カステラ本10章（10.9〜10.14）

※参考：

- 第10章後半「ブースティングと加法的木」
- GBDTの仕組みと手順を図と具体例で直感的に理解する
- Friedman, 1999

10.9 ブースティング木

（前提）木のモデル

- 重複のない木の終端頂点で表現される領域 R_j
- それぞれの領域に割り当てられる定数 γ_j
- 予測則： $x \in R_j \rightarrow f(x) = \gamma_j$

$$T(x;\Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j) \quad (10.25)$$

- パラメータの探索

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{x \in R} L(y_i, \gamma_j) \quad (10.26)$$

ブースティング木のモデル

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m) \quad (10.28)$$

- 前向き加法的モデリングで導出できる。

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \Theta_m)) \quad (10.29)$$

10.10 勾配ブースティングによる数値最適化

- 任意の微分可能な損失基準を用いてブースティング木を解くための高速な近似アルゴリズムは、数値最適化からの類推で導出できる。

(導入)

- 学習データ上の y の予測に $f(x)$ を用いることによる損失

$$L(f) = \sum_{i=1}^{N} L(y_i, f(x_i)) \qquad (10.33)$$

- これを f に関して最小化することを考える

$$\hat{f} = \arg\min_f L(f) \qquad (10.34)$$

- 数値最適化の手順：式(10.34)を、各増分ベクトル h_m （ステップ）の和として解く。

$$f_M = \sum_{m=0}^{M} \mathbf{h_m}, \qquad \mathbf{h_m} \in \mathbb{R}^N$$

10.10.1 最急降下法

- 貪欲的な探索法
- 増分：$h_m = -\rho_m \mathbf{g}_m$
- **ステップの長さ**：

$$\rho_m = \arg\min_\rho L(\mathbf{f_{m-1}} - \rho \mathbf{g_m}) \qquad (10.36)$$

- $\mathbf{f} = \mathbf{f_{m-1}}$ で評価された**勾配**：

$$g_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)} \qquad (10.35)$$

- これらをもとに、

$$\mathbf{f_m} = \mathbf{f_{m-1}} - \rho_m \mathbf{g_m}$$

での更新を繰り返す。

10.10.2 勾配ブースティング

- 前向き段階的ブースティングとの類似性
  - 現在のモデル f_{m-1} とその当てはめ $f_{m-1}(x_i)$ に対し、
    $\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \Theta_m))$ (10.29)
    を最大限減少させる木を構築
  - 木の予測 $T(x_i; \Theta_m)$ は、負の勾配要素（10.35）に対応

最急降下法のデメリット

- ロバストな規準（指数損失に対する逸脱度、二乗誤差損失に対する絶対値誤差損失）が微分不可能で、勾配の計算が難しくなる。
- (10.35)で定義される勾配は正解ラベル y ありきで計算されるため、テストデータでは y が存在せず勾配を定義できない。

対処策

- 木を負の勾配に当てはめる（＝訓練データから負の勾配を予測する木を構築する）

損失関数と勾配

問題設定	損失関数	勾配
回帰	$\frac{1}{2}[y_i - f(x_i)]^2$	$y_i - f(x_i)$
回帰	$\lvert y_i - f(x_i)\rvert$	$sign[y_i - f(x_i)]$
回帰	フーバー	$\lvert y_i - f(x_i)\rvert \leq \delta_m$ に対しては $y_i - f(x_i)$, $\lvert y_i - f(x_i)\rvert > \delta_m$ に対しては $\delta_m sign[y_i - f(x_i)]$

勾配ブースティングのアルゴリズム（10.3）

1. 初期化：$f_0(x) = \arg\min_\gamma \sum_{i=1}^N L(y_i, \gamma)$
2. $m = 1$ から M に対して以下を行う。
   (a) $i = 1, 2, ..., N$ に対し、
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$
   (b) 終端領域 $R_{jm} \quad (j = 1, 2, ...J_m)$ を与える回帰木を目標 r_{im} に適合させる。
   (c) $j = 1, 2, ...J_m$ に対し、
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$
   (d) $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$ のように更新する。
3. $\hat{f}(x) = f_M(x)$ を出力する。

- 勾配ブースティングに埋め込まれるパラメータ
  - 構成する木の大きさ J
  - 繰り返し回数 M

10.11 ブースティングのための木の適切な大きさ

一般的な木の構築での課題と対処

- 各木の最適な大きさを一般的な方法で独立に推定
- 各木で非常に大きな木を導出し、これをボトムアップ的な手法で、
  最適な終端頂点数になるまで刈り込み
    - 各木の繰り返しの初期に構築される木が大きくなりすぎることで、
      性能の低下、計算量の増加が起こる。
- 回避策
    - 全ての木を同じ大きさに制限して（$J_m = J, \quad \forall m$）J を調節する。

