Interpreting Machine Learning Models

Key Formulas & Definitions

Pierre Pudlo

1 Model Interpretation: Overview

Machine learning models present a fundamental trade-off between predictive accuracy and interpretability. Whilst complex architectures (neural networks, ensemble methods) achieve superior predictive performance, their decision-making processes remain inherently opaque—a phenomenon often termed the "black box" problem. Interpretation methods endeavour to elucidate these processes, thereby addressing scientific, regulatory, and practical imperatives across diverse application domains.

Two principal approaches:

- **Intrinsic interpretability:** Models whose structure renders them inherently transparent by design (e.g., linear models, decision trees)
- **Post-hoc interpretability:** Techniques applied to arbitrary models subsequent to training (e.g., SHAP, LIME, permutation-based feature importance)

Applications encompass: Scientific discovery and hypothesis generation, regulatory compliance (notably GDPR Article 22), algorithmic fairness auditing, model debugging and validation, and stakeholder communication in high-stakes domains.

2 Intrinsically Interpretable Models

2.1 Linear and Logistic Regression

Linear regression:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Interpretation: β_j quantifies the change in $\hat{y}$ when x_j increases by one unit, *ceteris paribus.*

Logistic regression: For $p(x) = P(Y = 1|X = x)$:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Odds ratio: $\exp(\beta_j)$ represents the multiplicative change in odds for a one-unit increase in x_j.

2.2 Standardisation

Problem: Regression coefficients exhibit scale-dependence, thereby precluding meaningful comparison across features measured in disparate units.

Solution: Apply z-score standardisation to all covariates:

$$x_j^{\text{std}} = \frac{x_j - \bar{x}_j}{\text{sd}(x_j)}$$

Consequence: Standardised coefficients quantify the change in $\hat{y}$ per one standard deviation increase in x_j, facilitating direct comparison of relative feature importance.

Critical consideration: Standardisation parameters ($\bar{x}_j$, $\text{sd}(x_j)$) must be estimated exclusively from training data; the identical transformation is subsequently applied to test data to prevent information leakage.

2.3 Limitations of Linear Models

Multicollinearity: When covariates exhibit substantial correlation, regression coefficients become unstable and their interpretation unreliable, as small perturbations in the data may induce disproportionate changes in estimated parameters.

Interaction effects: Linear models presume strictly additive feature contributions; any interaction effects must be explicitly incorporated through manually specified product terms.

Non-linearity: The assumption of linear relationships may prove untenable; polynomial terms, splines, or basis function expansions are required to capture non-linear patterns in the data-generating process.

2.4 Decision Trees

Feature importance via impurity reduction:

$$\text{Importance}_j = \sum_{t:\text{split on } X_j} \text{Impurity decrease at node } t$$

Advantages: Excellent visual interpretability through graphical representation of decision rules; naturally accommodates categorical predictors without encoding; captures interaction effects and non-linearities without explicit specification.

Limitations: High variance renders individual trees unstable; performance typically inferior to ensemble methods; interpretability diminishes rapidly with tree depth; prone to overfitting without appropriate regularisation.

3 Global Interpretation Methods

3.1 Permutation Importance

Principle: Assess feature importance by quantifying predictive performance degradation induced by random permutation of feature values, thereby destroying the association between the feature and the response whilst preserving marginal distributions.

Algorithm:

1. Establish baseline performance on validation set: Score_0
2. For each feature X_j:
   - Randomly permute values of X_j across observations in validation set
   - Evaluate perturbed model performance: Score_j
   - Compute importance as degradation: $\text{Importance}_j = \text{Score}_0 - \text{Score}_j$
3. Repeat permutation procedure K times to obtain stable estimates via averaging

Advantages: Model-agnostic framework applicable to arbitrary learning algorithms; grounded in actual predictive performance metrics rather than model-internal statistics; naturally accounts for feature interactions and non-linear relationships.

Limitations: Correlated features pose interpretational challenges, as the model may substitute one feature for another; computationally intensive for large datasets; estimates exhibit variance across different random permutations, necessitating multiple repetitions.

