決定木アルゴリズム

経済学のための機械学習入門

川田恵介

Table of contents

Get started

- アルゴリズム ≃ 推定手法
- 決定木 = 非常に優れた出発点
  - OLS との高い補完性
  - 直感的

サブグループ分析

- “もっとも” よく使われるデータ活用方法
- X 上にサブグループ A_j を定義し、予測モデル $g(X)$ を以下のルールで生成

$$g(X_i) = E[Y|X_i \in A_j]$$

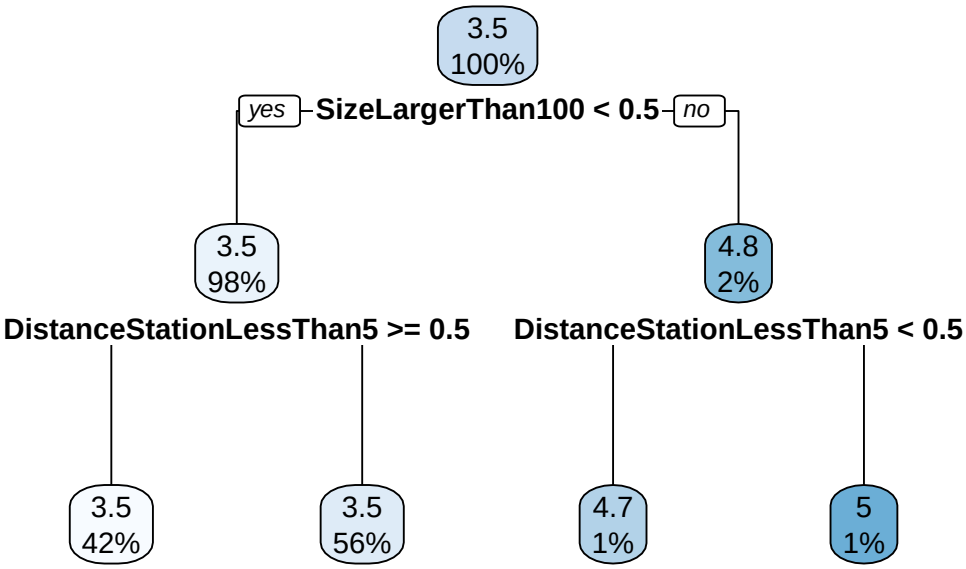
“伝統的” VS Data adaptive modelling

- 伝統的アプローチ: 研究者が事前 (データを見る前に) に A_j を定義
- Data adaptive: データが決定
- (注) “Bad practice” : 研究者がデータ $\{Y, X\}$ を見ながら A_j を決定

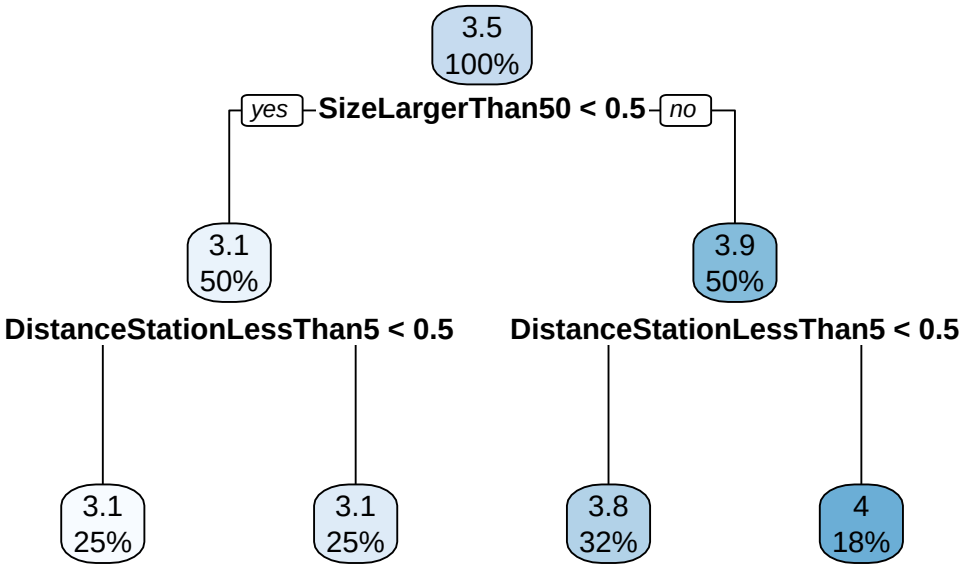
伝統的アプローチ

1. 研究者が事前に A_j を定義
2. 各 A_j について、サンプル平均を計算し、予測モデルを構築

実例: 伝統的アプローチ



実例: 伝統的アプローチ



伝統的アプローチの問題点

- 予測研究においては、サブグループを定義する際の、**Practical** guide line が限られている
  - 比較・因果研究であれば、研究課題により自動的に決まる部分がある (例: 大卒高卒間賃金格差 = 少なくとも大卒/高卒でグループ分け)
- 予測結果は、グループの定義に決定的な影響を受ける