最適な木の大きさ J とテスト誤差

- 実用的な J の値は、目標関数

$$\eta = \arg\min_{f} E_{XY} L(Y, f(X)) \qquad (10.39)$$

の性質を考えることにより得られる。

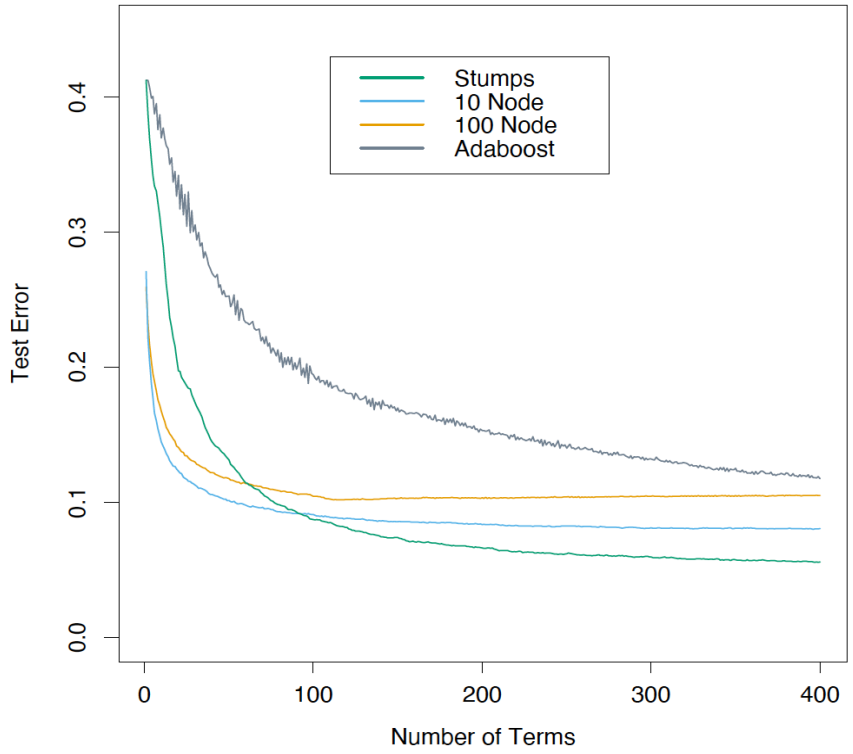
- $\eta(X)$ の性質： $X^T = (X_1, X_2, ..., X_p)$ の相互作用

$$\eta(X) = \sum_j \eta_j(X_j) + \sum_{jk} \eta_{jk}(X_j, X_k) + \sum_{jkl} \eta_{jkl}(X_j, X_k, X_l) + ... \qquad (10.40)$$

- 以下の分類におけるテスト誤差

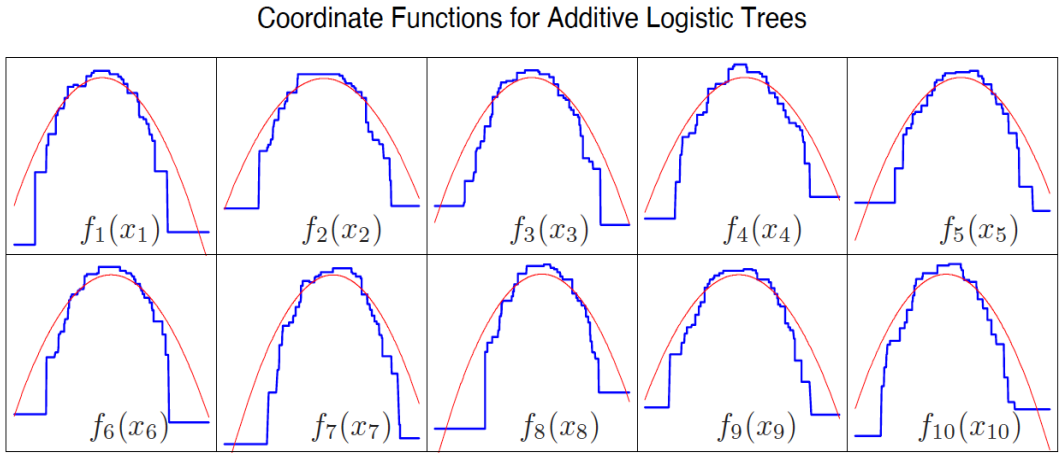
$$Y = \begin{cases} 1 & \sum_{j=1}^{10} \mathrm{X}_j^2 > \chi_{10}^2(0.5) \\ -1 & other \end{cases}$$

- $J > 10$ が必要な場合はほぼない



メタパラメータ J の調整

- $4 \leq J \leq 8$ の場合にうまくいくことが多い



10.12 正則化

導入

- 勾配ブースティングでの過学習回避による精度向上やノイズに対するロバスト性向上を目的としたアプローチを論じる。
- ブースティングの繰り返し回数 M について
  - 繰り返しを増やすほど訓練リスク $L(f_M)$ を減少させることができる。
  - 合わせすぎると過学習が生じ、未知のデータに対する予測精度が下がる。
  - 最適な M^* は未知のリスクを最小化するもの