3.2 Gini/Gain Importance (Tree-Based Models)

For tree-based ensembles comprising B trees:

$$\text{Importance}_j = \frac{1}{B} \sum_{b=1}^{B} \sum_{t:\text{split on } X_j} \text{Impurity decrease at node } t$$

Advantages: Computationally efficient as importance scores are calculated during training without additional model evaluations; readily available in standard implementations (e.g., `feature_importances_` in scikit-learn).

Disadvantages: Exhibits systematic bias favouring high-cardinality features and variables with numerous possible split points; unreliable in the presence of correlated features; applicability restricted exclusively to tree-based architectures.

3.3 Partial Dependence Plots (PDP)

Definition:

$$\text{PDP}_j(x_j) = \mathbb{E}_{X_{\setminus j}}\left[\hat{f}(x_j, X_{\setminus j})\middle|\hat{f}\right]$$

Estimation via Monte Carlo approximation:

$$\widehat{\text{PDP}}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_j, x_i^{(\setminus j)})$$

where $x_i^{(\setminus j)}$ denotes the i-th observation's values for all features except X_j.

Interpretation: The PDP represents the marginal effect of X_j on predictions, obtained by averaging over the empirical distribution of all remaining features.

Limitations:

- **Independence assumption:** Presumes features are independent; produces misleading results when features exhibit correlation, as the procedure generates counterfactual observations lying outside the support of the joint distribution
- **Heterogeneity masking:** Displays only average effects, thereby obscuring potentially important heterogeneous responses across different subpopulations
- **Extrapolation artefacts:** May evaluate the model at unrealistic feature combinations, particularly problematic in regions of sparse data coverage

3.4 Individual Conditional Expectation (ICE)

Definition:

$$\text{ICE}_i(x_j) = \hat{f}(x_j, x_i^{(\setminus j)})$$

Interpretation: The ICE curve represents the predicted response trajectory for observation i as X_j varies whilst all other features remain fixed at their observed values $x_i^{(\setminus j)}$.

Relationship to PDP: One curve per observation; the PDP constitutes the pointwise average of all ICE curves, i.e., $\text{PDP}_j(x_j) = n^{-1}\sum_{i=1}^{n}\text{ICE}_i(x_j)$.

Advantage: Reveals heterogeneity in feature effects and potential interactions obscured by the averaging inherent in PDPs; parallel ICE curves indicate homogeneous effects, whilst divergent curves suggest interactions or subgroup-specific responses.

3.5 Accumulated Local Effects (ALE)

Motivation: PDPs generate counterfactual observations by marginalising over features, thereby creating unrealistic feature combinations when variables exhibit correlation. This violates the support of the true joint distribution and can yield misleading interpretations.

Principle: Quantify feature effects through accumulated local changes evaluated exclusively within the empirical correlation structure of the data, thereby respecting the manifold upon which observations lie.

For interval $I_k = [a_k, a_{k+1}]$ partitioning the range of X_j:

$$\widehat{\text{ALE}_j}(k) = \mathbb{E}\Big[\hat{f}(a_{k+1}, X_{\setminus j}) - \hat{f}(a_k, X_{\setminus j})\Big|X_j \in I_k\Big]$$

The ALE function is then estimated by accumulating these local differences across intervals, yielding a curve that represents the integrated effect of X_j:

$$\text{ALE}_j(x_j) = \int_{z=\min}^{x_j}\mathbb{E}\Big[\frac{\partial\hat{f}(z, X_{\setminus j})}{\partial z}\Big|X_j = z\Big]dz - \mathbb{E}\left[\int_{z=\min}^{X_j}\mathbb{E}\Big[\frac{\partial\hat{f}(z, X_{\setminus j})}{\partial z}\Big|X_j = z\Big]dz\right]$$

(The second term centres the ALE function to have zero mean: $\mathbb{E}[\text{ALE}_j(X_j)] = 0$.)

Advantages over PDP:

- **Respects correlation structure:** Evaluates model predictions only at realistic feature combinations observed in the data
- **Unbiased effect estimation:** Provides valid marginal effect estimates even in the presence of substantial feature correlation

- **Computational efficiency:** Often faster than PDP, particularly for large datasets, as local averaging requires fewer model evaluations

Recommendation: Employ ALE when pairwise correlations exceed $|\rho| > 0.3$; for weakly correlated features, PDP remains adequate and more intuitive.