Data adaptive アプローチ

0. 停止条件を設定
1. データに適合するように、A_j を設定
2. 各 A について、サンプル平均を計算し、予測モデルを構築

- 課題: データに適合?、具体的には？

Squared error

- “不適合度” を図る代表的指標

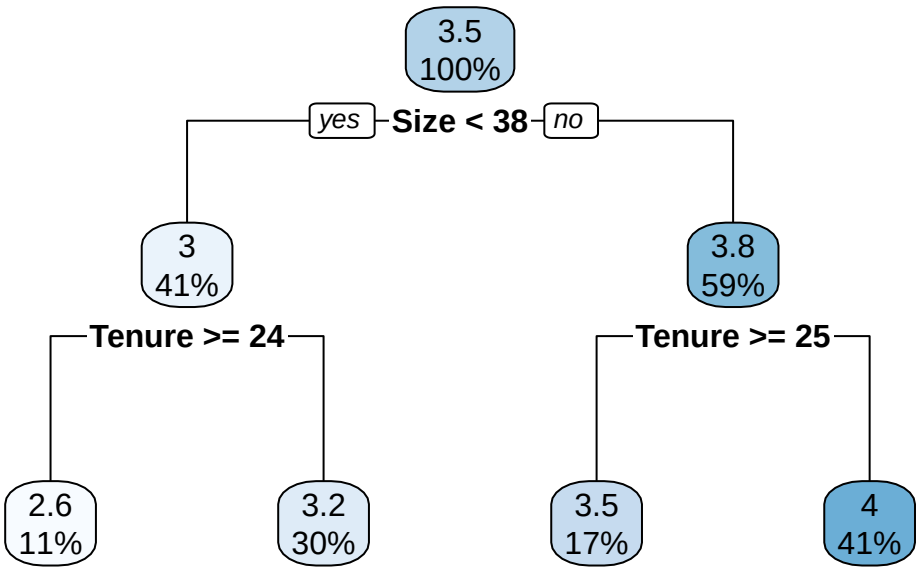
$$E[(Y_i - g(X))^2]$$

- 教師付き学習においても人気

Recursive アルゴリズム

0. 停止条件 (最大分割回数、最小サンプルサイズなど) を設定
1. 2 分割する: データ内二乗誤差を最小化するように一つの変数、閾値を選ぶ
2. 1 度目の分割を” 所与” として、 2 度目の分割を行う
3. 停止条件に達するまで、繰り返す

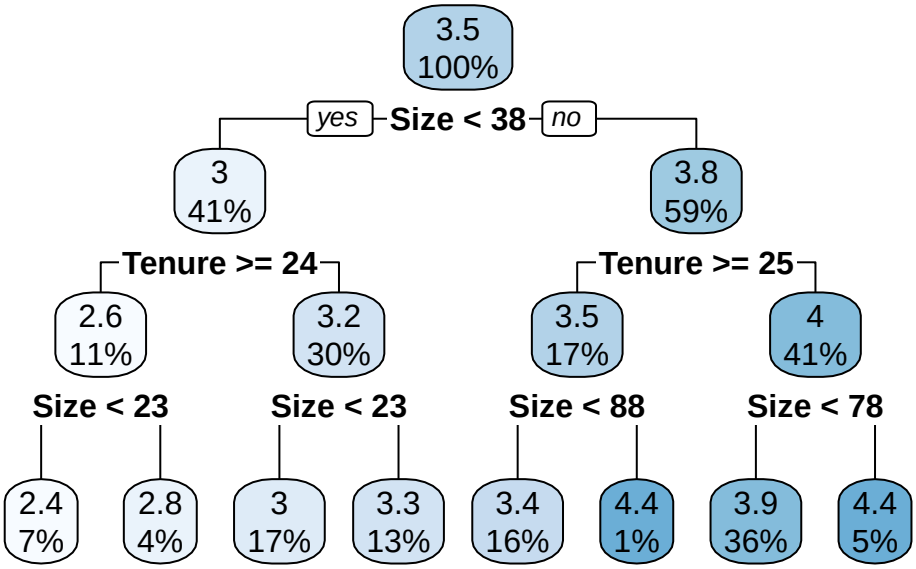
実例: 停止条件 $= 2$ 回分割



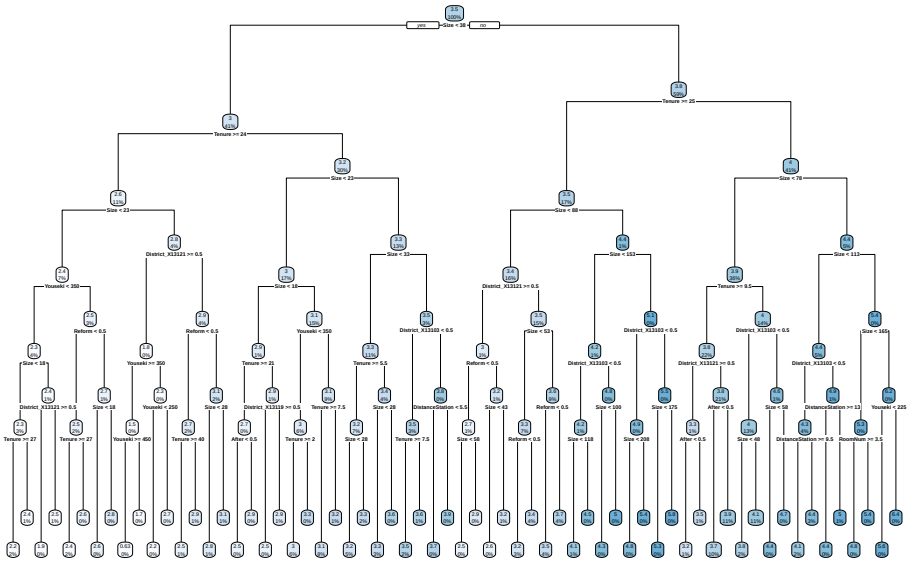
Data adaptive アプローチの課題

- モデルが停止条件に決定的な影響を受ける
- 停止条件を緩める (最大分割回数を増やす, 最小サンプルサイズを減らす) と巨大な (複雑な) 決定木が生成される

