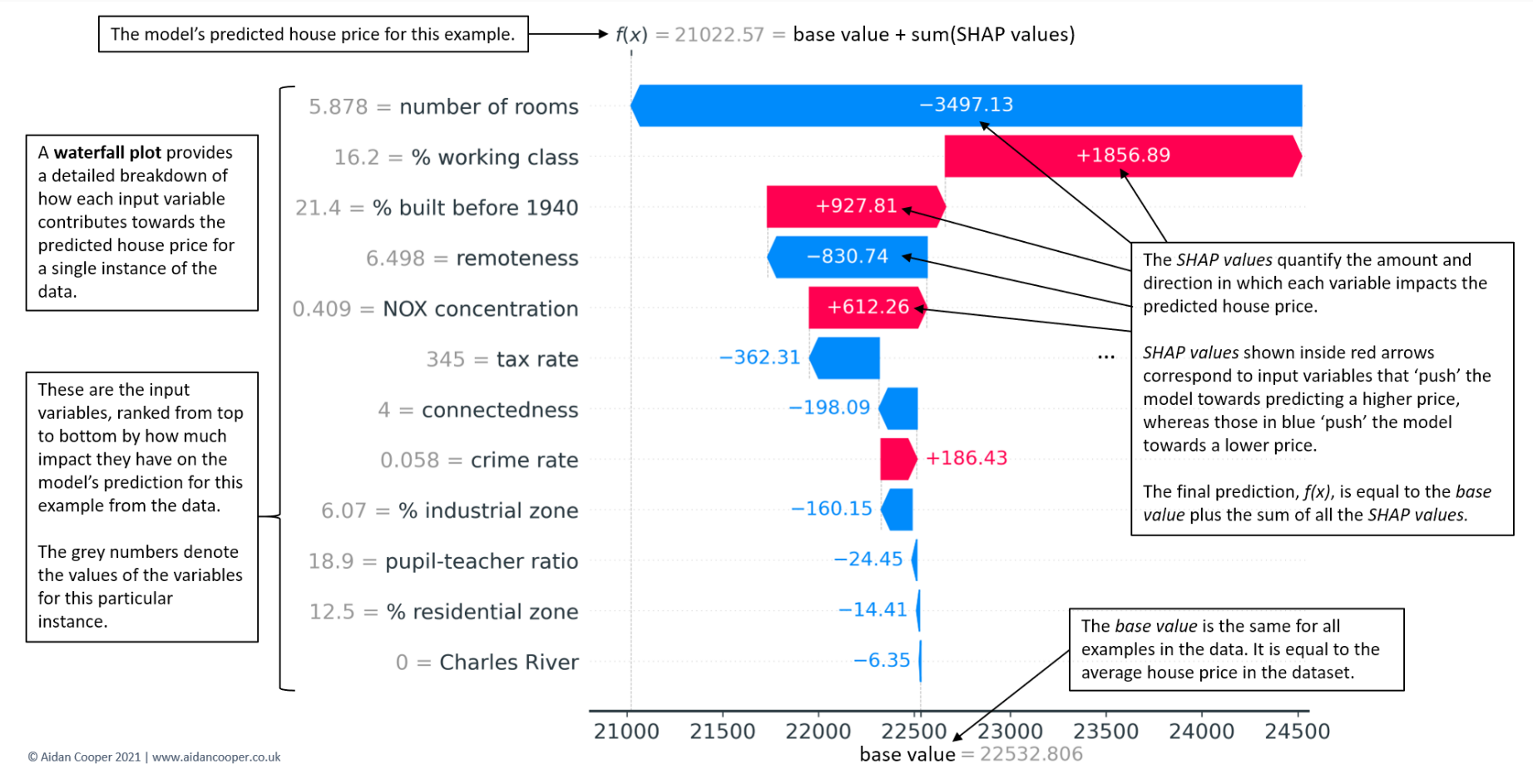


Figure 14: Source: Aidan Cooper

Force Plot

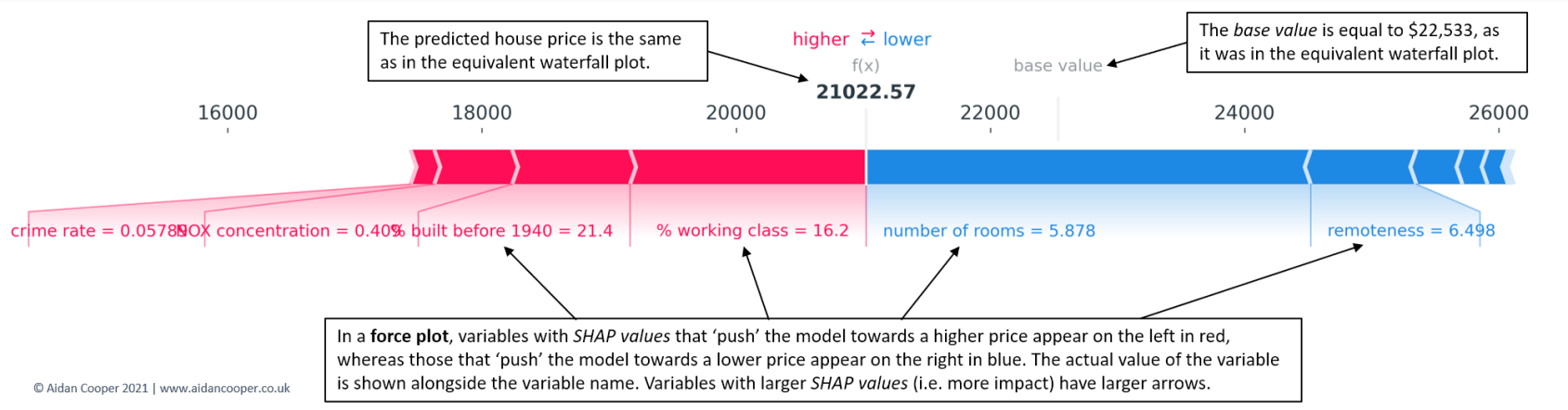
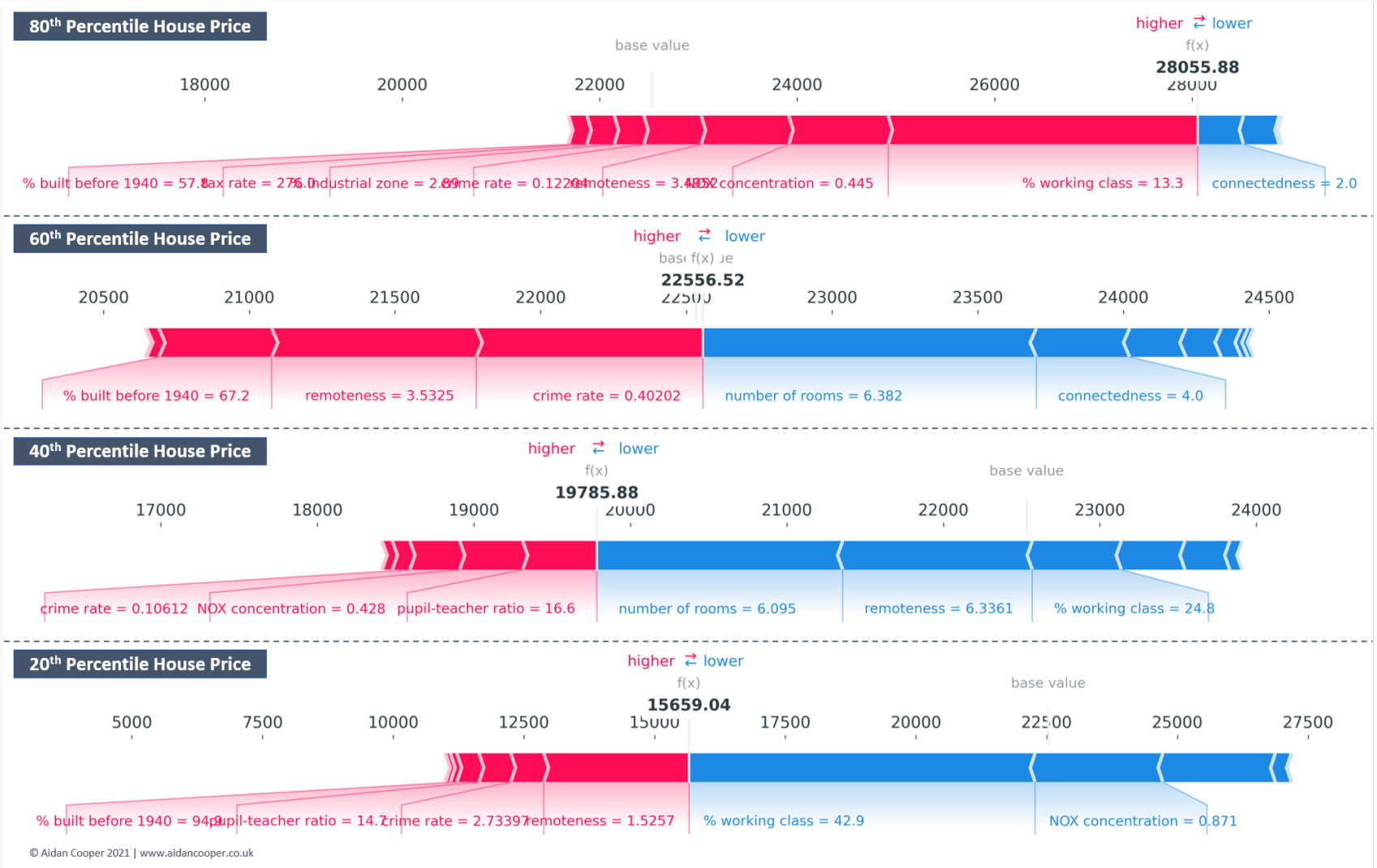