3.6 Two-Dimensional Partial Dependence

Definition:

$$\widehat{\mathrm{PDP}}_{jk}(x_j, x_k) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_j, x_k, x_i^{(\setminus jk)})$$

Purpose: Visualise and quantify interaction effects between features X_j and X_k by examining how their joint variation influences predictions.

Interpretation:

- **Parallel contours:** Indicate additive (non-interacting) effects; the influence of one feature does not depend upon the value of the other
- **Non-parallel or curved contours:** Signify interaction effects; the marginal effect of one feature varies systematically with the level of the other

3.7 Friedman's H-statistic

Definition: Provides a scalar measure quantifying interaction strength between features j and k:

$$H_{jk}^2 = \frac{\sum_i \Big(\mathrm{PDP}_{jk}(x_{ij}, x_{ik}) - \mathrm{PDP}_j(x_{ij}) - \mathrm{PDP}_k(x_{ik})\Big)^2}{\sum_i \mathrm{PDP}_{jk}(x_{ij}, x_{ik})^2}$$

The numerator quantifies departure from additivity, whilst the denominator normalises by total variation.

Interpretation:

- $H_{jk} = 0$: Perfect additivity; effects are strictly non-interacting
- $H_{jk} > 0$: Interaction detected; magnitude indicates relative strength
- Typical threshold: $H_{jk} > 0.1$ warrants further investigation

Application: Efficient screening mechanism to identify salient interactions amongst $\binom{p}{2}$ potential pairs prior to computationally expensive 2D PDP visualisation.

4 Fairness and Model Auditing

4.1 Fairness Metrics

Let A denote a protected attribute (e.g., gender, race, ethnicity) with $A \in \{0, 1\}$ representing group membership.

Demographic Parity (Statistical Parity):

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

Interpretation: The rate of positive predictions (e.g., loan approvals, hirings) must be identical across protected groups, irrespective of potentially differing base rates in the population. This criterion demands equal treatment but ignores whether such equality is justified by differential qualifications.

Equalized Odds:

$$P(\hat{Y} = 1|Y = y, A = 0) = P(\hat{Y} = 1|Y = y, A = 1) \quad \text{for } y \in \{0, 1\}$$

Interpretation: Both true positive rates (sensitivity) and false positive rates must be equalised across groups, conditional on the actual outcome. This constitutes a more stringent requirement than demographic parity, ensuring predictive equality for both qualified and unqualified individuals.

Equal Opportunity:

$$P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1)$$

Interpretation: True positive rates (recall) must be equal across groups; qualified individuals should receive positive predictions at equal rates regardless of group membership. This represents a relaxation of equalised odds, focusing exclusively on error rates for the positive class.

4.2 Disparate Impact

Legal definition (US Equal Employment Opportunity Commission):

$$\text{Disparate Impact Ratio} = \frac{P(\hat{Y}=1|A=1)}{P(\hat{Y}=1|A=0)}$$

Four-fifths rule: If the ratio falls below 0.8, prima facie evidence of adverse impact exists, triggering further investigation. This threshold derives from US employment discrimination law and lacks statistical justification.

Note: Disparate impact constitutes a legal screening tool rather than definitive proof of unlawful discrimination; contextual factors and business necessity defences may apply.

4.3 Fairness-Accuracy Trade-off

Impossibility theorem (Kleinberg, Mullainathan & Raghavan, 2017):

Except in degenerate cases where base rates are identical across groups or the classifier achieves perfect accuracy, one cannot simultaneously satisfy:

- **Calibration:** Predicted probabilities equal true conditional probabilities, i.e., $P(Y=1|\hat{f}(X)=s, A=a)=s$ for all s, a
- **Equal false positive rates:** $P(\hat{Y}=1|Y=0,A=0)=P(\hat{Y}=1|Y=0,A=1)$
- **Equal false negative rates:** $P(\hat{Y}=0|Y=1,A=0)=P(\hat{Y}=0|Y=1,A=1)$

Implication: Practitioners must make normative choices regarding which fairness criterion best aligns with the ethical requirements and legal constraints of their specific application domain. No universal solution exists; transparency regarding trade-offs is essential.

5 Local Interpretation Methods

5.1 LIME (Local Interpretable Model-Agnostic Explanations)

Principle: Approximate the behaviour of a complex model locally in the vicinity of a prediction of interest through a sparse, interpretable surrogate model.