実例: 停止条件: 3 回分割



実例: 停止条件: 6 回分割



Data adaptive アプローチの課題

- “停止条件をどのように決める?”
- 異なる条件のもとで、モデルの試作と中間評価を繰り返し、最善の条件を探す
  - 独立して抽出されたデータへの当てはまり
  - 理論的評価指標 (AIC など)
  - データへの当てはまり

サンプル分割法

0. 検証する停止条件群を決める
1. データをモデル試作用データ (Training データ) と中間評価用データ (Validation データ) にランダム 2 分割 (8:2 など)
2. ある停止条件について、Training データ**のみ**を用いて、$g(X)$ を構築
3. Validation データに当てはめ、予測値を獲得し、二乗誤差を推定
4. 2-3 を繰り返し、最も二乗誤差が小さくなる停止条件を探索

例

```
# A tibble: 6 x 3
      X      Y Type
  <int>  <dbl> <chr>
1     9  35.3  Training
2     4 -18.8  Training
3     7   2.43 Training
4     1  -2.89 Training
5     2   5.88 Validation
6     7  69.1  Validation
```

例

```
# A tibble: 6 x 5
      X      Y Type         Mean TreeDepth1
  <int>  <dbl> <chr>       <dbl>      <dbl>
```

```
1     9  35.3  Training    4.01      35.3
2     4 -18.8  Training    4.01     -10.8
3     7   2.43 Training    4.01       2.43
4     1  -2.89 Training    4.01     -10.8
5     2   5.88 Validation  4.01     -10.8
6     7  69.1  Validation  4.01       2.43
```

例

```
# A tibble: 6 x 7
      X      Y Type        Mean TreeDepth1 MeanMSE TreeMSE
  <int>  <dbl> <chr>      <dbl>      <dbl>   <dbl>   <dbl>
1     9  35.3  Training    4.01      35.3      979       0
2     4 -18.8  Training    4.01     -10.8      520      63
3     7   2.43 Training    4.01       2.43       2       0
4     1  -2.89 Training    4.01     -10.8       48      63
5     2   5.88 Validation  4.01     -10.8        4     280
6     7  69.1  Validation  4.01       2.43    4236    4444
```

例

- Validate データ
  - Mean: 2120
  - Tree: 2362
- Training データ
  - Mean: 387
  - Tree: 32

まとめ

- モデルの複雑さを決めることは難しい
- 背景情報や理論、“現実をよく見て” 決める?
  - 一般に複雑なので、常により巨大な決定木を支持
- データへの当てはまりで決める?
  - 注意しないと、常により巨大な決定木を支持

– データ外も記述できる理論的枠組みを用いて、問題構造を理解する必要がある

予測問題

問題設定

- 母集団 $f_P(Y, X)$ よりランダムに抽出したデータ $\{X_i, Y_i\}_{i=1,..,N}$ を用いて、同じ母集団から**新たに**ランダム抽出する事例を予測するモデル $g(X)$ を構築
- Population risk を減らす

Population risk minimization

$$\min_{g(X)\in\mathbb{G}} E_P[L(Y, g(X))]$$

- $\mathbb{G}$：関数の集合
- $L(Y, g(X))$ = Loss function (研究者が指定)
  - 以下 $L(Y, g(X)) = (Y - g(X))^2$ と定式化

Decomposition on Population Risk

$$E_p[(Y - g(X))^2] = \underbrace{E_p[(Y - E_P[Y|X])^2]}_{Irreducible}$$

$$+ \underbrace{E_p[(E_P[Y|X] - g(X))^2]}_{Reducible}$$

- Irreducible: X が決まった時点で、**どうしようもない**
  - データから観察できない個人差がある以上
- Reducible: (古典的な) 推定問題

Decomposition on Reducible term

$$E_P[Y|X] - g(X)$$

$$= \underbrace{E_P[Y|X] - g_\infty(X)}_{Approximation\ \ Error}$$

$$+ \underbrace{g_\infty(X) - g(X)}_{Estimation\ Error}$$

- $g_\infty(X)$: 無限大のサンプルサイズで推定した予測モデル

推定問題

- 一般に

$$Y_i = E_P[Y|X_i] + \underbrace{u_i}_{Y_i - E_P[Y|X_i]}$$

- $\{Y_i, X_i\}$ を観察したとしても、u_i から $E_P[Y|X_i]$ を区別できない
  - Estimation error の源泉
  - 評価にも悪影響

Empirical Risk

- 理想的な推定方法は、Population Risk を直接最小化する
- できないのでどうするか?
  - データ上でのリスクを最小化する

Empirical Risk Minimizaiton

- 実現可能な大体案は、データ上の Risk (Empirical Risk) を最小化する

$$\min_{g(X)\in\mathbb{G}} E[(Y - f(X))^2] \coloneqq \sum_i (Y_i - f(X_i))^2/N$$

- OLS (古典的なサブグループ分析も含む) は、$\mathbb{G}$ をかなり研究者が制約した元で、Empirical Risk Mimization の解としてパラメタを推定している。

例: 古典的アプローチ

- A_j を事前に設定し、サブサンプル平均としてモデルを推定
- 以下と同値

$$\min_{g(X)} E[(Y_i - g(X_i))^2]$$

$$g(X_i) = \beta_1 \times \underbrace{I(X_i \in A_1)}_{Indicator} + .. + \beta_L \times I(X_i \in A_L)$$

- OLS と同じ!!!