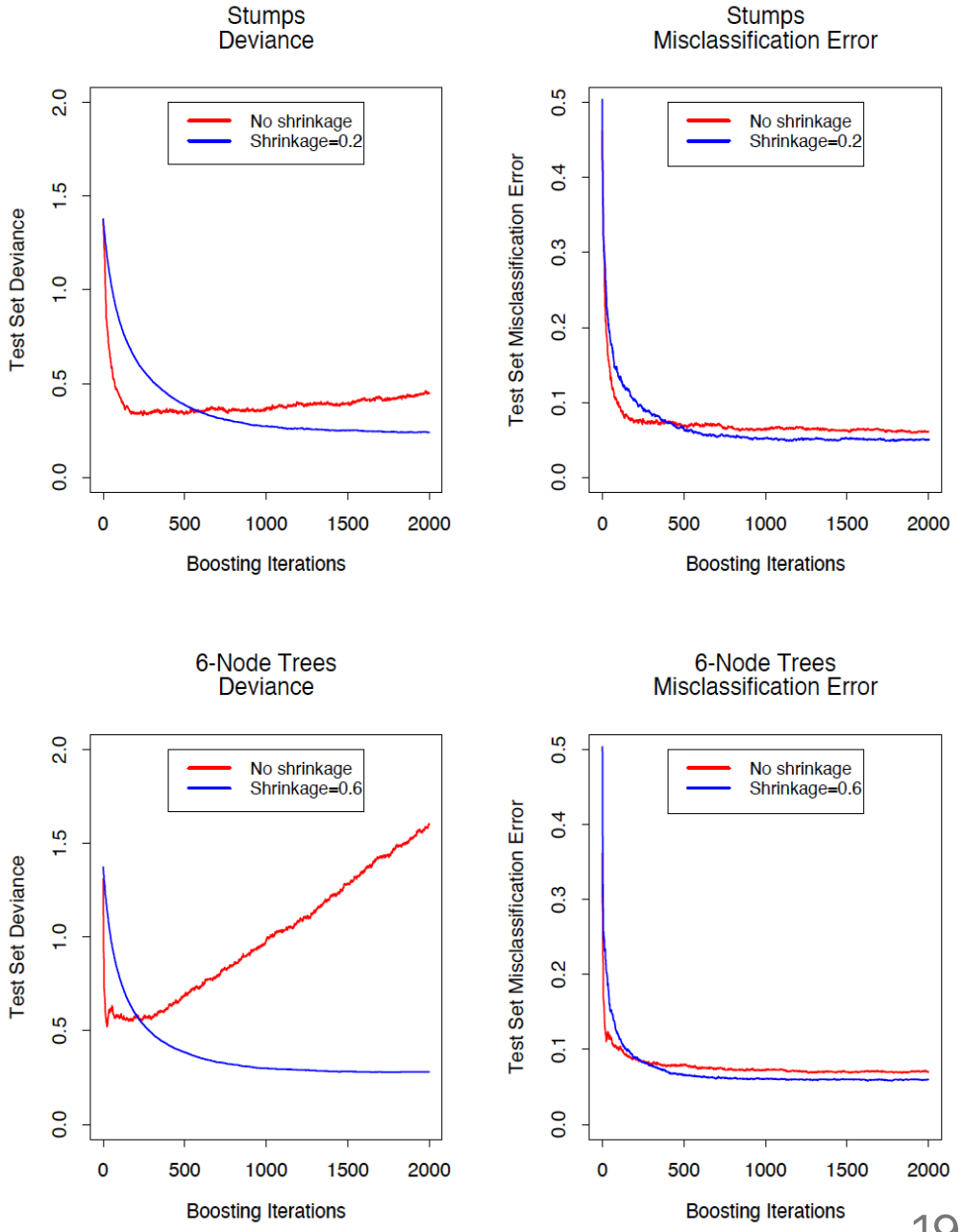
10.12.1 縮小法

- 現在の近似に新たに追加する際に、それぞれの木の貢献度を $0 < \nu < 1$ の倍率で掛け合わせる。
- つまり、アルゴリズム10.3ステップ2(d)が以下の通り書き換わる

$$f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$$

  - パラメータνにより、ブースティング手順の学習率が制御される。
  - ν と M にトレードオフ関係
    - ν の値が小さい場合、テスト誤差が小さくなるが、これに対応して大きなMが必要になる。

- 縮小法を用いると、各テスト誤差曲線はより小さい値になり、繰り返し回数が多くなっても低い値に留まっている。



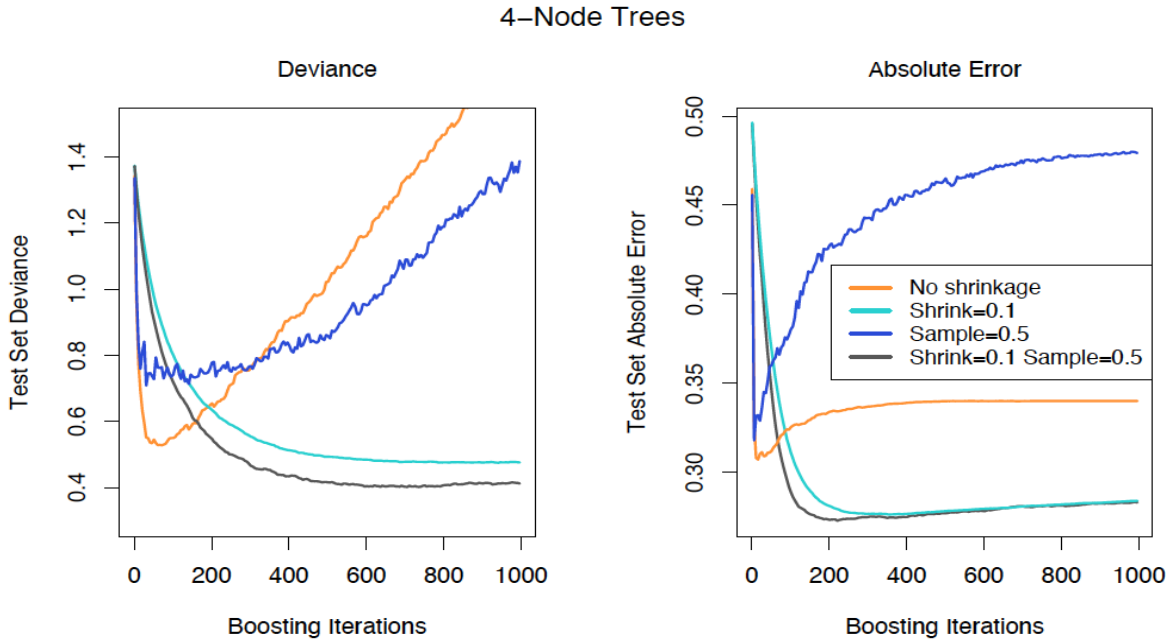
10.12.2 部分標本化

- アルゴリズム(10.3) 2(a)

$$\{\eta(i)\}_1^N = randperm\{i\}_1^N$$

$$r_{\eta(i)m} = -\left[\frac{\partial L(y_{\eta(i)} f(x_{\eta(i)}))}{\partial f(x_{\eta(i)})}\right]_{f=f_{m-1}}$$

- 繰り返しの各ステップで訓練用の観測値の割合 η を標本化（非復元抽出）
- 一般的に η はN に対して $\frac{1}{2}$ 程度
- 計算時間を η の比率分短縮でき、より正確なモデルを作れる



10.13 説明性

10.13.1 予測変数の相対的重要性

- 多くの場合、入力（予測変数）のうち一部の変数のみが応答に決定的な影響を与えている。
- したがって、応答を予測する場合に、各入力変数の相対的重要度、貢献度を学習するのが有効になる。

単独の決定木

- 予測変数 X_l の相対的重要度：

  $$\mathrm{I}_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = l) \qquad (10.42)$$

  - 変数 X_l の相対的重要度の2乗 = 全ての内部頂点上での2乗改善の和
- 加法的木展開(10.28)に一般化できる。

  $$\mathrm{I}_l^2 = \frac{1}{M} \sum_{m=1}^{M} \mathrm{I}_l^2(T_m) \qquad (10.43)$$