Optimisation objective:

$$\xi(x_0) = \arg\min_{g\in\mathcal{G}} \mathcal{L}(f, g, \pi_{x_0}) + \Omega(g)$$

where:

- f: black-box model to be explained
- g: interpretable surrogate model (typically linear or sparse additive)
- $\mathcal{G}$: hypothesis class of interpretable models
- $\mathcal{L}(f, g, \pi_{x_0})$: locally weighted loss measuring fidelity of g to f
- π_{x_0}: proximity kernel determining the neighbourhood of x_0
- $\Omega(g)$: complexity penalty encouraging sparse explanations

Algorithm:

1. Select instance x_0 requiring explanation
2. Generate synthetic neighbourhood via perturbations (e.g., Gaussian noise, feature sampling)
3. Obtain black-box predictions $f(x')$ for perturbed instances x'
4. Weight observations by proximity: $w_i = \pi_{x_0}(x'_i)$
5. Fit interpretable model g via weighted regression on synthetic data
6. Extract and interpret coefficients as feature attributions

Key parameters:

- `num_samples`: Number of synthetic perturbations (default 5000; increase for stability)
- `kernel_width`: Bandwidth parameter σ controlling locality (default $\sqrt{p} \times 0.75$)
- `num_features`: Sparsity level K for explanation (5–10 recommended for human interpretation)

Strengths: Model-agnostic framework applicable to arbitrary models; provides locally faithful approximations; yields intuitive explanations via interpretable models; computationally efficient.

Limitations: Explanations may exhibit instability across repeated executions; sensitive to kernel bandwidth selection; quality contingent upon perturbation strategy; local fidelity does not guarantee global consistency.

Best practice: Generate multiple explanations with different random seeds and verify consistency; report confidence intervals; validate local fidelity (R^2) of surrogate model.

5.2 SHAP (SHapley Additive exPlanations)

Theoretical foundation: Shapley values, originating from cooperative game theory (Shapley, 1953), provide an axiomatic approach to distributing total payoff amongst cooperating players based on their marginal contributions.

Shapley value for feature j:

$$\phi_j = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} \left[\hat{f}(S \cup \{j\}) - \hat{f}(S)\right]$$

where:

- S: arbitrary subset (coalition) of features excluding j
- $\hat{f}(S)$: model prediction using exclusively features in S
- $\hat{f}(S \cup \{j\}) - \hat{f}(S)$: marginal contribution of feature j to coalition S
- Weighting factor: $\frac{|S|!(p-|S|-1)!}{p!}$ ensures all orderings of feature inclusion receive equal consideration

Interpretation: ϕ_j quantifies the average marginal contribution of feature j across all possible coalitions, weighted by the probability of each coalition forming. This constitutes a fair attribution scheme grounded in axiomatic principles.

5.3 SHAP Properties

Axioms satisfied by Shapley values:

1. **Efficiency (Local accuracy):** $\sum_{j=1}^{p} \phi_j = \hat{f}(x) - \hat{f}(\emptyset)$ Feature attributions sum precisely to the difference between the prediction and baseline
2. **Symmetry:** If features j and k contribute identically to all coalitions, then $\phi_j = \phi_k$ Equivalent contributions receive equivalent attributions
3. **Dummy (Null player):** If feature j contributes nothing, then $\phi_j = 0$ Irrelevant features receive zero attribution
4. **Additivity (Linearity):** For composite models $\hat{f} = \hat{f}_1 + \hat{f}_2$, attributions sum: $\phi_j(\hat{f}) = \phi_j(\hat{f}_1) + \phi_j(\hat{f}_2)$

Uniqueness theorem: Shapley values constitute the *unique* attribution method satisfying these four axioms simultaneously, providing theoretical justification for their use.

5.4 SHAP Approximation Methods

Computational challenge: Exact Shapley value computation necessitates evaluating 2^p coalitions, rendering brute-force calculation infeasible for high-dimensional problems.

Tractable approximations:

- **KernelSHAP:** Model-agnostic approximation via weighted linear regression on coalition samples; theoretically grounded but computationally intensive
- **TreeSHAP:** Polynomial-time exact computation leveraging tree structure; exploits conditional independence in tree-based models
- **DeepSHAP:** Efficient approximation for deep neural networks via compositional structure and reference values
- **LinearSHAP:** Analytical solution for linear models; coefficients scaled by feature values

Practical recommendation: TreeSHAP for tree ensembles (random forests, gradient boosting); KernelSHAP for other architectures when computational budget permits.