Figure 15: Source: Aidan Cooper

Force Plot

Compare multiple instances simultaneously

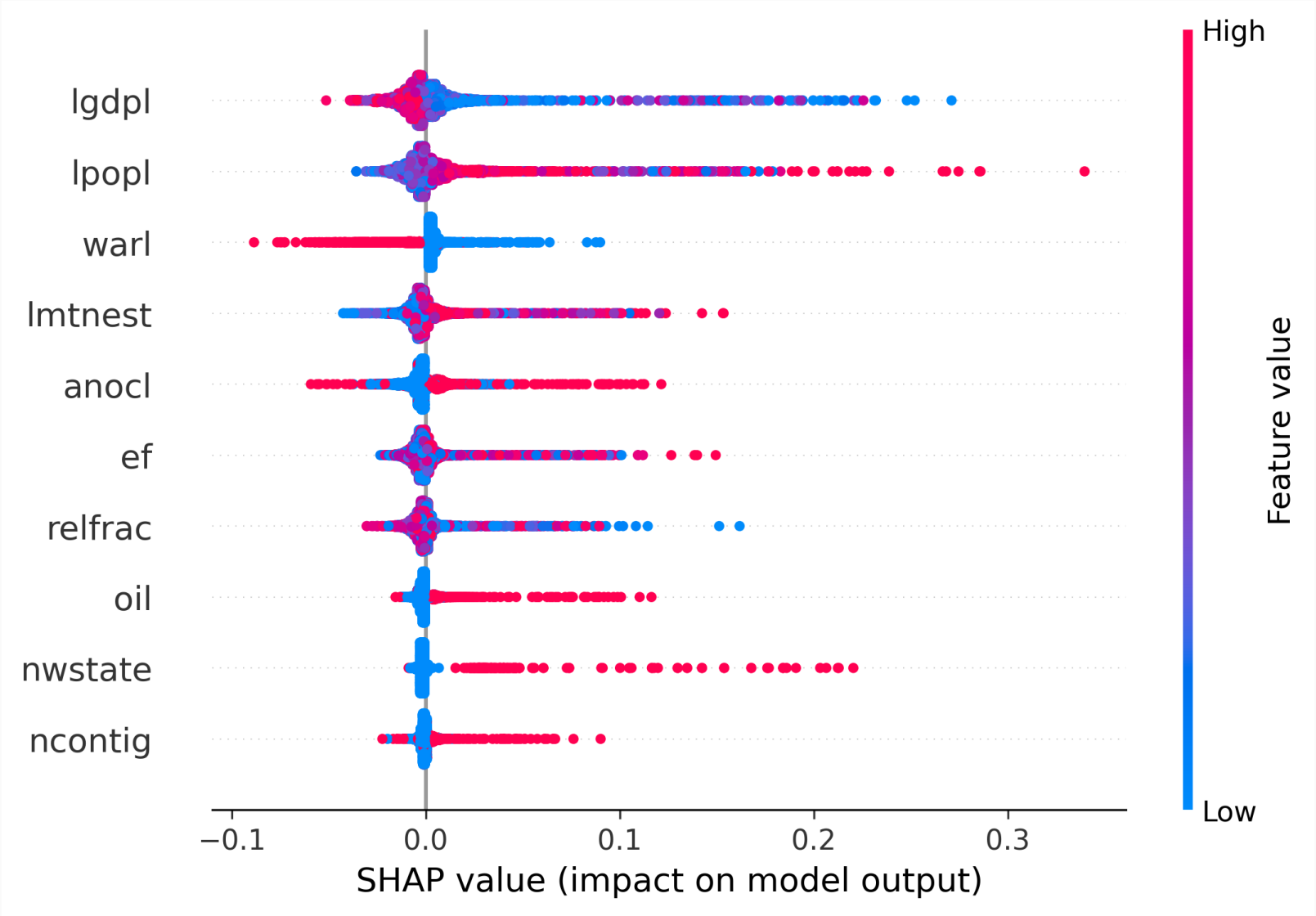


Fearon and Laitin (2003) Example

- What predicts civil war?
- What predicted a particular civil war?
- Base Value:
  - Global (pooled)
  - Within-country
  - Individual country

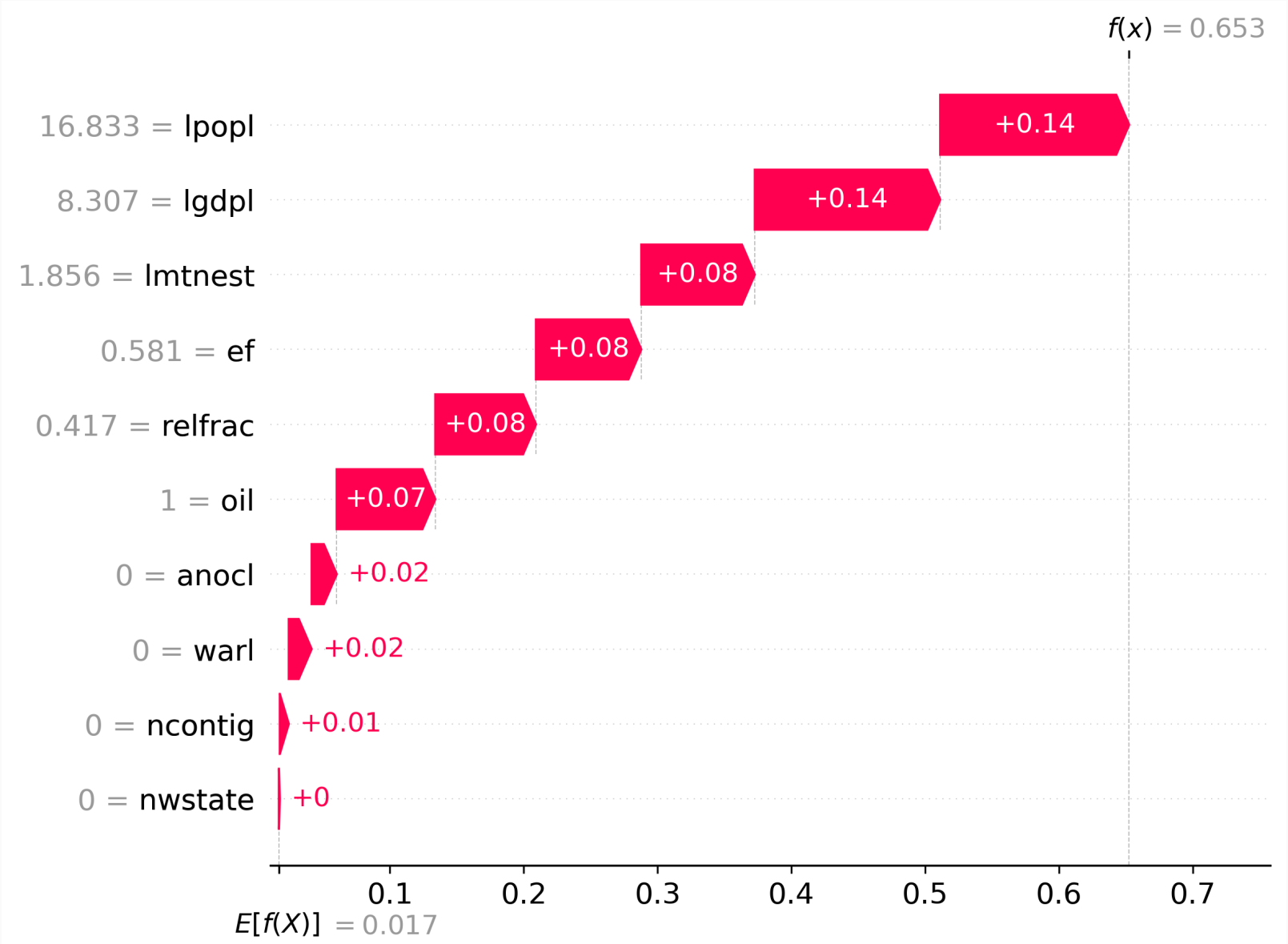
Beeswarm Plot Example

Example: Random Forest Model of Fearon and Laitin (2003)