性質

- 一般に $Y_i = E_P[Y_i|X_i \in A_j] + \underbrace{u_i}_{Y_i - E_P[Y_i|X_i \in A_j]}$
  - u_i の分布 = データによって異なる
  - 誤差が存在しない/誤差とそれ以外を区別できるのであれば、巨大な決定木が最善
- 経済学の応用では" 常に" 個人差が残る (X は不十分)
  - 「細かくサブグループを作ることで u_i を消去する」を" 諦める"

大標本性質: 一致性

- IID なので、

$$\lim_{n\to\infty} \sum_{i|X_i \in A_j} \frac{Y_i}{N_{A_j}} = E_P[Y_i|X_i \in A_j] + \underbrace{\sum_{i|X_i \in A_j} \frac{u_i}{N_{A_j}}}_{\to 0}$$

- N_{A_j} が小さいと、u_i の (データ上での) 分布の影響を強く受ける

例: Data adaptive アプローチ

- A_j 内のサブグループ平均として予測値を推定
  - EmpiricalRisk 最小化
- **上記を所与として、A_j も** EmpiricalRisk を最小化するように決定

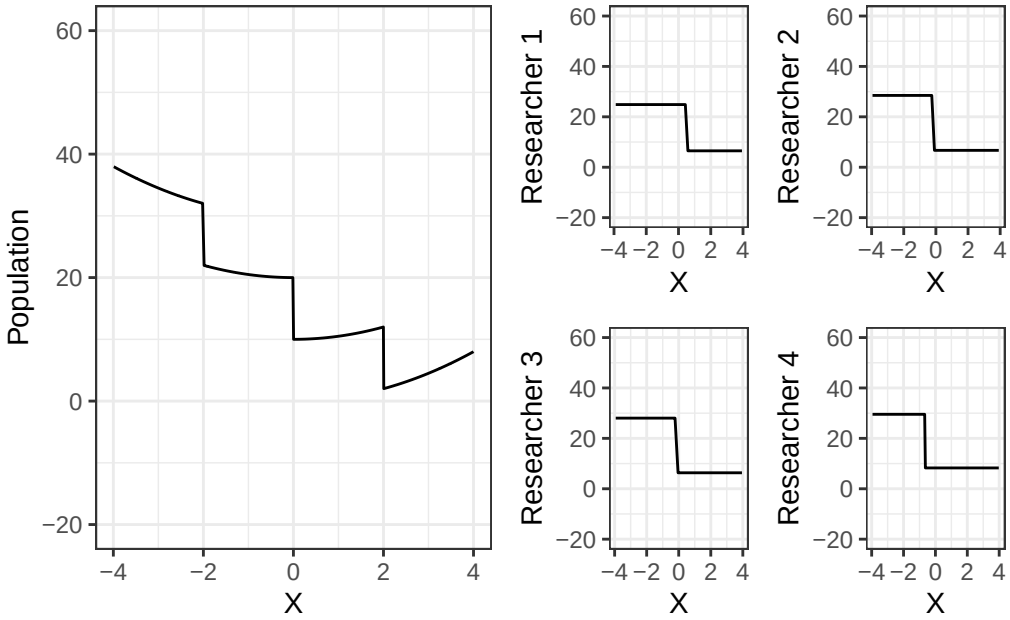
トレードオフ

- A_j を細かくすれば、
- $E_\infty[Y_i|X_i \in A_j] \to E_P[Y|X_i]$
  - Approximation error の縮小
- $E_\infty[Y_i|X_i \in A_j]$ と $g(X_i)$ のギャップ拡大