  - 平均をとることによる安定化効果により、信頼性が高まる。
  - 縮小法により、高い相関を持つ重要変数の別変数による隠蔽（多重共線性？）の問題もかなり少ない。

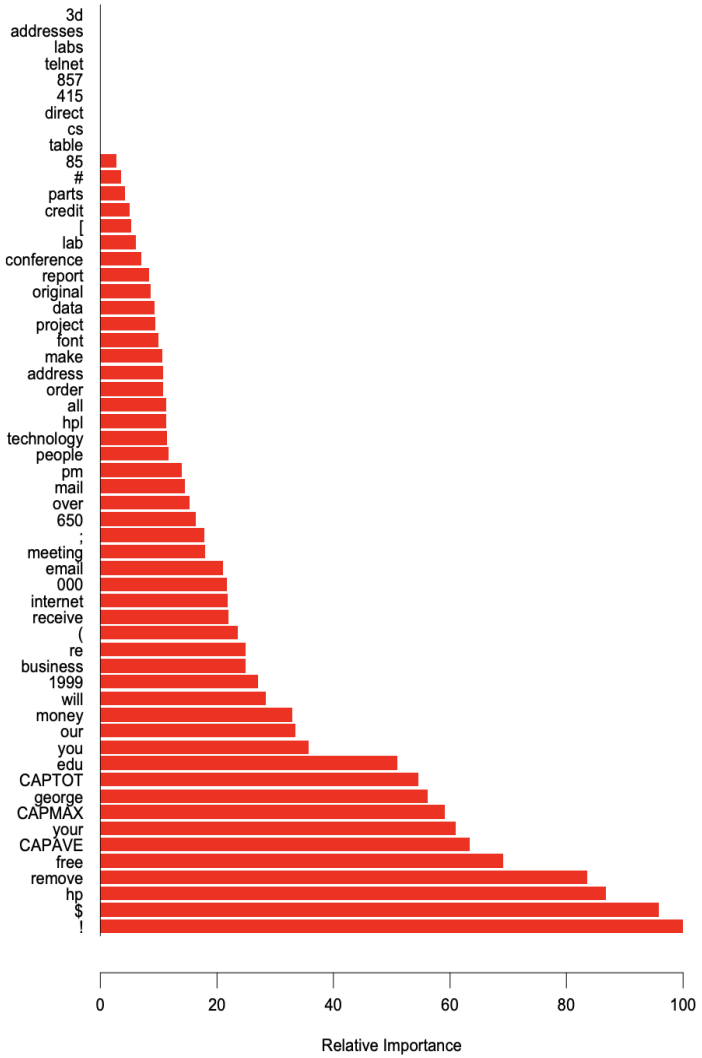


FIGURE 10.6. *Predictor variable importance spectrum for the* `spam` *data. The variable names are written on the vertical axis.*

Kクラス分類

- K分割モデル $f_k(x) \quad (k = 1, 2, ..., K)$ に対して、それぞれが木の和

$$f_k(x) = \sum_{m=1}^{M} T_{km}(x) \qquad (10.44)$$

- 式(10.43)のKクラス分類における一般化は、

$$\mathrm{I}_{lk}^2 = \frac{1}{M} \sum_{m=1}^{M} \mathrm{I}_l^2(T_{km}) \qquad (10.45)$$

※　I_{lk} はクラス k の観測を他のクラスから分割する場合の X_l の関連度を表す。

- X_l の全体の関連度は全クラスに対して平均化することで得られる。

$$\mathrm{I}_l^2 = \frac{1}{K} \sum_{k=1}^{K} \mathrm{I}_{lk}^2 \qquad (10.46)$$

10.13.2 部分依存図

- 入力変数の結合値に近似 $f(X)$ がどのように依存しているかを理解したい
- 有効な打ち手は、$f(X)$ を入力引数の関数として図示し、入力変数の結合地への依存性の全体像を俯瞰することだが、そうした図は低次元でないと得ることが難しい。
    - 入力変数の部分集合を選択し、その部分集合上での近似 $f(X)$ の部分依存性を示す一連のグラフを描画するのがいい。（特に低次元の相互作用が支配的である場合には、有益な情報が得られる可能性がある。

部分依存関数

- 部分依存関数は、任意の「ブラックボックス」の学習手法の結果を説明するのに利用できる。
- 定義
  - インデックス $S \subset \{1, 2, ..., p\}$ のついた入力予測変数 $X^T = (X_1, X_2, ..., X_p)$
  - $S \bigcup C = \{1, 2, ..., p\}$ となる補集合 C
  - 一般的な関数 $f(X)$ は全ての入力変数に依存し、 $f(X) = f(X_S, X_C)$
  - $f(X)$ の X_S への部分的な依存性であり、周辺化された平均として、

$$f_S(X_S) = E_{X_C} f(X_S, X_C) \qquad (10.47)$$

- 部分依存関数は次のように推定することができる。

$$\bar{f}_S(X_S) = \frac{1}{N}\sum_{i=1}^{N} f(X_S, x_{iC}) \qquad (10.48)$$

  - $\{x_{1C}, x_{2C}, ..., x_{NC}, \}$ は訓練データ中に現れる X_C の値で、計算コストが大きい。
  - しかし決定木を用いることで、 X_C に対してデータ参照を行わなくても良いので、高速に計算できる。

$$f(X_S, x_{iC}) = \sum_{j \in S} \gamma_j I(x_{iC} \in R_{x_j})$$

$$\rightarrow \qquad \bar{f}_S(X_S) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j \in S} \gamma_j I(x_{iC} \in R_{x_j})$$