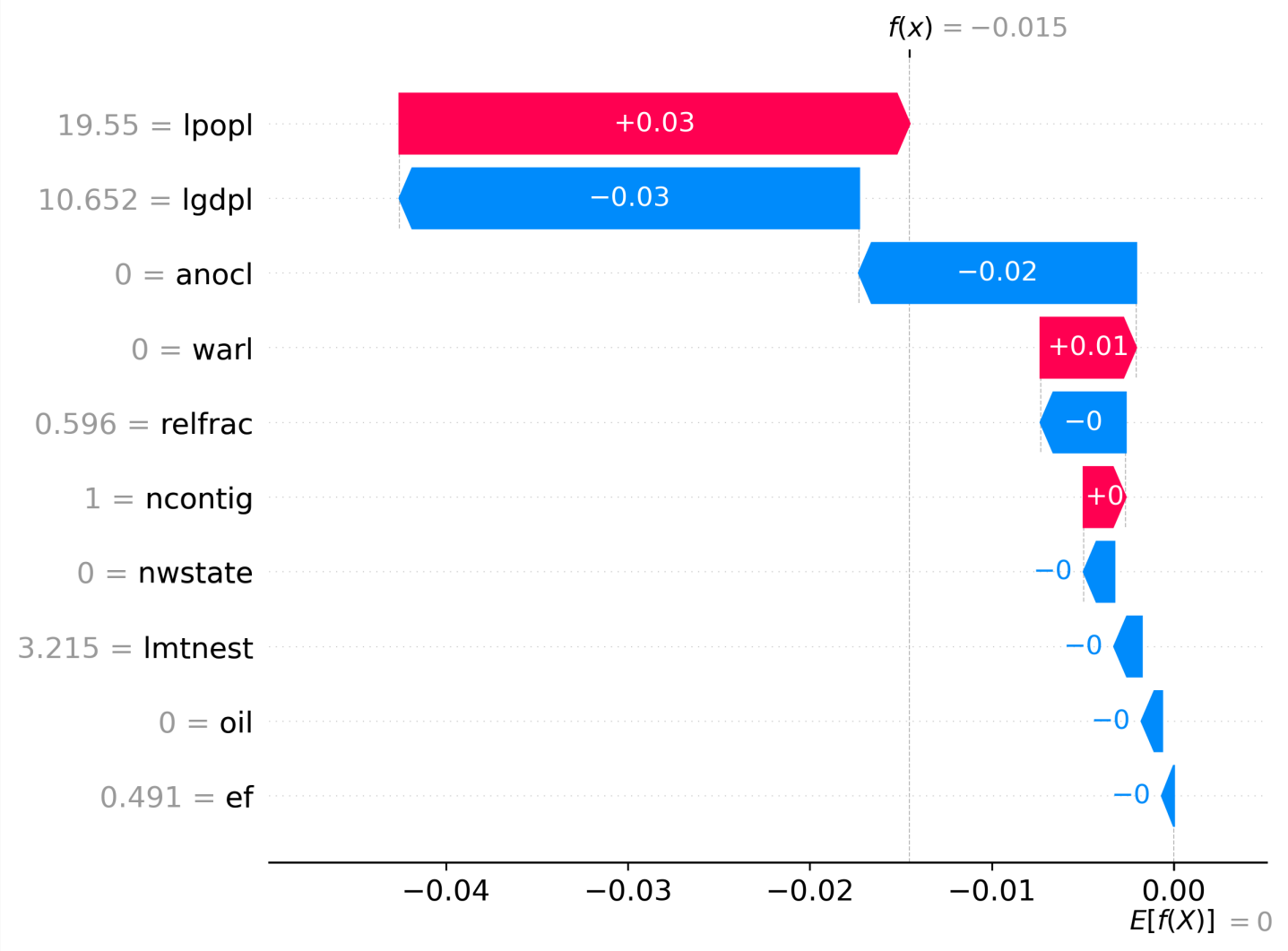
Waterfall Plot Example

What predicted Syria civil war in 2011?



Waterfall Plot Example

What was U.S. civil war risk in 2011?



Explainability: Why did the model predict y_i?

Method	Goal	Strengths	Limitations
Partial Dependency Plot (PDP)	Estimate average marginal effect of a feature	Simple, visual, model-agnostic	Assumes feature independence; sensitive to correlated features
Accumulated Local Effects (ALE)	Estimate feature effect without assuming independence	Handles correlated features; faster than PDP	Harder to interpret; less well-known; assumes smooth local effects
Statistical ML (Multiple Testing, Kullback-Leibler, Meta-Learners)	Hypothesis testing, inference, meta-learners	Marries inference and ML worlds; controls false discovery rate (FDR)	May over-report false positives; standard inference problems

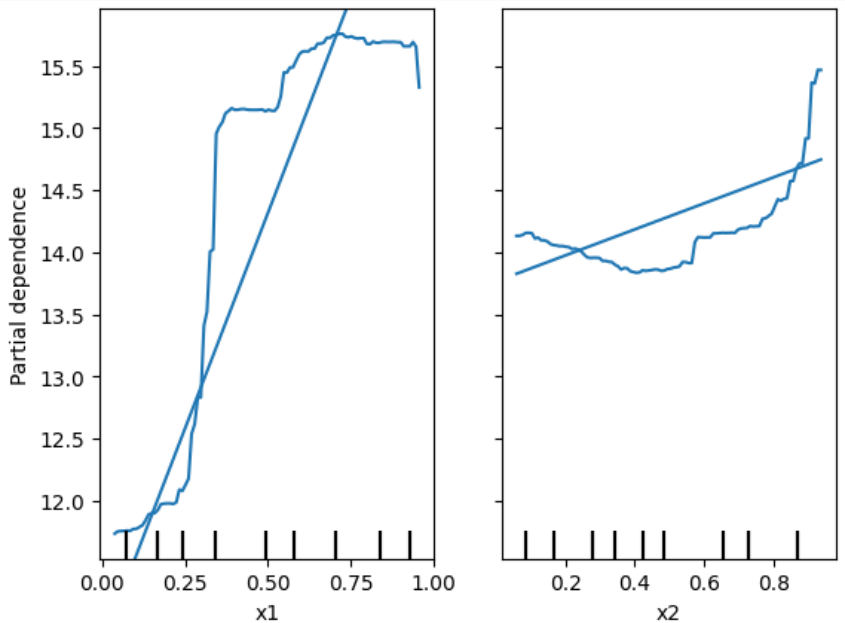
Partial Dependence Plot (PDP)

Goal: Understand the marginal effect of a feature on a model's prediction.

Procedure:

- Choose the feature(s) of interest (e.g., "GDP").
- For each unique value or range of the chosen feature:
  - Fix the value of the feature.
  - Compute the model predictions (e.g., 'civil war') while varying other features in the data.
  - Take the average of these predictions for each feature's fixed value.
- Plot the averaged predictions against the fixed feature values to get range of values

Partial Dependence Plot (PDP)



Interpretation: X-axis tells us values of feature and y-axis is prediction given X-value.

- x1 is positively impacting the model's performance
- x2 is initially pushing down then pushing up the prediction (non-linear)

Fearon and Laitin Partial Dependence Plot (PDP)