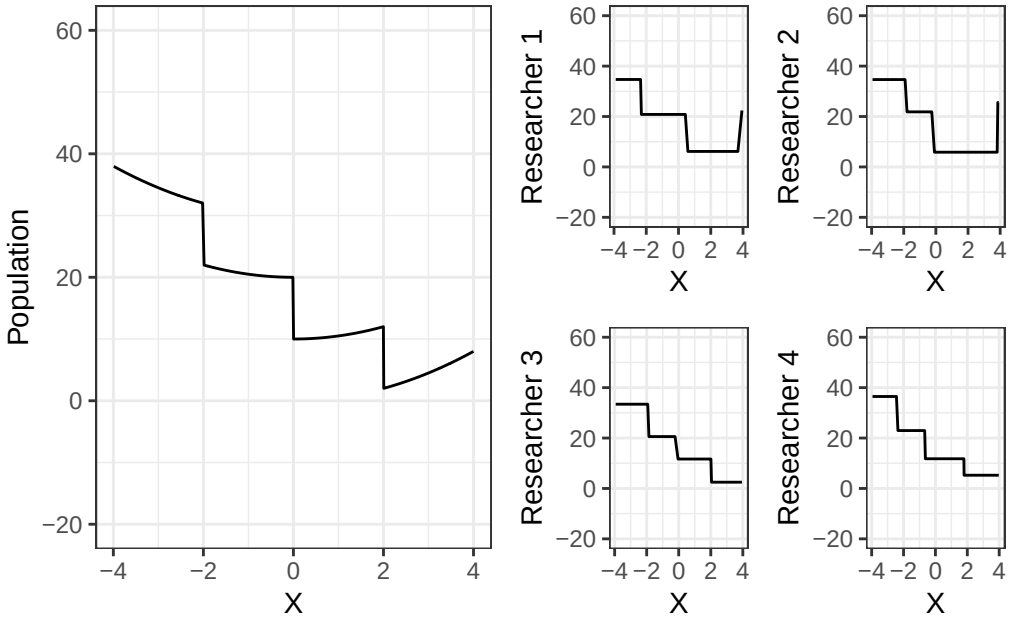
– Estimation error の拡大

– データ依存

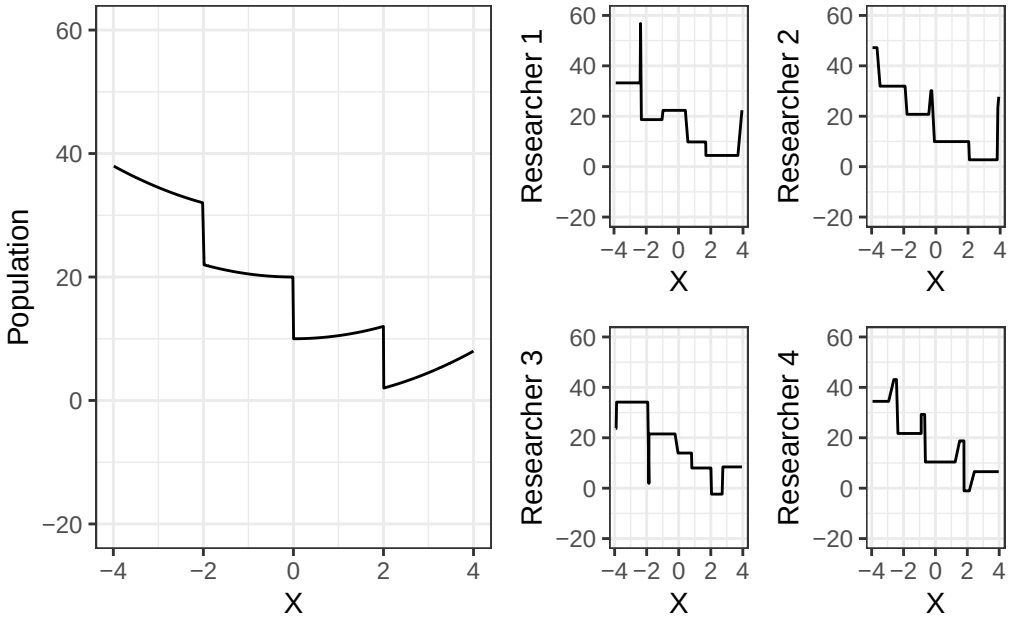
数値例: 1 回分割



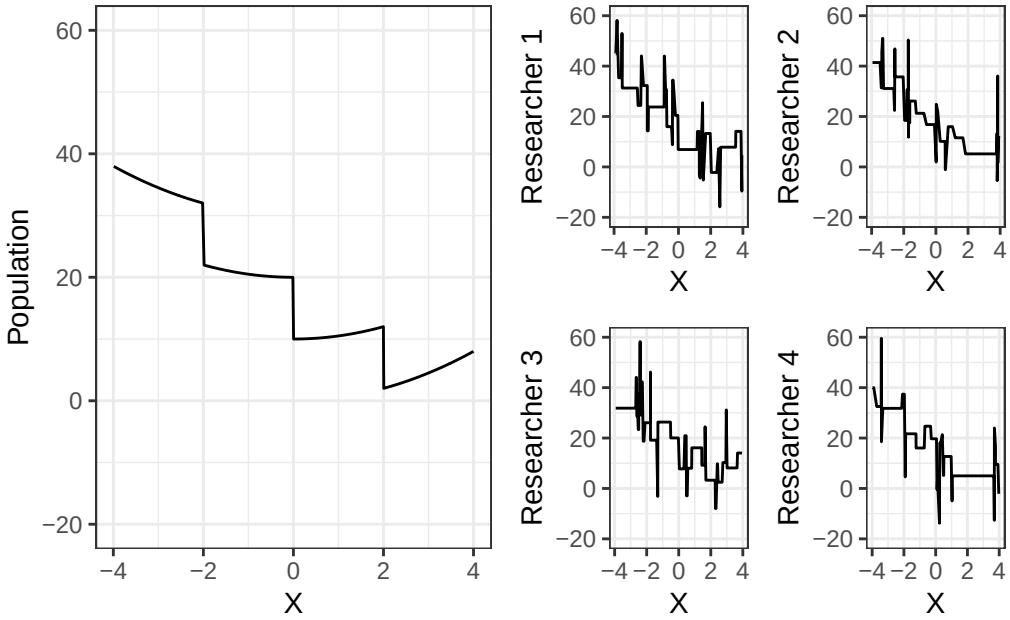
数値例：２回分割



数値例：３回分割



数値例: ５回分割



モデルの評価

- 観察できない個人差 u_i の存在のために、Empirical Risk は**使えない**

理想の評価

- 新しく独立した事例を大量にサンプルし、モデルを評価
- 予測問題: “新しい” 事例について予測したいので、新しいデータで評価するのは自然
- 母平均関数への Fitting: 最善の予測モデル = 母平均との二乗誤差最小化なので、予測にうまくいくモデル = 母平均をよりよく捉えるモデル
  - !? ← 後述