部分依存関数の条件付き期待値の部分依存関数としての不適切性

- 部分依存関数(10.47)は、 $f(X)$ 上の変数 X_C の影響（平均）を計算した後の $f(X)$ 上の X_S の影響を示しているのであり、 X_C の影響を無視した X_S の影響ではない。
- X_S 単体の影響は、

$$\tilde{f}_S(X_S) = E(f(X_S, X_C)|X_S) \qquad (10.49)$$

で与えられ、 X_S と X_C が独立である場合にのみ $\tilde{f}_S$ と $\bar{f}_S$ が等しくなるがほとんどない。
- これは、(10.47)では以下が成り立つが、(10.49)では成り立たないからである。
  - 選択された変数の部分集合の効果が完全に加法的な場合
    - $f(X) = h_1(X_S) + h_2(X_C)$ より、 $f_S(X_S) = h_1(X_S)$
  - 選択された変数の部分集合の効果が完全に乗法的な場合
    - $f(X) = h_1(X_S) \times h_2(X_C)$ より、 $f_S(X_S) = h_1(X_S)$

Kクラス分類の部分依存関数

- 各クラスに一つずつ合計K個のモデルが存在

$$f_k(X) = \sum_{m=1}^{M} T_{km}(x), \qquad k = 1, 2, ..., K$$

- 各モデルとそれぞれの確率 $p_k(x) = \frac{e^{f_k(x)}}{\sum_{l=1}^{K} e^{f_l(x)}}$ との関係

$$f_k(X) = logp_k(X) - \frac{1}{K} \sum_{l=1}^{K} logp_l(X) \qquad (10.52)$$

- $f_k(X)$ はそれぞれの確率に対数を適用した単調増加関数
- 最も関連性が高い予測変数 I_l^2 に対する各 $f_k(X)$ の部分依存図を見ると、各クラスの対数オッズがどのように各入力変数に依存しているかを理解しやすい。

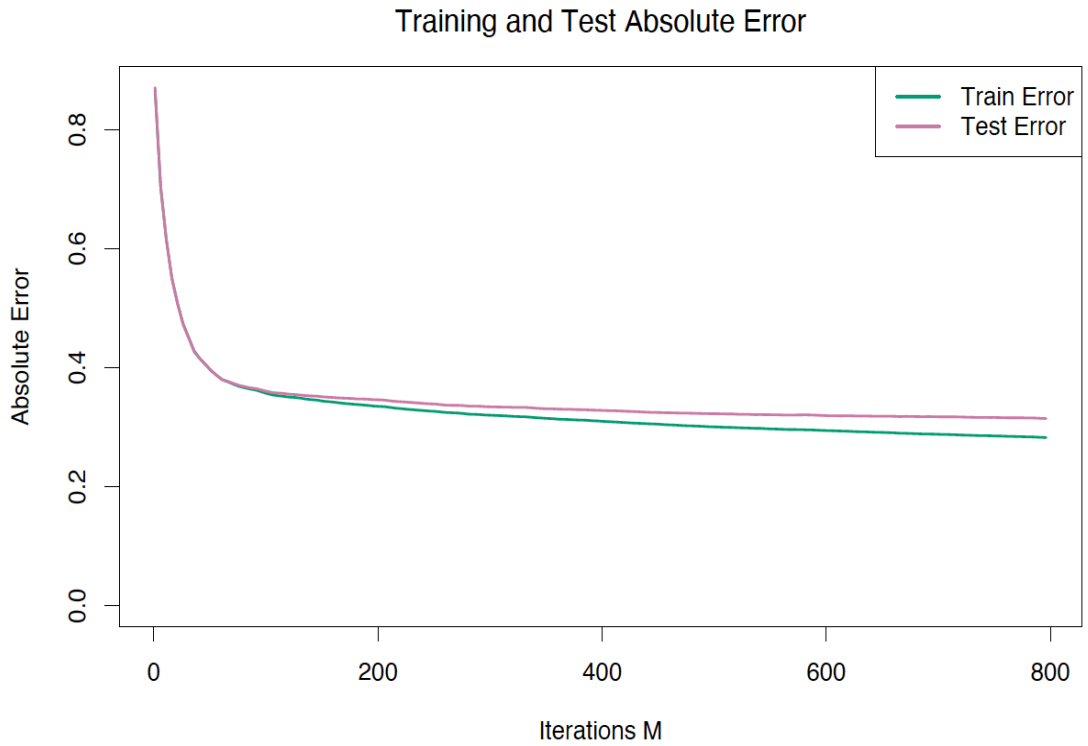
10.14 具体例

10.14.1 カリフォルニアの住宅

- 応答変数：各地区の家屋価値（10万ドル単位）の中央値
- 予測変数
  - 収入の中央値 `MedInc`
  - 家の数を反映した住宅密度 `House`
  - 各住居の平均居住率 `AveOccup`
  - 平均部屋数 `AveRooms`
  - 寝室数 `AveBedrms`
    など全8種類（全て数値型）

MARTを用いた勾配ブースティングモデルの当てはめ

- パラメータ
  - 終端頂点　$J = 6$
  - 学習率 (10.41) $\nu = 0.1$
  - 損失関数：フーバー損失
- 誤差
  - $AAE = E[y - \hat{f}_M(x)]$
- モデルの当てはまり
  - R^2：0.84