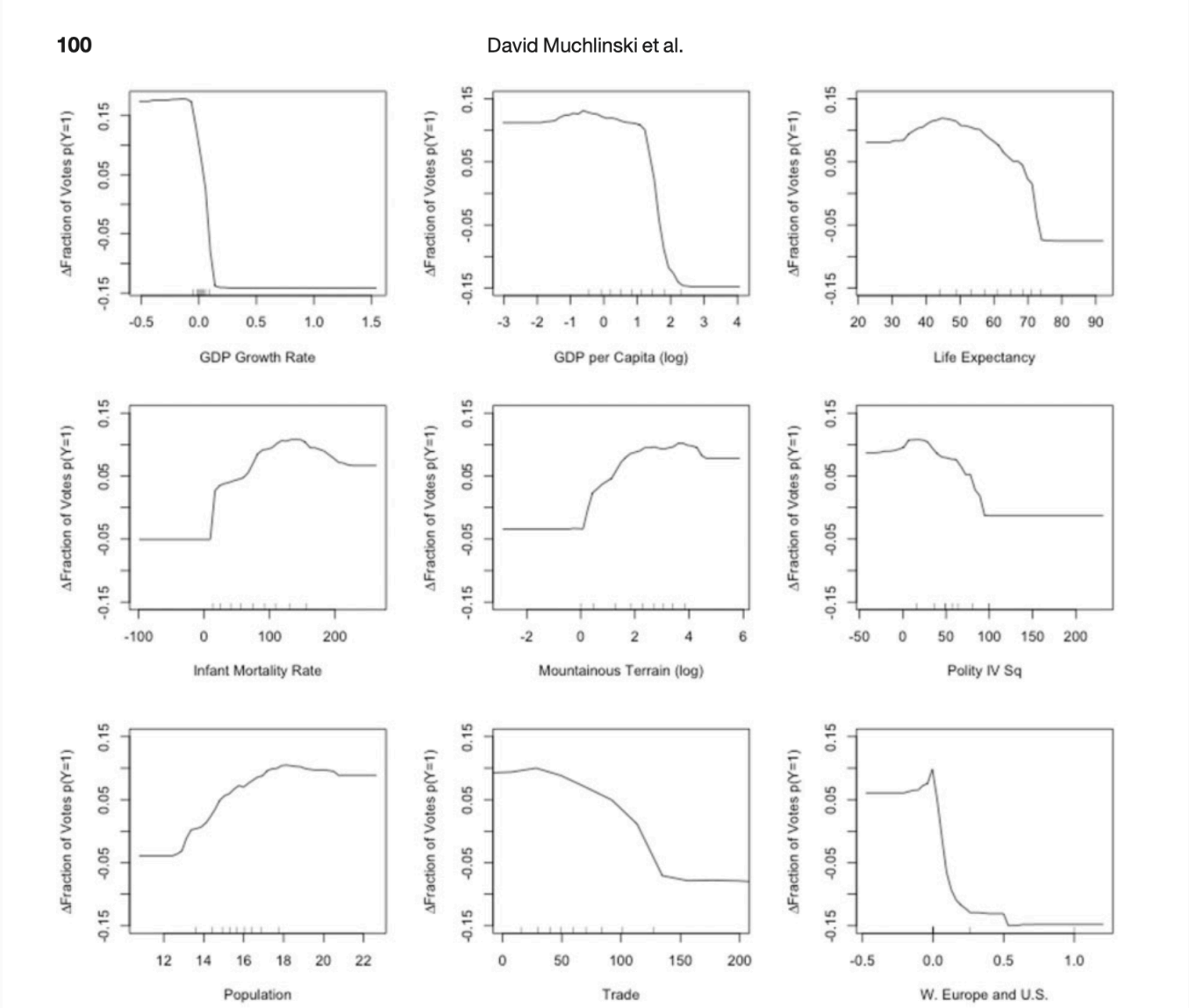
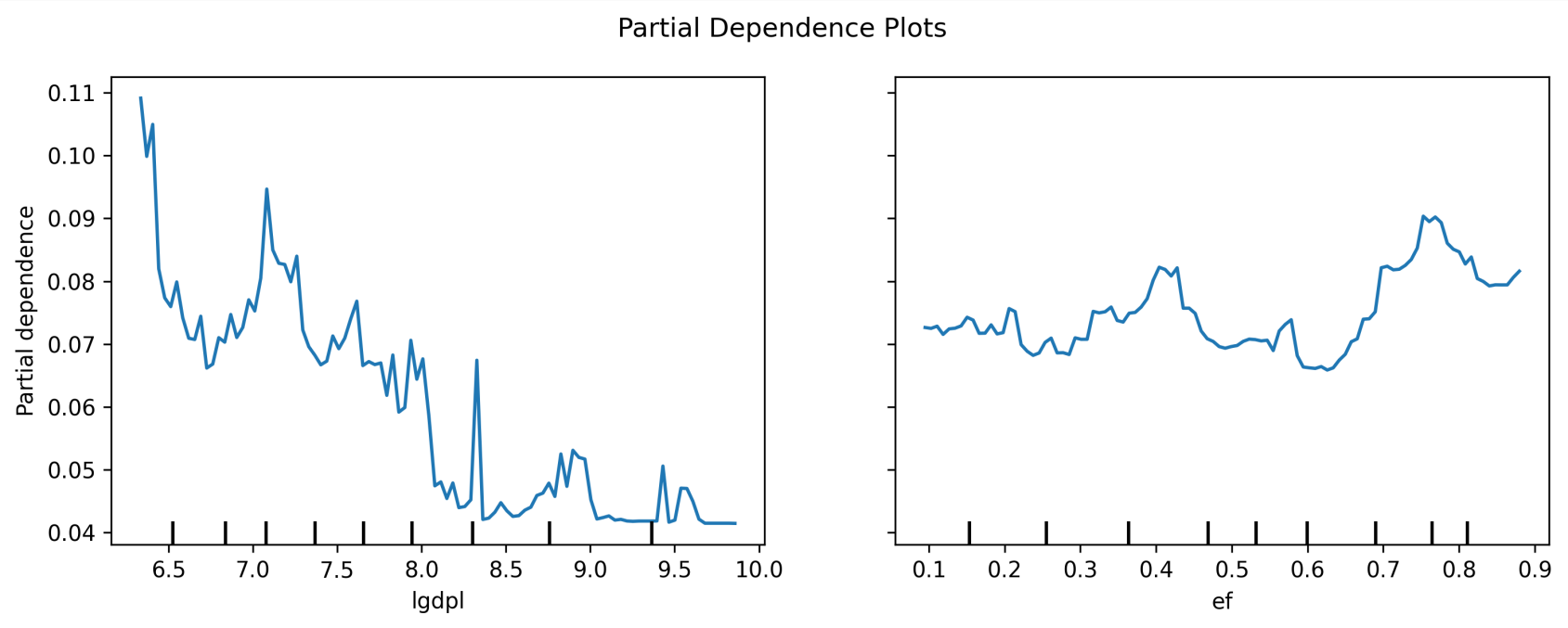


Fig. 5 Partial dependence plots.

Figure 18: Source: Muchlinski et al

Fearon and Laitin Partial Dependence Plot (PDP)



Accumulated Local Effects

Goal: Understand how a feature affects predictions controlling for relationships with other variables.

Main Idea: It averages the **local differences** in predictions over small intervals (bins) of the feature. The resulting plot is a sum of these local effects.

Benefits:

- Features may be correlated or interactive (limit to PDP)
- Conditional: ALE plots show the average change in prediction when a feature changes slightly, while **conditioning on the data distribution**.
- Bounded: Avoids unrealistic combinations of feature values (limit to permutation-methods)
- **Interpretation:** Conditional on feature 1 value, the relative effect of changing feature 2 on the prediction is fill (sum) value.

Debugging

Debugging is key to detect ML Shortcuts

- **Problem:** “Clever Hans”

Debugging is key to detect ML Shortcuts

- **Problem:** “Clever Hans”
  - Recall: ‘Greedy” algorithms take ML shortcuts
  - Another shortcut can occur when ML focuses on non-meaningful patterns (the noise) rather than the meaningful patterns (the signals)
  - Overly flexible model backfires to create incoherent pattern

Debugging is key to detect ML Shortcuts

- **Problem:** “Clever Hans”
  - Recall: ‘Greedy” algorithms take ML shortcuts
  - Another shortcut can occur when ML focuses on non-meaningful patterns (the noise) rather than the meaningful patterns (the signals)
  - Overly flexible model backfires to create incoherent pattern
- Consequence:
  - Incoherent or idiosyncratic patterns in data.
  - Undermines trust in machines
  - Unclear generalizability

ML shortcuts are more likely as models become more complex

Examples: Wolf vs husky classifier (white vs green background), cats vs dogs (pink vs blue leashes), U.S. election model (October surprise)

(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Figure 19: Source: Ribeiro, Singh, and Guestrin (2016)

Debugging Methods to Catch Shortcuts

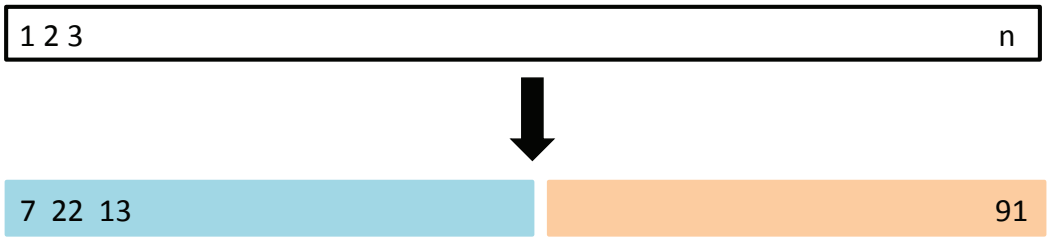
- **Adversarial examples:** If we try to deceive the model (bad data), how does it respond?
- **Data leakage monitoring:** How stable are results across different folds? How does model do once deployed?
- **Ceteris paribus:** How sensitive is the model to current DGP? (similar to PDP)
- **Cross-validation error:** Are results consistent across different iterations?

