5.Global Model-Agnostic Methods

What Are Global Methods?

•Explain how model behaves across entire dataset
•Not focused on individual predictions
•Show average or overall patterns

Two Methods Covered:

1. Partial Dependence Plot (PDP) - Friedman 2001

•Shows marginal effect of features on predictions
•Averages over all instances

2. Surrogate Models

•Approximates black box with interpretable model
•Global interpretation through simpler model

Common Characteristics:

•Model-agnostic (work with any ML model)
•Provide global insights
•Help understand overall model behavior

Partial Dependency Plots

What PDP Shows:

- Relationship between feature(s) and predicted outcome
- Whether relationship is linear, monotonic, or complex
- Average effect across all instances

Mathematical Definition:

$$\hat{f}_S(\mathbf{x}_S) = \mathbb{E}_{\mathbf{X}_C}\left[\hat{f}(\mathbf{x}_S, X_C)\right] = \int \hat{f}(\mathbf{x}_S, X_C) d\mathbb{P}(\mathbf{X}_C)$$

Components:

- xS: Feature(s) for which PDP is plotted (usually 1-2 features)
- XC: Other features (treated as random variables)
- f: Machine learning model
- Marginalization: Average over distribution of XC

Key Idea:

- Fix feature(s) in S at certain values
- Average predictions over all possible values of other features C
- Get function depending only on S (with interactions included)

PDP - Estimation

Formula:

$$\hat{f}_S(\mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(\mathbf{x}_S, \mathbf{x}_C^{(i)}\right)$$

Algorithm:

Step 1: Select feature(s) S to analyze (e.g., temperature)

Step 2: Choose grid of values for xS (e.g., temperature from 0°C to 40°C)

Step 3: For each value in grid:

- Replace feature S with that value in ALL n instances
- Keep all other features C at their original values
- Get n predictions from model

Step 4: Average the n predictions

Step 5: Repeat for all values in grid

Step 6: Plot xS values (x-axis) vs. average predictions (y-axis)

Equivalent Interpretation:

- Averaging all ICE (Individual Conditional Expectation) curves
- PDP = average marginal effect

PDP - Categorical Features

Process for Categorical Features:

•Very simple compared to numerical features
•For each category, get one PDP estimate

Algorithm:

Step 1: Identify all categories (e.g., seasons: spring, summer, fall, winter)
Step 2: For each category:
•Force ALL data instances to have that category
•Get predictions for all instances
•Average the predictions
Step 3: Result = one average prediction per category
Example - Bike Rental by Season:
•Winter: Average prediction = 3500 bikes
•Spring: Average prediction = 4200 bikes
•Summer: Average prediction = 5800 bikes
•Fall: Average prediction = 5500 bikes

Visualization:

•Bar chart or point plot
•Shows effect of each category on prediction

PDP Examples - Weather Features

Temperature Effect:

•Hotter → more bikes rented
•Trend increases up to 20°C, then flattens
•Slight drop around 30°C

Humidity Effect:

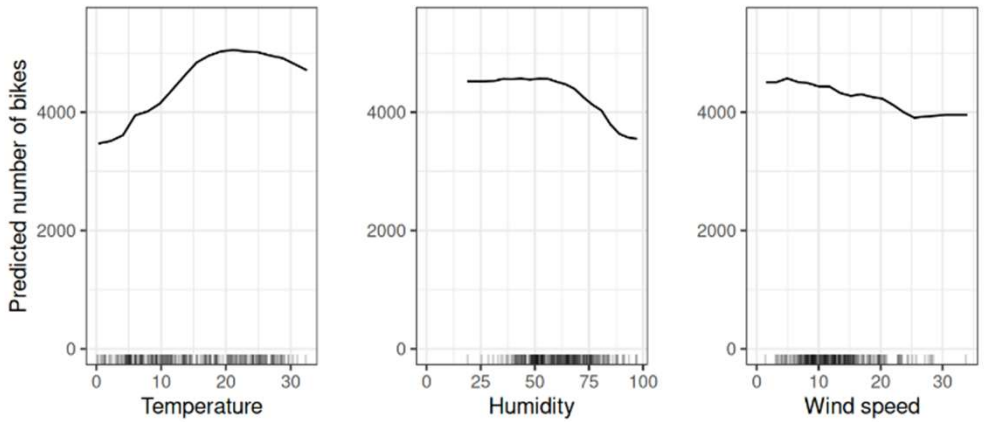
•Humidity > 60% → fewer bike rentals
•Clear negative relationship
•Bikers increasingly inhibited above 60%