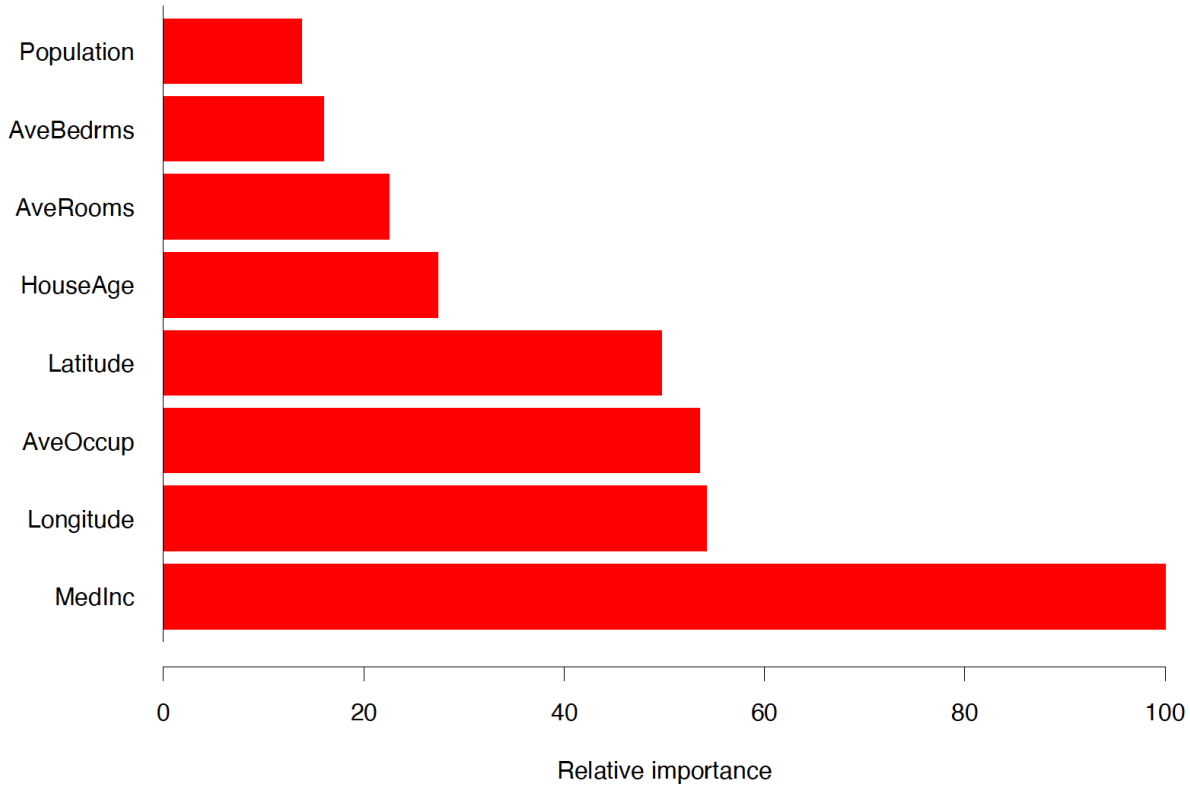


Pace and Barry(1997)におけるモデルの調整

- 以下のモデル式に調整
  $ln(Medval) = \alpha + \beta_2 MedInc + \beta_3 MedInc^2 + \beta_4 MedInc^3 + \beta_5 ln(med(Age)) + \beta_6 ln(Total.rooms/Population) + \beta_7 ln(Bedrooms/Population) + \beta_8 ln(Population/Households) + \beta_9 ln(Households)$
- 勾配ブースティングにより $R^2 = 0.86$ を実現