Cross-Validation (CV) error analysis

Goal: Identify most robust predictors based on cross-validation error; make inferences about 'important' contributors based on sampling distributions.
Recall: Cross-validation aims to estimate the validation (test) error for a supervised learning method

Strategy:

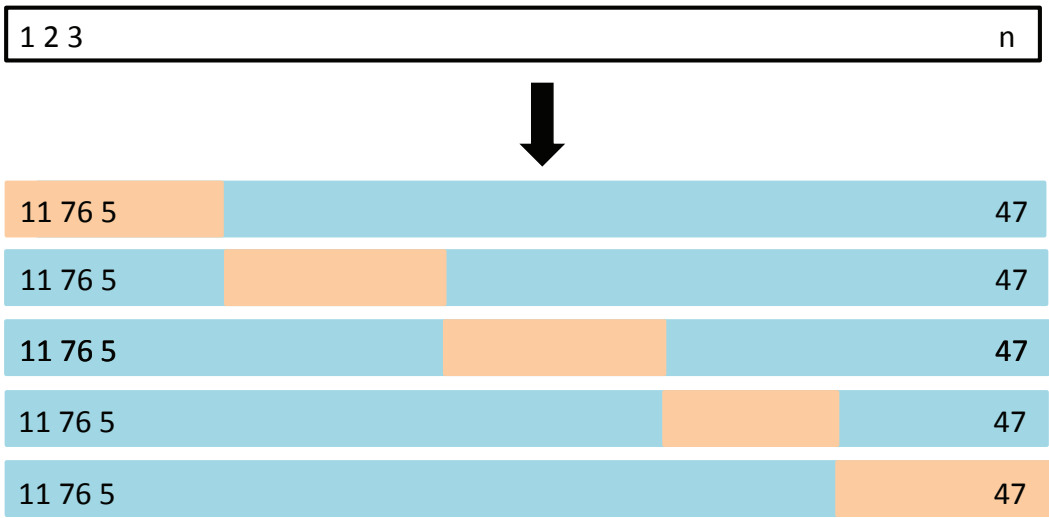
- Split the data in two parts
- Train the model on the first part (training set)
- Compute the error on the second part (validation set)



k-fold Cross-Validation

General k-fold CV Procedure

- Split the data into k subsets or folds
- For every $i = 1, 2, \ldots, k$:
  - Train the model on every fold except the i^{th} fold
  - Compute the test error on the i^{th} fold
- Average the test errors



CV Error Analysis

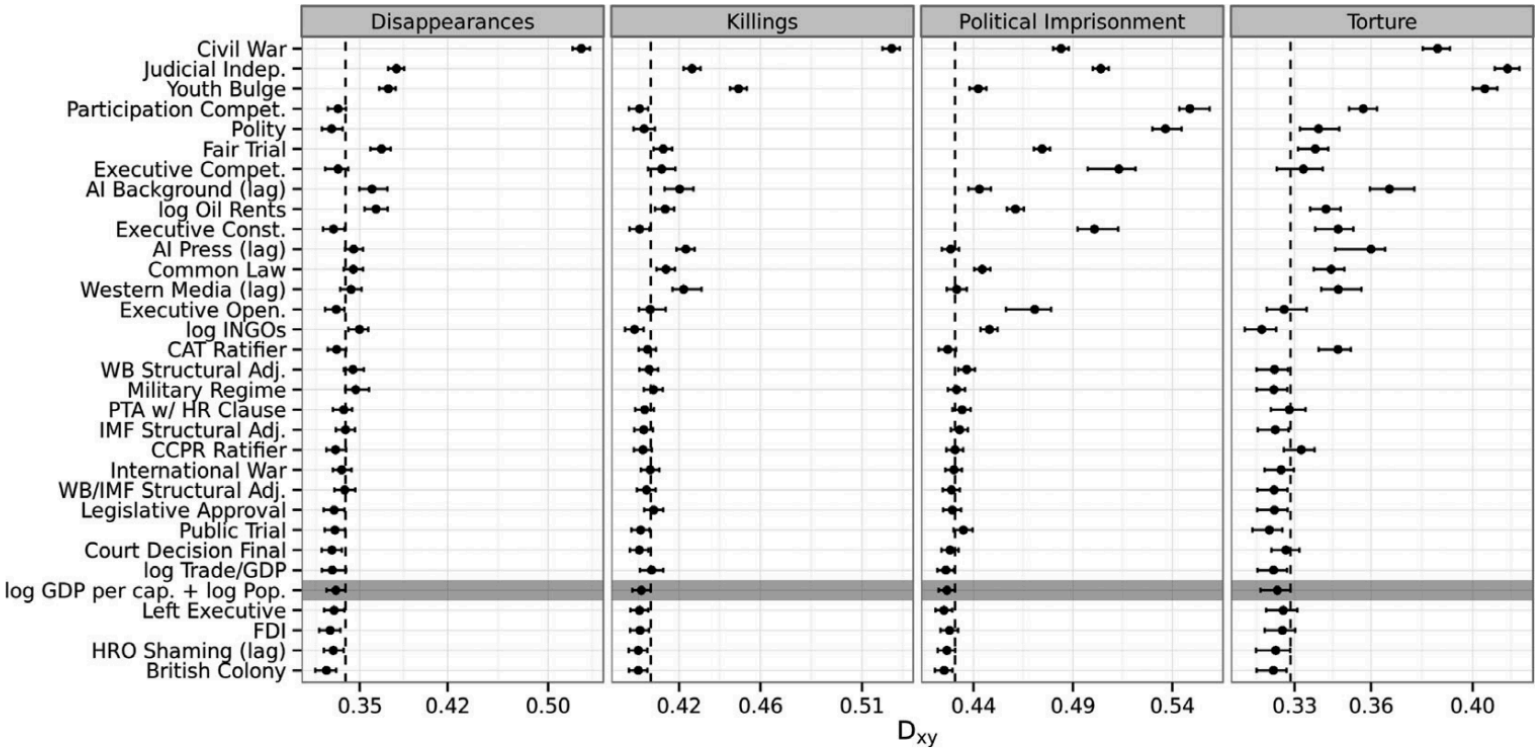
Procedure:

1. Create baseline regression model
2. Iteratively add one candidate variable at a time to model
3. For each iteration, apply 10-fold cross-validation.
4. Estimate mean prediction error across iterations where distribution of prediction errors for each feature based on out-of-bag error
5. **Inference Rule:** "...to determine whether a covariate is an important predictor of state repression: if the lower bound (the .025 quantile) of the prediction error for the model including that covariate is above the upper bound (the .975 quantile) of the prediction error for the baseline mod el, then the covariate is marginally important."

Example: Hill and Jones (2014)

Hill and Jones (2014) debug correlates of state repression (CIRI Physical Integrity Score) using 10-fold CV

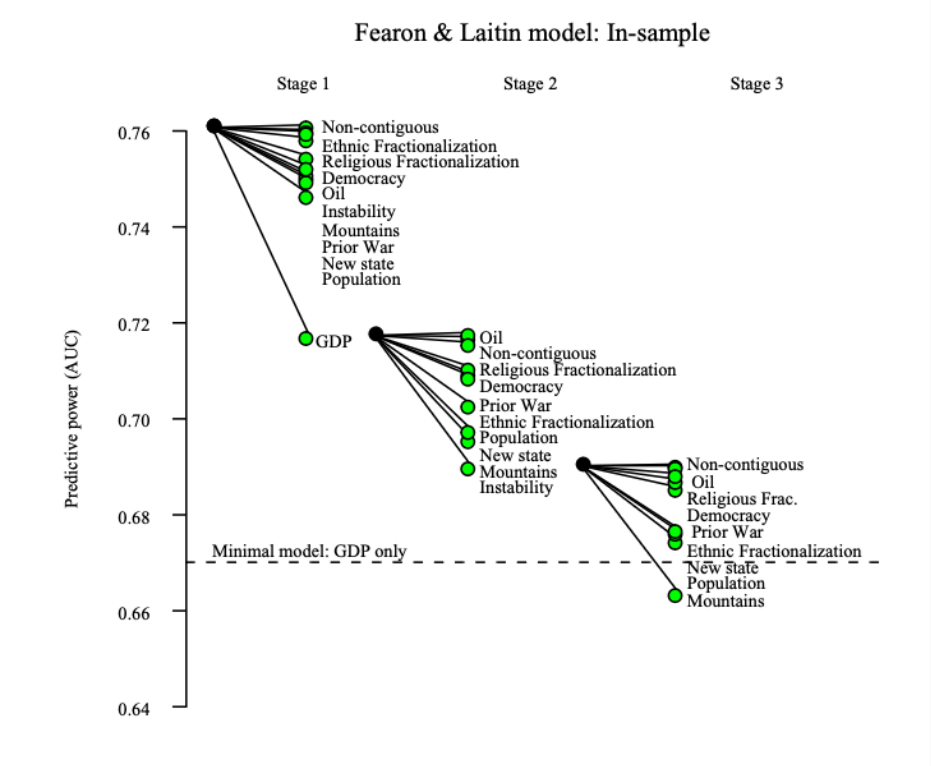
FIGURE 3. Results from Cross-validation of Error from Ordinal Logistic Regression Models of State Repression using (the natural logs of) GDP per Capita and Population as a Base Specification



Notes: The *x* axis shows Somer's D_{xy}, a rank correlation coefficient that ranges from −1 to 1. The *y* axis represents model specifications which are composed of a base model, which is indicated by the gray band, and the variable indicated on the *y* axis. The dots show the median of the sampling distribution of the Somer's D_{xy} statistic, along with the .025 and .975 quantiles. The dotted line shows the .975 quantile of the sampling distribution of the D_{xy} statistic for the base model. Model specifications whose intervals overlap with this line do not add significantly to the fit of the model compared to the base specification.