5.5 SHAP Global Importance

Aggregated feature importance via mean absolute attribution:

$$\text{Importance}_j = \frac{1}{n} \sum_{i=1}^{n} \left|\phi_j^{(i)}\right|$$

Advantage: Provides coherent global-local interpretation framework; global importance rankings derived consistently from instance-level explanations; naturally accounts for both magnitude and sign of contributions across the dataset.

5.6 LIME vs SHAP: Comparative Assessment

Aspect	LIME	SHAP
Theoretical foundation	Heuristic local approximation	Axiomatic (game-theoretic)
Consistency	May produce inconsistent attributions	Provably consistent
Computational complexity	$O(K \cdot n_{\text{samples}})$	$O(2^p)$ exactly; $O(n_{\text{trees}} \cdot p^2)$ for TreeSHAP
Stability	Sensitive to sampling	More robust
Scope	Local explanations only	Unified local-global framework
Interpretability	Highly intuitive	Requires theoretical understanding

Recommendation: Employ SHAP when feasible, particularly for tree-based architectures where TreeSHAP enables efficient exact computation. Resort to LIME for rapid prototyping, exploratory analysis, or when SHAP proves computationally prohibitive.

6 Practical Recommendations

6.1 Best Practices

Model selection hierarchy:

1. Commence with intrinsically interpretable architectures (linear models, shallow decision trees, GAMs)
2. Escalate to complex models solely when interpretable alternatives demonstrably underperform
3. Rigorously document the rationale underlying architectural choices and performance trade-offs

Interpretation workflow:

1. Establish global feature importance via permutation-based methods
2. Conduct detailed attribution analysis using SHAP values
3. Visualise marginal effects through PDPs or ALEs (selecting the latter for correlated features)
4. Validate all quantitative findings against domain expertise and substantive knowledge

Critical considerations:

- Deploy multiple complementary interpretation methods to triangulate findings
- Systematically verify consistency across different methodological approaches
- Report confidence intervals, standard errors, and stability diagnostics
- Screen for problematic feature correlations ($|\rho| > 0.3$)
- Conduct fairness audits across protected demographic groups

6.2 Common Pitfalls

Interpretation Causation:

- Supervised learning models capture associational patterns, not causal mechanisms
- Feature importance quantifies predictive utility, not causal influence
- Causal inference necessitates substantially stronger assumptions: experimental randomisation, conditional ignorability, valid instrumental variables, or structural causal models
- Confounding variables may induce spurious attributions

Methodological limitations to acknowledge:

- No singular interpretation method furnishes complete understanding
- All techniques embody implicit assumptions and possess identifiable blind spots
- Correlated predictors systematically confound interpretation and attribution
- Computational resource constraints materially constrain methodological choices in production environments

Fairness considerations:

- Merely excluding protected attributes from feature sets proves insufficient; proxy variables inevitably encode protected information
- Multiple competing fairness criteria exist; simultaneous satisfaction typically proves mathematically impossible
- Practitioners must explicitly select fairness criteria aligned with normative values and legal requirements
- Transparently document fairness desiderata, trade-offs accepted, and validation procedures employed

7 Resources

Essential references:

- Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book/
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. https://hastie.su.domains/ElemStatLearn/
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?: Explaining the Predictions of Any Classifier." *KDD*
- Lundberg, S. M., & Lee, S. I. (2017). "A Unified Approach to Interpreting Model Predictions." *NIPS*
- Apley, D. W., & Zhu, J. (2020). "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models." *JRSS-B*

Python libraries:

- `scikit-learn`: https://scikit-learn.org/ (built-in interpretation tools)
- `shap`: https://github.com/slundberg/shap
- `lime`: https://github.com/marcotcr/lime
- `PyALE`: https://github.com/DanaJomar/PyALE
- `fairlearn`: Fairness assessment and mitigation

Documentation:

- SHAP documentation: https://shap.readthedocs.io/
- Interpretable ML book: https://christophm.github.io/interpretable-ml-book/