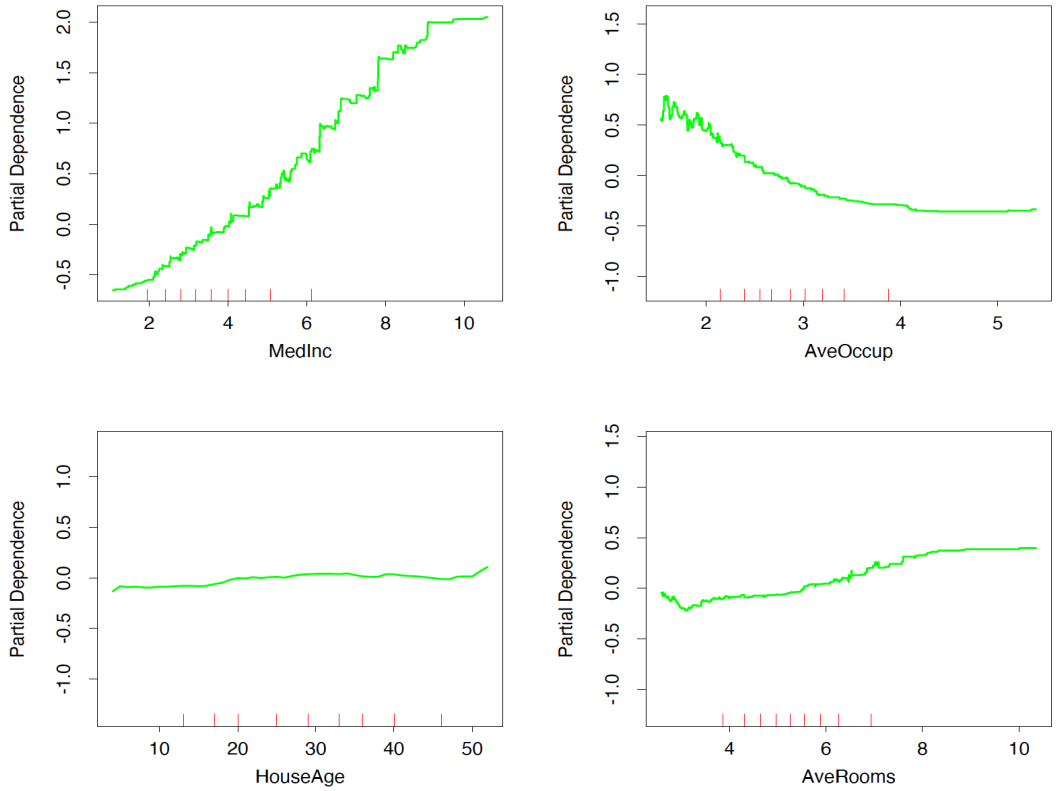
家屋価値に対する予測変数の相対重要度

- `近隣整体の収入の中央値` が最も関連が強く、`経度` ・ `緯度` ・ `平均居住率` がそれに続く。



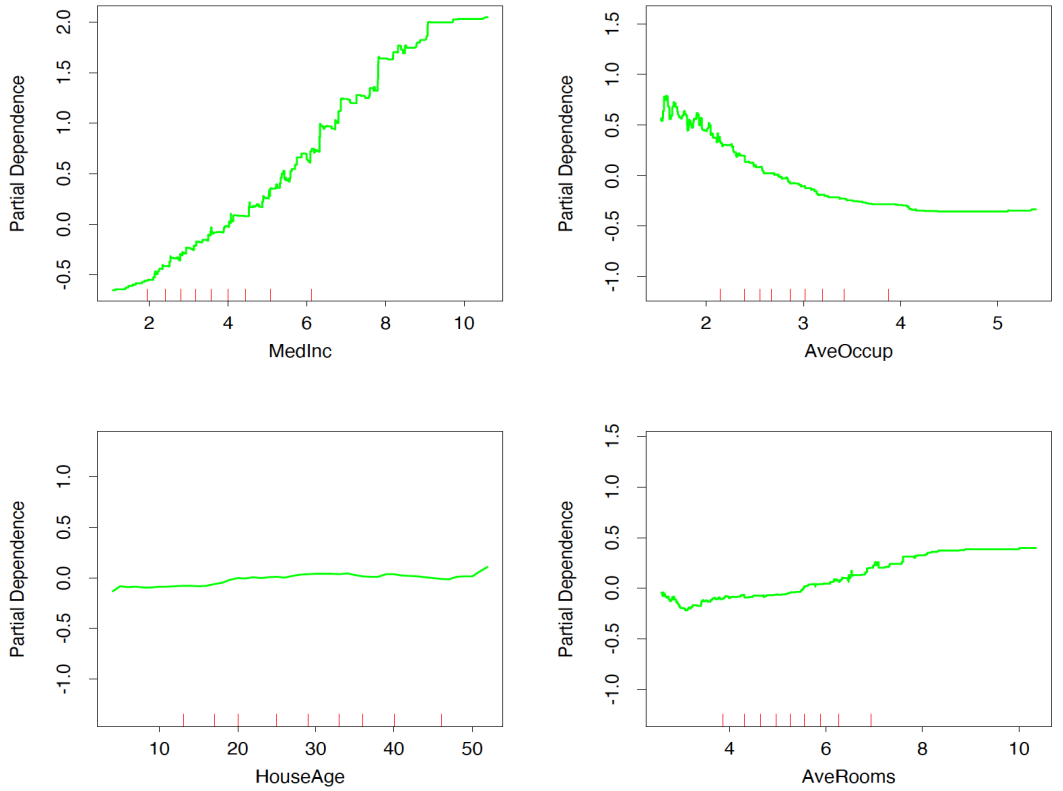
家屋価値に対する予測変数の部分依存性

- 地理的要素を含まない予測変数の中で関連度が大きい4変数をピックアップ。
- 収入の中央値 に対する 家屋価値
  → 単調増加
- 平均居住率 に対する 家屋価値
  → 概ね単調減少



家屋価値に対する予測変数の部分依存性

- 地理的要素を含まない予測変数の中で関連度が大きい4変数をピックアップ。
- 平均部屋数 に対する 家屋価値
  → およそ3部屋で最低
  それ以上でもそれ以下でも増加
- 築年数 に対する 家屋価値
  → 非常に弱い部分依存性
  （図10.14の重要度順位と矛盾）
  築年数 の弱い主効果が他の変数とのより強い相互効果を隠している可能性

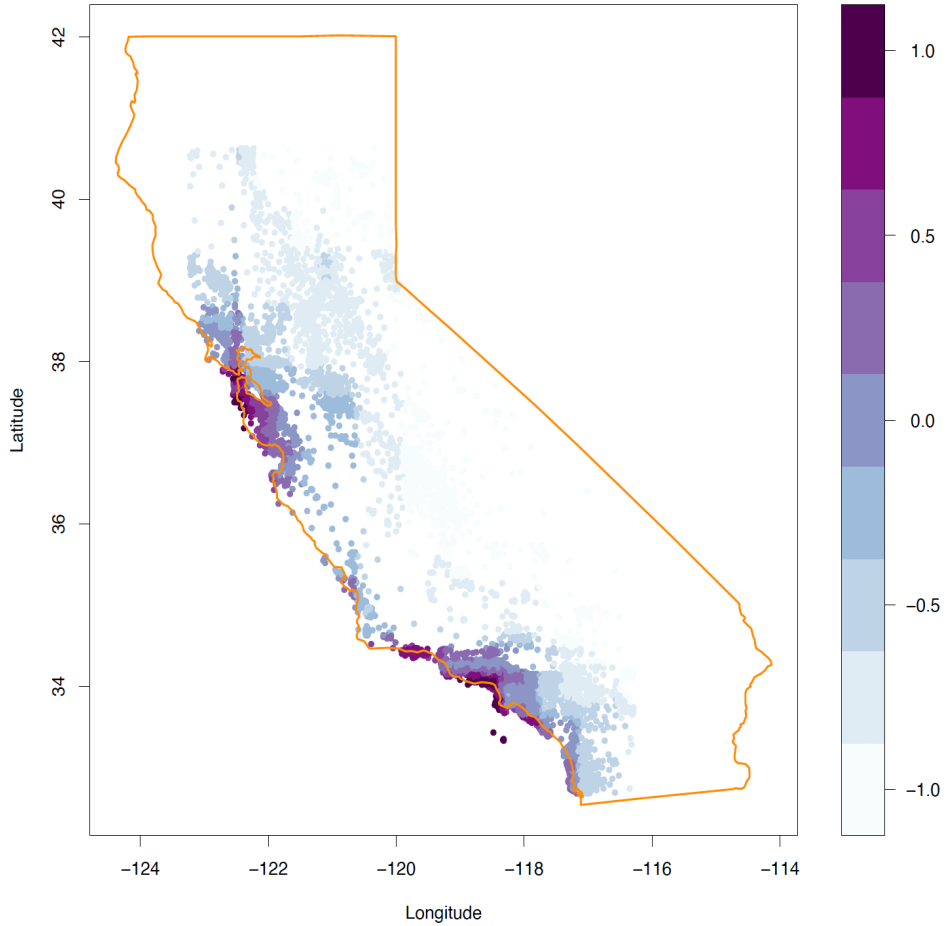


家屋価値に対する予測変数の部分依存性

- 築年数 と 平均居住率 に対する
  家屋価値 の2変数部分依存性
- 平均居住率 > 2
  家屋価値 は 築年数 に対して独立
- 平均居住率 ≤ 2
  築年数 と強い依存関係