性質

- ランダムサンプリングであれば、u_i は独立無相関

$$E_P[(Y_i - g(X_i))^2] = E_P[(E_P[Y|X] + \underbrace{u_i - g(X_i)}_{Independent})^2]$$

$$= E_P[(E_P[Y|X] - g(X_i))^2] + E[u_i^2]$$

- 完璧なモデルでも、$E_P[(Y_i - g(X_i))^2] = E[u_i^2] > 0$

データ" ランダム" 分割法

- ランダムに分割すれば、母集団から” 独立に抽出された” と見做せる二つのデータを作り出せる
- 理想的な評価法を近似: 無限大のデータで評価できているわけではないが、
  - u_i の分布が独立しているデータで評価できている

同一データで評価

$$E[(Y_i - g(X_i))^2] = E[(E_P[Y|X] + \underbrace{u_i - g(X_i)}_{Dependent})^2]$$

- u_i の影響を**強く受けた** (Estimation error が大きい) 予測モデル の方が高評価されてしまう!!!
  - **過剰適合** を悪用すれば、“完璧” にデータに合うモデルができる

過剰適合/過学習問題

- Empirical Risk Minimization を突き進めると、Estimation error が爆発し、$E_P[Y|X]$ からかけ離れたモデルが生成されてしまう
  - 過剰にデータに適合した (学びすぎた) モデル

例: 丸暗記モデル

- Learning by memorization
- 「最も X の値が近い事例を予測値とする」
  - 極めて深い決定木で生成可能
- X の組み合わせが十分に多いと、1 事例しかないサブグループを生成できる
  - $g(X_i) = Y_i$ であり、**データに**完璧に適合する
  - 一般に予測性能は極めて悪い

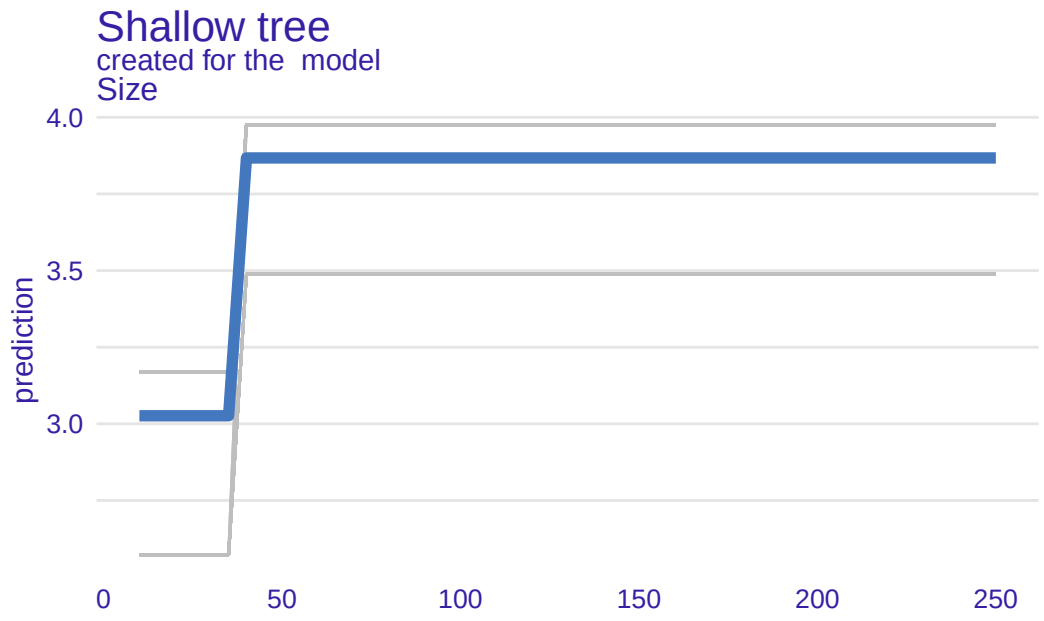
直感

- $Y_i = E_P[Y|X_i]$ であれば問題ないが、
  - 多くの応用で、観察できない要因による上振れ・下振れが生じる
- 例: 一卵性の双子
  - 大数の法則を用いた、観察できない要因の影響緩和が必須
- 丸暗記が有効なケース: 観察不可能な要因の影響を排除しているケース
  - パソコンの挙動理解? 判決予測?