Wind Speed Effect:

•More wind → fewer bike rentals
•Exception: 25-35 km/h range shows no decrease
•Likely due to limited training data in that range
•Intuitively, higher wind should decrease rentals

Key Insight:

•PDPs reveal what model learned
•Can identify data-sparse regions
•Can validate sensible patterns



Categorical Feature Analysis

Setup:

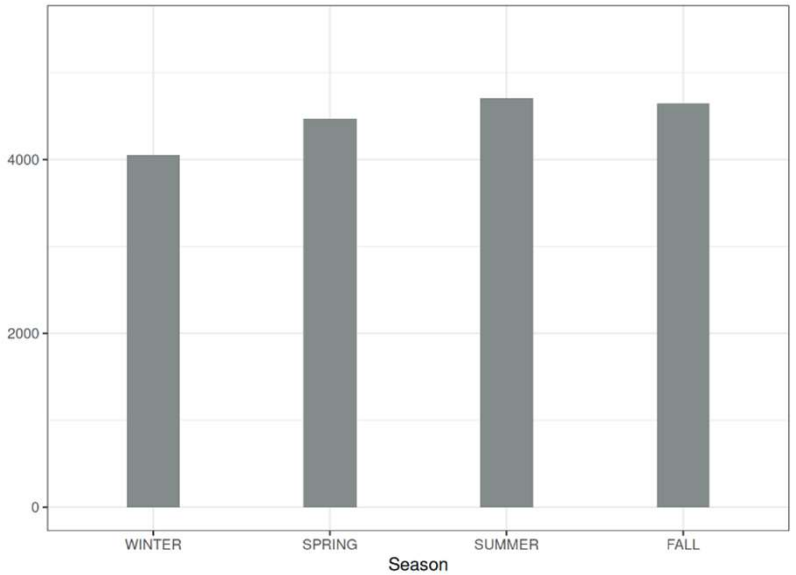
•Feature: Season (categorical)
•Model: Random forest
•Target: Daily bike rentals

Results:

•Winter: Lowest predicted rentals (clear negative effect)
•Spring: Slightly fewer than average
•Summer: Similar effect to fall
•Fall: Similar effect to summer

Interpretation:

•All seasons show similar effect on predictions
•Only winter stands out with notably fewer rentals
•Model learned seasonal patterns appropriately



PDP Example - Penguin Classification

Setup:

•Predict P(female) from body measurements
•Model: Random forest
•Features analyzed: Body mass, bill depth

Body Mass Effect:

•Heavier penguin → lower P(female)
•Clear negative relationship

Bill Depth Effect:

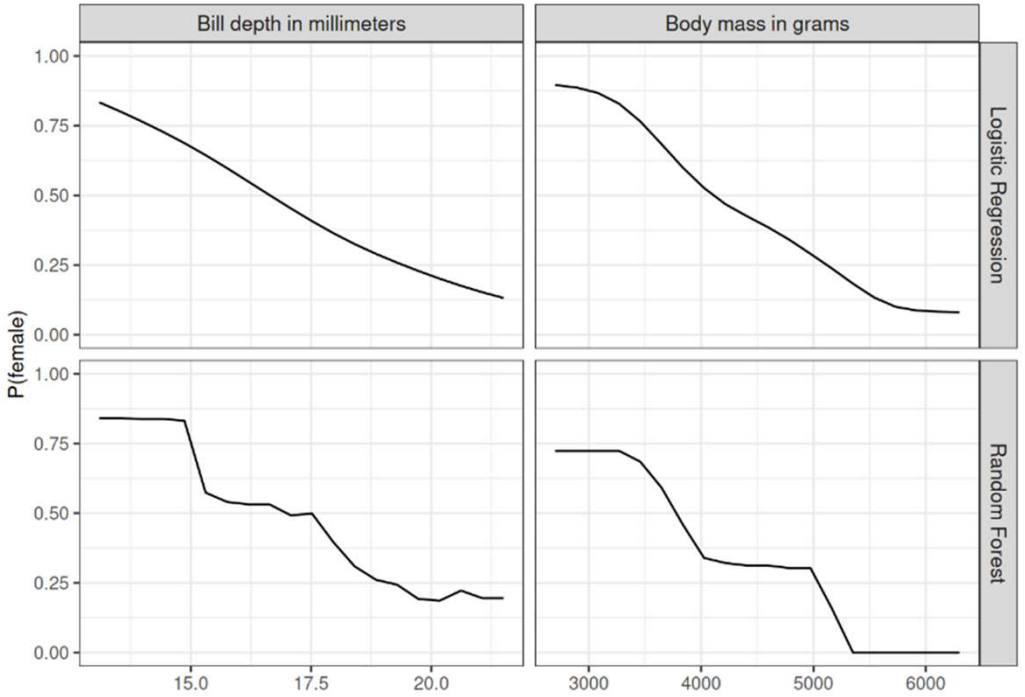
•Deeper bill → lower P(female)
•Similar pattern to body mass