Example: CV Analysis of Fearon and Laitin (2003) AUC Changes

4-fold CV: IS GDP-conflict link robust?



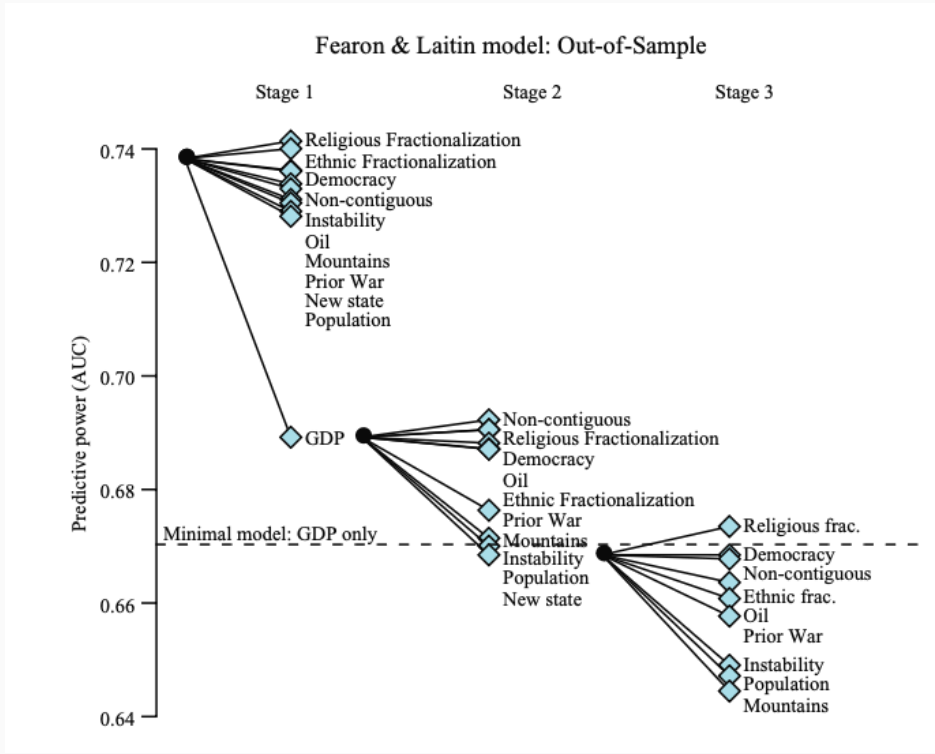
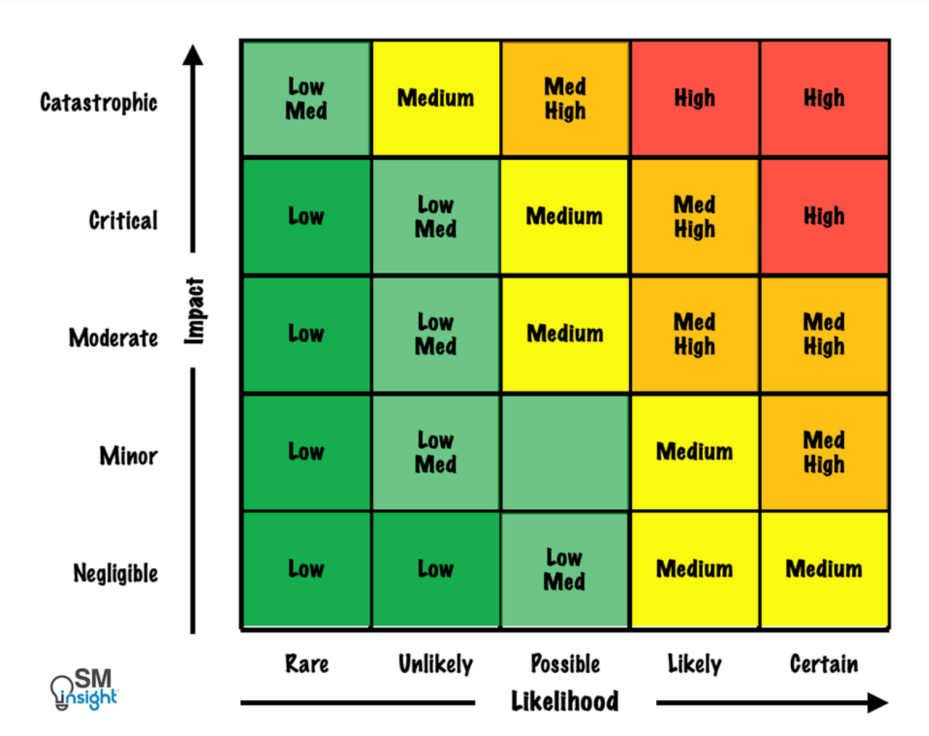


Figure 21: Source: Ward, Greenhill, and Bakke (2010)

Transferability: What if?

Motivation:

- If history doesn't repeat itself, then we need:
  - Risk management
  - Contingency planning
- Stress-test requirements, ex. 2025 European Banking Authority



Political risk management and contingency planning tools

1. **Scenario Analysis:** Model domestic and geopolitical tail risks (e.g., Taiwan Strait crisis) using forward-looking assumptions.

Political risk management and contingency planning tools

1. **Scenario Analysis:** Model domestic and geopolitical tail risks (e.g., Taiwan Strait crisis) using forward-looking assumptions.
2. **Synthetic Data:** Generate artificial datasets to simulate rare or unseen edge cases (e.g., regime collapse, realistic disaster scenarios)

Political risk management and contingency planning tools

3. **Simulations**:
    - **Stress-Tests:** Evaluate model outputs under extreme or adverse conditions

Political risk management and contingency planning tools

3. **Simulations**:
    - **Stress-Tests:** Evaluate model outputs under extreme or adverse conditions
    - **Monte Carlo methods:** Randomly sample from input distributions to:
        - Estimate value at risk (VaR)
        - Quantify prediction uncertainty
        - Explore outcome distributions under uncertainty and alternative assumptions

Simulations

Question: “If X changes, how does the predicted outcome change?”
Answer:

- Point estimate: What is the new $\hat{y}$ given X change?
- Likelihood: How probable is this new estimate?
- Inferential uncertainty: How much confidence should I have in this estimate?

Method: Simulations use the model’s initial estimates to show how changes in variables influence predicted outcomes.