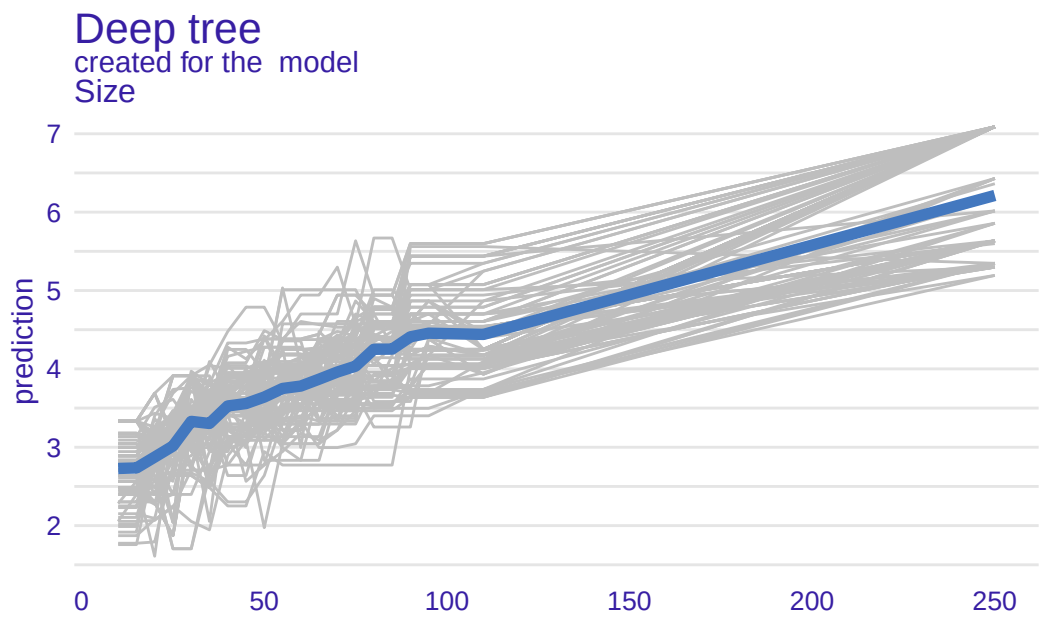
実例

- 取引年, 取引, 価格を予測するモデルを、最大 2 回、30 回分割する Recursive アルゴリズムで構築

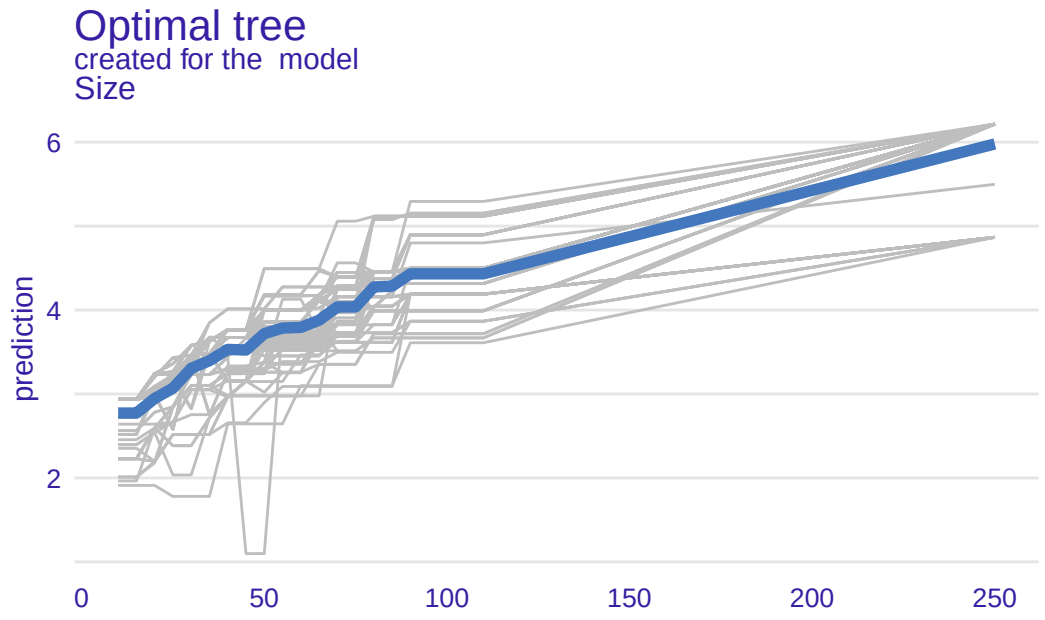
実例: Shallow Tree on Prie



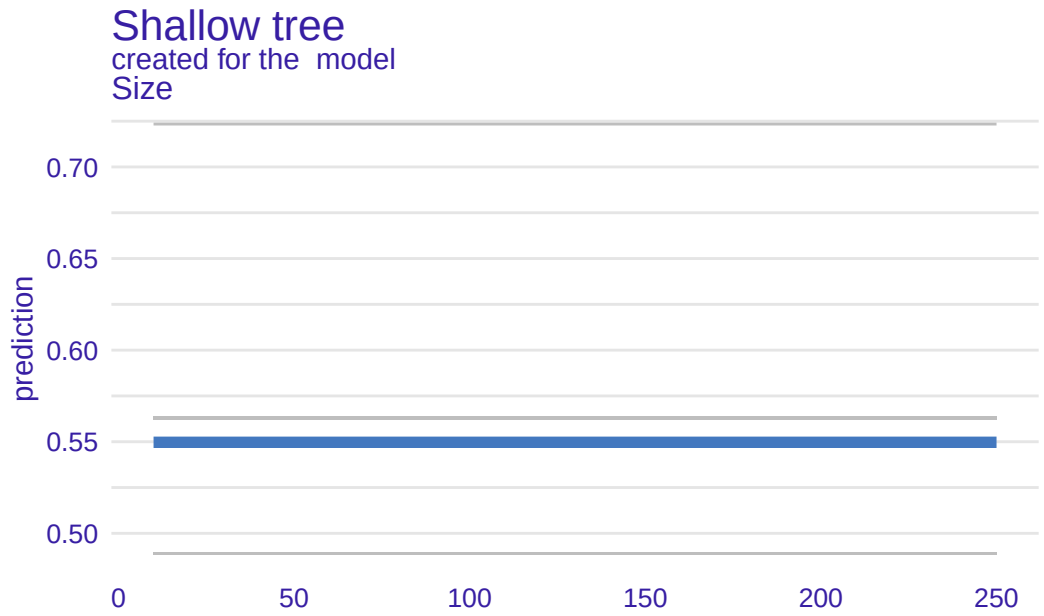
実例: Deep Tree on Prie



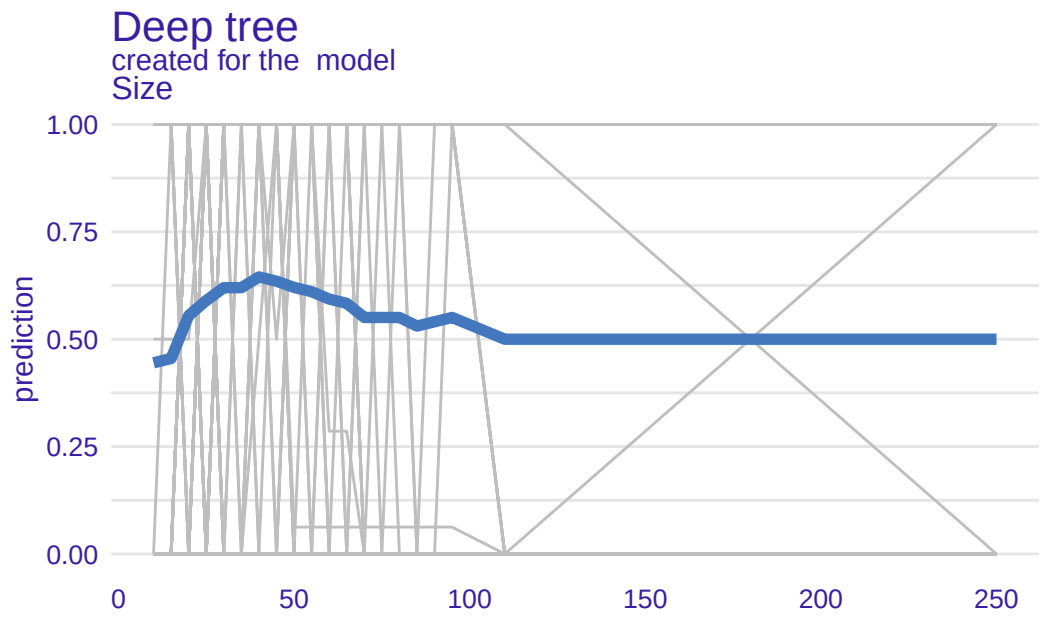
実例: Optimal Tree on Prie



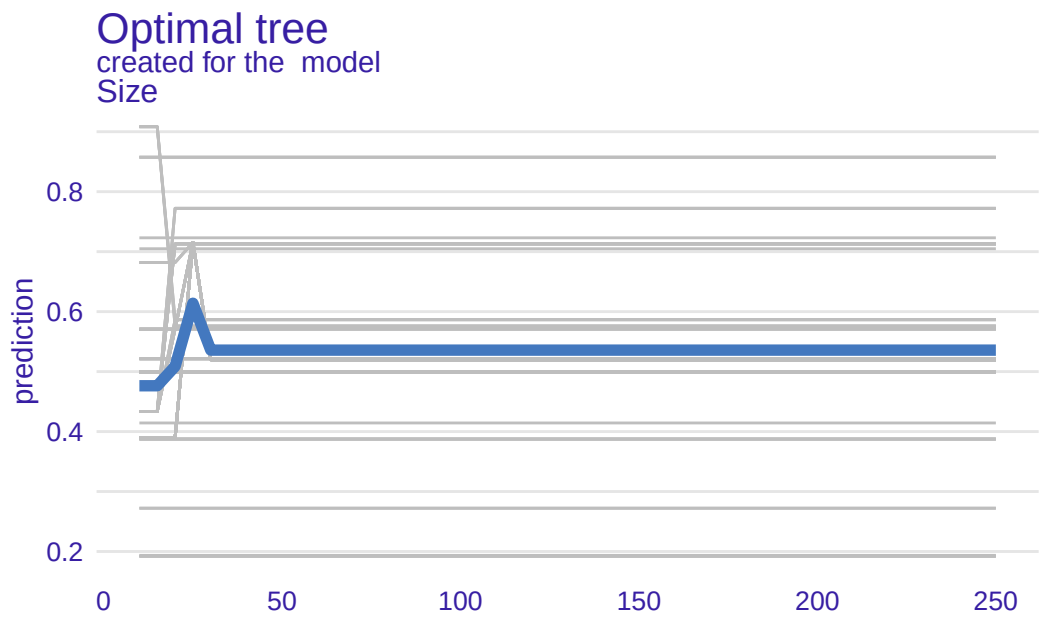
実例: Shallow Tree on Period



実例: Deep Tree on Period



実例: Optimal Tree on Period



評価

nr	task_id	learner_id	resampling_id	iters	regr.rsq
1: 1	Price	DeepTree	holdout	1	0.69722453
2: 2	Price	ShallowTree	holdout	1	0.56028579
3: 3	Price	Optimized Tree	holdout	1	0.78750222
4: 4	Period	DeepTree	holdout	1	-0.23105951
5: 5	Period	ShallowTree	holdout	1	0.01200626
6: 6	Period	Optimized Tree	holdout	1	0.07184169

Hidden columns: resample_result