Visual Characteristic:

•Random forest PDP is rugged due to tree-based structure
•Shows complexity of learned relationships

Big Problem with Figure 19.3:

•Throws all penguin species together
•Species-specific patterns hidden in aggregation



PDP-Based Feature Importance

Motivation (Greenwell et al., 2018):

- Flat PDP → feature not important
- More PDP varies → more important feature

For Numerical Features:

$$I(\mathbf{x}_S) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\hat{f}_S(\mathbf{x}_S^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\hat{f}_S(\mathbf{x}_S^{(k)}))^2}$$

- K = number of unique feature values
- Measures standard deviation of PDP

For Categorical Features:

$$I(X_S) = \frac{\max_k(\hat{f}_S(\mathbf{x}_S^{(k)})) - \min_k(\hat{f}_S(\mathbf{x}_S^{(k)}))}{4}$$

- Range of PDP values divided by 4
- "Range rule" for rough variance estimate
- Denominator 4 from normal distribution (95% within ±2 SD)

Caution:

- Captures only main effect (ignores interactions)
- Feature could be important via interactions but have flat PDP
- Unique values weighted equally

PDP Strengths

Strength 1: Intuitive

- Easy concept: "Average prediction if we force feature to this value"
- Lay people grasp PDPs quickly
- No technical background needed

Strength 2: Clear Interpretation (Uncorrelated Features)

- If feature not correlated with others → perfect representation
- Shows how average prediction changes when feature changes
- More complicated when features correlated

Strength 3: Easy Implementation

- Simple algorithm
- Available in most ML libraries

Strength 4: Causal Interpretation

- We intervene on feature, measure prediction changes
- Analyzes causal relationship for MODEL
- Relationship is causal within model structure
- NOT necessarily causal for real world (Zhao and Hastie 2019)

Limitation 1: Maximum 2 Features

- Realistic visualization limit
- 1 feature → 2D plot
- 2 features → 3D surface/heatmap
- Beyond that: Cannot visualize meaningfully
- Not PDP's fault, but representation constraints

Limitation 2: Feature Distribution Often Missing

- Some PD plots don't show distribution
- Can overinterpret regions with almost no data
- Solution:
    - Show rug (data point indicators on x-axis)
    - Show histogram
    - Prevents misleading interpretation in sparse regions

When to Use PDP:

•Understand overall feature effects
•Features not strongly correlated
•Check if model learned sensible relationships
•Compare different model behaviors

When to Be Cautious:
•Strongly correlated features → use ALE plots
•Suspected interactions → check 2D PDPs or ICE curves
•Heterogeneous effects expected → use regional PDP or ICE

Tips for Better Interpretability:
•Reduce model complexity:
- Lower tree depth for tree-based models
- Add monotonicity constraints

•Optimize for lower interactions and sparsity (Molnar et al., 2020)
•Simpler models → simpler, more reliable PDP interpretation

Surrogate Models - Introduction

Definition:

- A surrogate model is an **interpretable model** trained to mimic the predictions of a black-box model
- By examining the surrogate, we can draw conclusions about how the black-box model behaves.