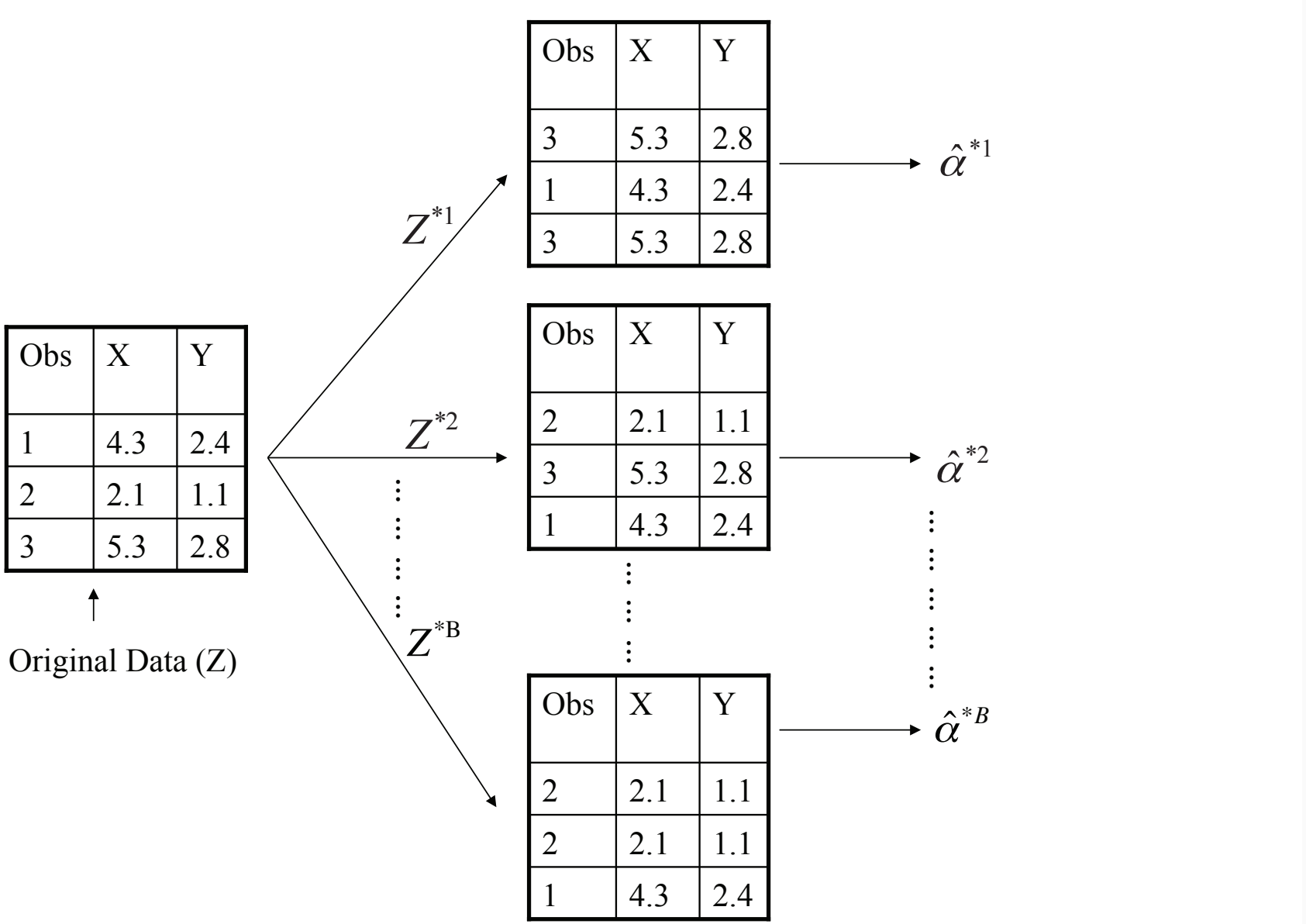
Simulations rely on bootstrap methods

Recall: The bootstrap estimates the variability around a parameter by repeatedly sampling observations from the dataset

Procedure:

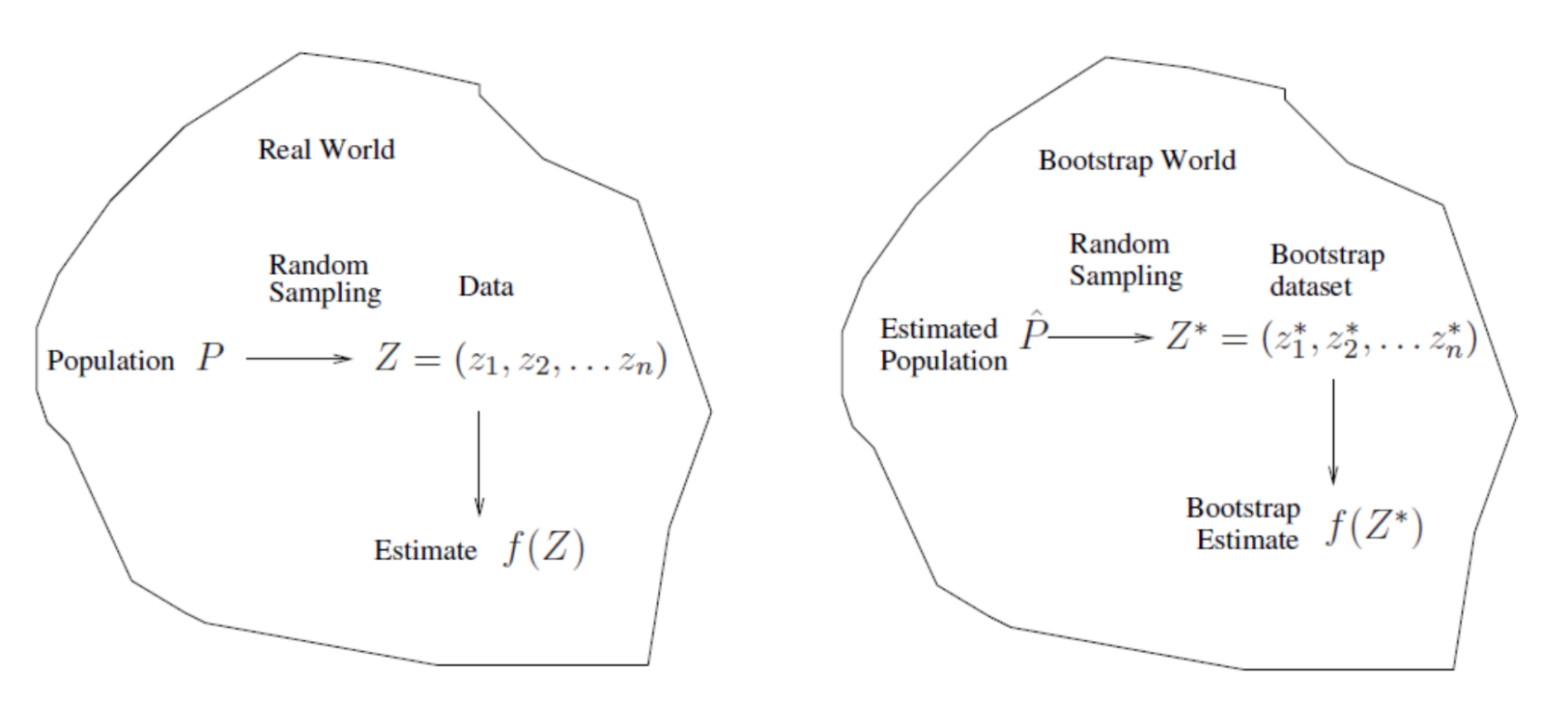
1. Resample the data by drawing B samples with replacement from training set to create new dataset Z^*
2. Estimate model on each sample of the data to get coefficient estimates $\hat{\alpha}$
3. Use variability in coefficient estimates to estimate standard error $\hat{\sigma}_\alpha$

Bootstrap Schematic



Comparison

Asymptotics mean bootstrapping recovers approximate population estimates



Simulations leverage bootstrap to measure uncertainty

Set-Up:

Coefficients are assumed to follow a multivariate normal distribution:

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{c}} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$$

- $\hat{\boldsymbol{\beta}}$ is the vector of estimated coefficients
- $\hat{\boldsymbol{\Sigma}}$ is the estimated variance-covariance matrix

Procedure:

1. Pick a given feature of interest to vary, keeping all others constant
2. Bootstrap $\tilde{\alpha}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{c}}$ from the variance-covariance matrix for M simulations

Simulations leverage bootstrap to measure uncertainty

Procedure:

3. Re-estimate the new predicted outcome $\tilde{Y}_c$ given $\tilde{\boldsymbol{\beta}}_{\boldsymbol{c}}$ given the relevant link function, e.g. logit:

$$\text{logit}^{-1}(\alpha + \mathbf{X}\boldsymbol{\beta}) = \frac{1}{1 + \exp\left(-(\alpha + \mathbf{X}\boldsymbol{\beta})\right)}$$

   Bootstraps generate a new distribution of possible outcomes $\hat{\tilde{Y}}_c \sim \mathcal{N}(\mu, \sigma)$

4. Average over the fundamental uncertainty by calculating the mean of the m simulations:

$$\tilde{E}[Y_c] = \frac{1}{m} \sum_{k=1}^{m} Y_c$$

   This gives the simulated expected value of the outcome variable.

Example: Simulations of Fearon and Laitin (2003)

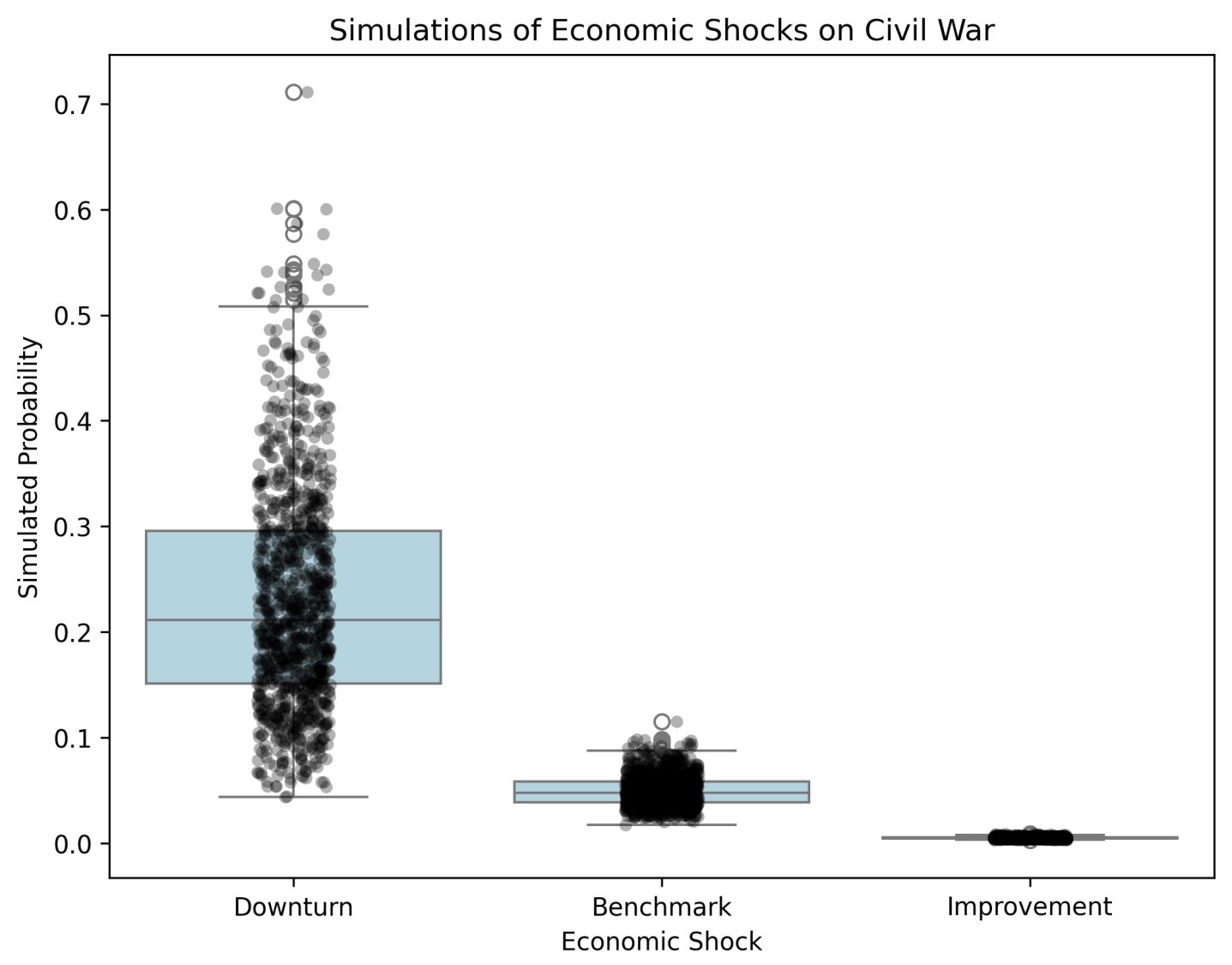
What if...

- Syria democratized in 2010?
- Sectarian violence worsens in 2025?
- U.S. faces an economic downturn?

Figure 23: Barbara Walter, “How Civil Wars Start”

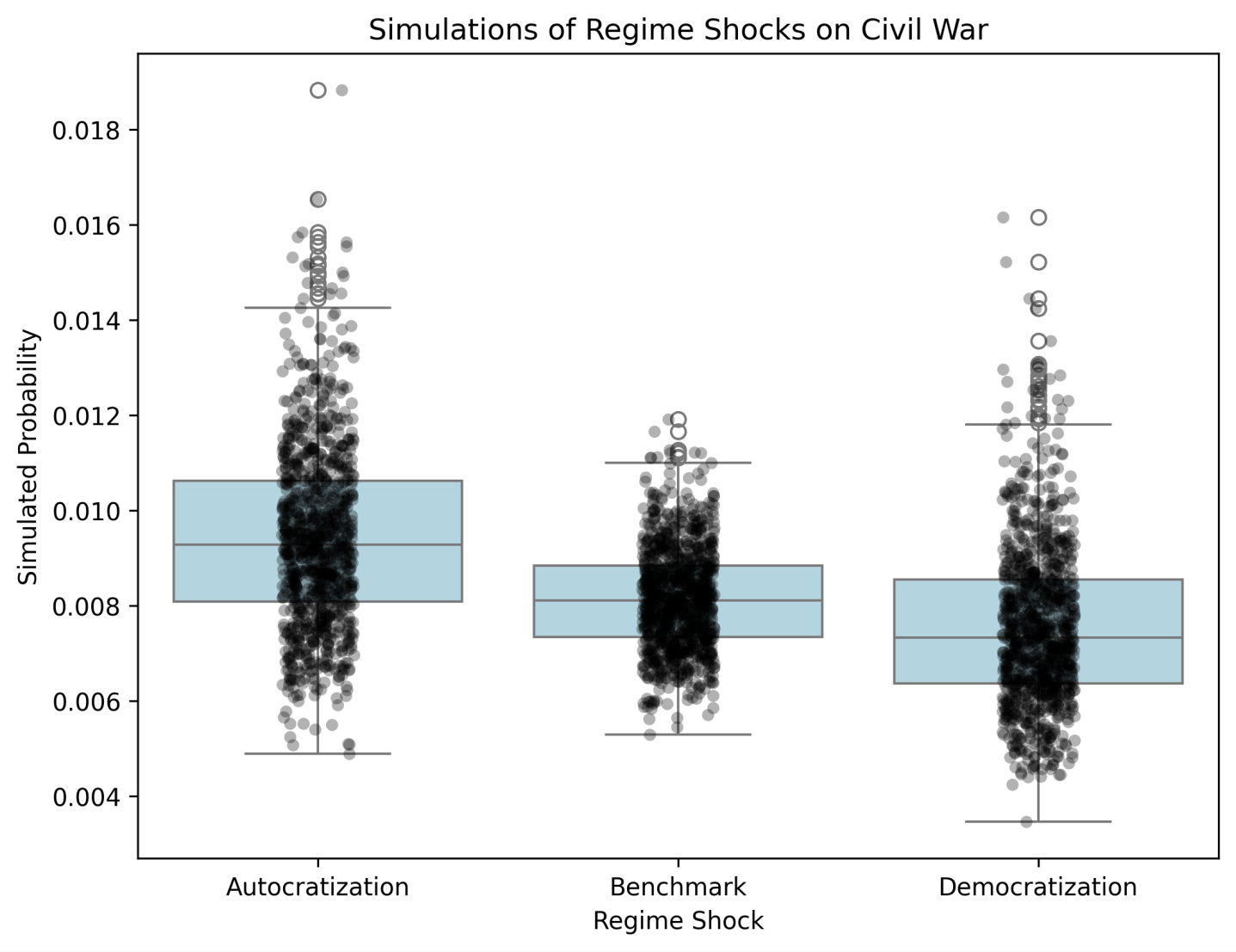
Example: Fearon and Laitin (2003)

Simulate: How would an economic shock affect the likelihood of civil war?



Example: Fearon and Laitin (2003)

Simulate: How would a regime shock affect the likelihood of civil war?



Conclusion

- Interpretable ML is key to increase trust, transparency, and explain why the model behaves the way it does
- Permutation importance and SHAP show why an output is changing
- Simulations show what-ifs and quantify uncertainty
- XAI requires technical know-how and domain expertise – your superpower

Course Wrap-Up

ML is powerful tool for prediction problems

Main Idea: ML is a form of artificial intelligence increasingly used in social science to solve complex prediction problems. It involves a set of computer algorithms which 'learn' patterns in existing data to assist in prediction and inference.

- Corporates
- Government
- Finance
- Academia

What We've Covered

1. Country and political risk analysis
2. Unsupervised Learning

3. Supervised Learning

What We've Covered

1. Country and political risk analysis
2. Unsupervised Learning
    - 2.1 Principal Component Analysis
    - 2.2 Clustering
    - 2.3 Topic Models (Text Analysis)
3. Supervised Learning

What We've Covered

1. Country and political risk analysis
2. Unsupervised Learning
    - 2.1 Principal Component Analysis
    - 2.2 Clustering
    - 2.3 Topic Models (Text Analysis)
3. Supervised Learning
    - 3.1 Class imbalance
    - 3.2 Random Forests, Boosting, Bagging
    - 3.3 Interpretability
    - 3.4 Bootstrap

ML is still a black box sometimes...

- No star gazing allowed
- No estimable f
- Non-parametric modeling
- Lots of hyperparamaters

...but it can be unpacked

Figure 26: Picasso self-portraits over the years

Quantitative political risk is going to grow

- Quantification political concepts
- Explosion big data
- Growth in computation processing

Our Forecasts...

Industry

- GenAI and ML will disrupt conventional industry...
- ...but AI-driven products still unclear (content, CoT models, agentic, forecasts, ...?)
- Trust and transparency requirements will increase
- Risk of over-confidence

World Ahead

- EM-ification of US
- Heightened asset sensitivity to politics
- Misinformation and disinformation risk grows

Further resources

- Texts: Elements of Statistical Learning, Interpretable ML, Toward Data Science (Medium), ICML working papers
- Key influencers: Andy Gelman, Phil Schrodt, Havard Hegre (Uppsala Conflict Programme), Jake Shapiro (Empirical Studies of Conflict), CSIS Futures Lab
- Courses: Intro to MLOps (Weights and Biases), AI Engineering (IBM), Advanced NLP with spACY (spacy), GenAI with LLMs (DeepLearning.ai)

Social Science is Key to ML

Black box means we need social science expertise (more than ever) to...

- ask the right question
- assemble the right inputs
- validate the outputs
- refine and reiterate the model

You’ve got this!