Alternative Names:

- Approximation model
- Metamodel
- Response surface model
- Emulator

Origin:

- The idea originally comes from **engineering**, where surrogate models approximate expensive simulations.

Key Difference from Engineering:

- Underlying model is ML model (not physical simulation)
- Surrogate must be interpretable (additional constraint)

Surrogate Models - Theory

Goal:
- Approximate black box prediction function f̂ with surrogate g
- g must be interpretable
- g should mimic f̂ as closely as possible

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Framework:
- Black box: f̂(x) - any complex ML model
- Surrogate: g(x) - must be interpretable

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

Where:
Rm = rectangular regions, cm = constant prediction

Key Properties:
- Model-agnostic method
- Only needs: Data + prediction function
- No information about black box internals
- Choice of black box and surrogate decoupled
- Can replace either independently

Step 1: Select Dataset X

•Training data used for black box (can also use new dataset from same distribution or subset of original data or grid of feature values)

Step 2: Get Black Box Predictions

•For each x^(i) in X, get $\hat{y}^{(i)} = \hat{f}(x^{(i)})$

•These become surrogate's target values

Step 3: Select Interpretable Model Type

•Linear model (for coefficient interpretation)

•Decision tree (for rule interpretation)

•Rule-based model

Step 4: Train Surrogate

•Train g on dataset (X, $\hat{y}$)

•Critical: Target is $\hat{y}$ (black box predictions), NOT original labels y

•Surrogate learns to mimic black box

Step 5: Measure Fidelity

•Calculate R-squared (explained next slide)

•Quantifies approximation quality

Step 6: Interpret Surrogate

•Analyze coefficients / tree structure / rules

•Interpretations explain black box behavior

Step 7: Evaluate and Use

•Check if R^2 sufficient for needs

•Use for understanding and communication

Measuring Approximation Quality

Components:

•$\hat{y}^{(i)}$: Surrogate prediction for instance i
•$\hat{y}^{(i)}$: Black box prediction for instance i
•$\bar{\hat{y}}$: Mean of black box predictions
•SSE: Sum of squared errors (surrogate vs. black box)
•SST: Total sum of squares

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_*^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{\hat{y}})^2}$$

Interpretation:

•R^2 = 0.9 (low SSE): Surrogate captures 90% of variance
  - Excellent approximation
  - Might even replace black box
•R^2 = 0.5 (medium SSE): Moderate approximation
  - Use with caution
  - Some patterns captured
•R^2 close to 0 (high SSE): Poor approximation
  - Surrogate unreliable
  - Don't trust interpretations

Critical Note:

•R^2 measures surrogate fidelity to BLACK BOX
•NOT black box performance on true labels
•If black box performs badly, surrogate interpretations become irrelevant

Surrogate Models - Variants

Standard (Global) Surrogate:
- Uses entire dataset
- Equal weight to all instances
- Explains overall model behavior
- Interpretation applies globally

Weighted/Subset Surrogates:
- Use subset of data
- Re-weight instances differently
- Focus on specific region or subpopulation
- Changes in distribution cause changes in interpretation
- No longer truly global

Local Surrogate:
- Weight data by proximity to specific instance
- Closer instances get higher weight
- Explains individual prediction
- This is LIME (covered in earlier slides)

Flexibility:
- Same framework, different scopes
- Choose based on interpretation goal

Surrogate Example - Regression

Example: Bike Rental SVM

Setup:

- Black box: Support vector machine
- Surrogate: CART decision tree
- Data: Test data (not training data)
- Task: Predict daily bike rentals

Results:

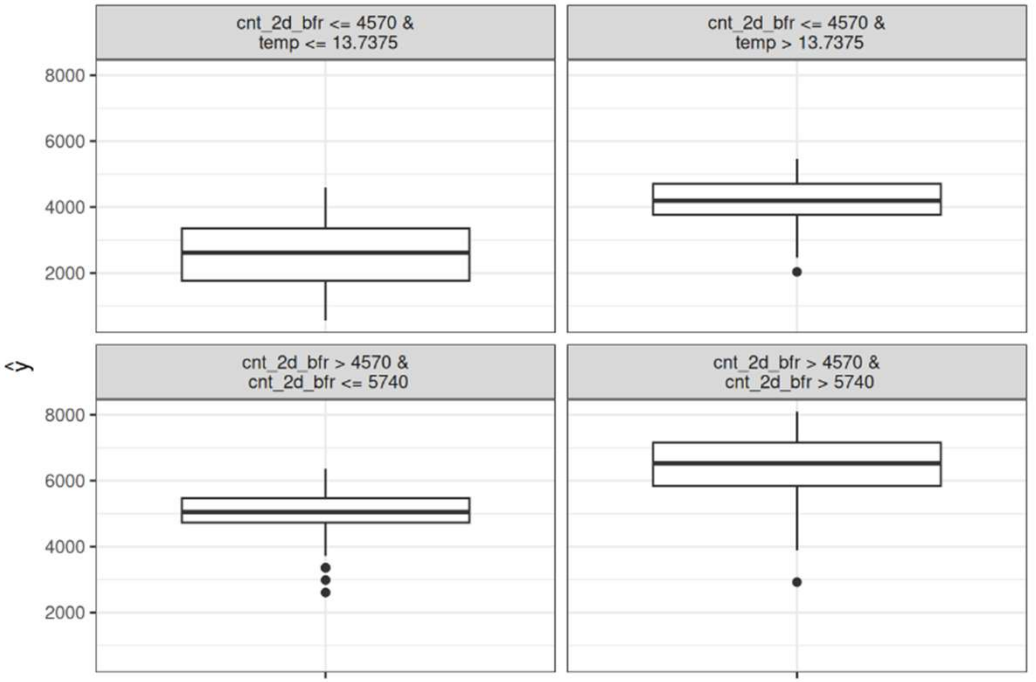
- $R^2 = 0.76$ on test data
- Good but not perfect approximation

Tree Interpretation:

- Temperature > 13°C: More bikes predicted
- Count 2 days before: Higher previous count → more bikes
- Tree captures main patterns SVM learned

Assessment:

- Surrogate provides reasonable approximation
- Can use for understanding SVM behavior
- 24% of variance not captured (acceptable trade-off for interpretability)



Surrogate Models - Strengths

Strength 1: Flexibility

- ANY interpretable model can be used as surrogate
- Can exchange interpretable model type
- Can exchange black box model
- One black box → multiple surrogate types

Example Use Case:

Team A understands linear models → linear surrogate

Team B understands trees → tree surrogate

Same black box, different explanations for different audiences

Strength 2: Intuitive Concept

- Easy to implement
- Easy to explain to non-experts
- "Train simple model to mimic complex model"
- Straightforward methodology

Strength 3: Measurable Fidelity

- R-squared quantifies approximation quality
- Objective metric for evaluation
- Know how much to trust interpretations

Strength 4: Model-Agnostic

- Works with any black box model
- No need for internal model access
- Only requires prediction function

Surrogate Models - Limitations

Limitation 1: Conclusions About Model, Not Data

- Surrogate trained on black box predictions $\hat{y}$
- Never sees real outcomes y
- Interprets MODEL behavior, not real-world relationships. If black box wrong, surrogate explains wrong behavior
- Interpretations not directly meaningful for reality

Limitation 2: No Clear R^2 Threshold

- No principled cutoff
- Context-dependent decision
- Need domain expertise

Limitation 3: Heterogeneous Approximation

- Might approximate well for one subset
- Widely divergent for another subset
- Overall R^2 can mask regional problems
- Interpretation not equally valid everywhere

Limitation 4: Inherits Surrogate Model's Issues

- Linear model: Assumes linearity
- Decision tree: Instability, axis-aligned splits
- Each interpretable model has limitations

Limitation 5: Interpretability Debate

- Some argue NO model is intrinsically interpretable, Even "simple" models can be opaque
- Illusion of interpretability could be dangerous

Use PDP When:

- Understanding specific feature effects
- Checking learned relationships
- Features relatively uncorrelated
- Quick visual insights

Use Surrogate When:

- Complete model approximation
- Interpretable replacement needed
- Multiple audiences
- Quantified fidelity required