家屋価値に対する予測変数の部分依存性

- `緯度` と `経度` に対して当てはめたモデルにおける2変数部分依存性
- `家屋価値` は明らかに地理的条件に強い依存性を持つ
- 地区や住居に関する属性の効果も考慮
    - 地理的条件に対して人々が払う**付加価値**を示している



10.14.2 ニュージーランドの魚

- クロマトウダイの存在と発生量のモデリング
- 分析の目的
  - ある漁網におけるおけるクロマトウダイの漁獲確率や漁獲高を推定する
- 予測変数
  - 漁綱の平均深度 `AvgDepth`
  - 水温/塩分濃度 `TempResid` / `SalResid`
  - 海表面の温度勾配 `SSTGrad`
  - 生態系の産出性指標 `Chla`
  - 海水内の浮遊粒子状物質 `SusPartMatter`

モデリング

- モデル式

$$E(Y|X) = E(Y|Y > 0, X) \times P_r(Y > 0|X) \qquad (10.54)$$

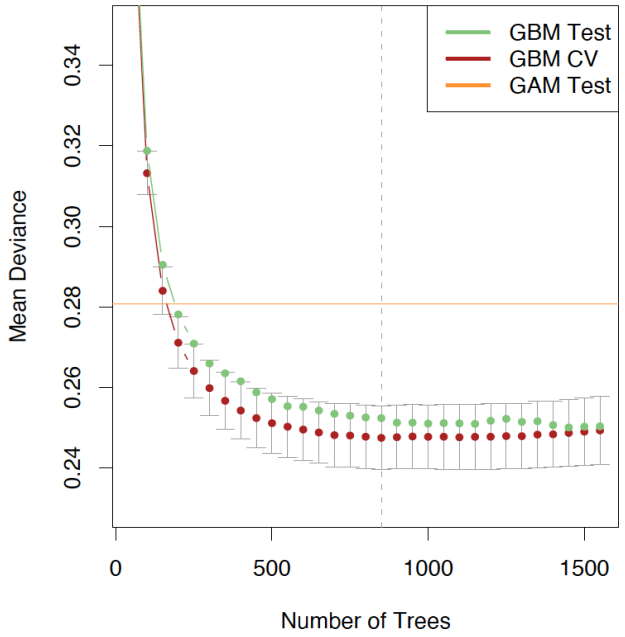
  - Y：非負の漁獲高
  - 第1項は17,000箇所の底引網のうち、クロマトウダイの漁獲があった2,553箇所を用いた推定
  - 第2項はロジスティック回帰で推定

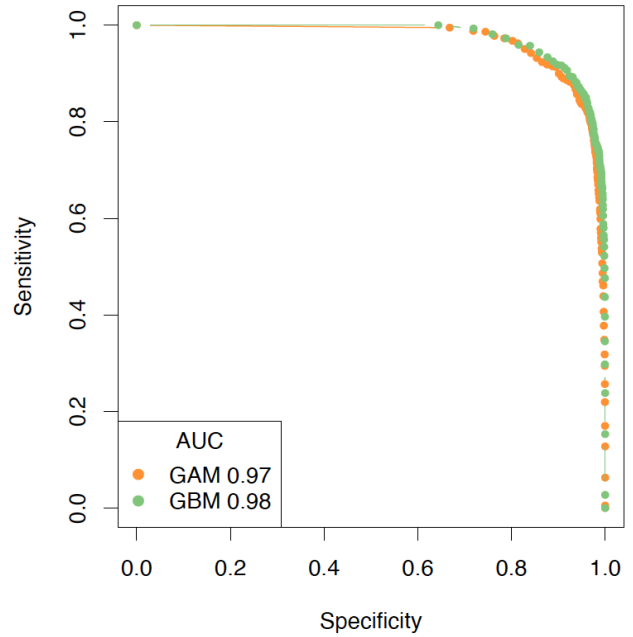
モデリング

- 第1項のGBMモデル
  - 損失関数：二項逸脱度
  - 木の深さ：10
  - 縮小率：$\nu = 0.025$
- 第2項のGBMモデル
  - 損失関数：2乗誤差損失
  - 木の深さ：10
  - 縮小率：$\nu = 10$
  - $log(Y)$ をモデリングし、予測の時に対数演算を外す
- どちらの場合にも、10分割CVで項数と縮小率を決定

モデルの当てはまり

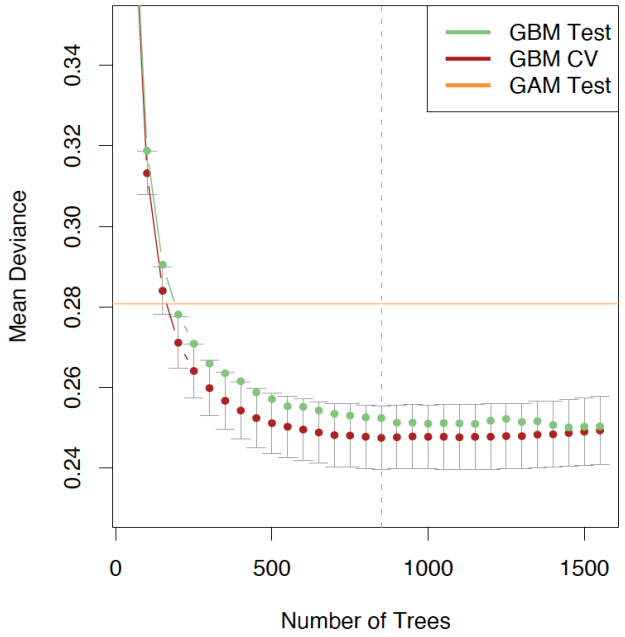
- 左側：
  - 以下3つの平均二項逸脱度
    - GBM (10 Fold CV)
    - GBM (テストデータ)
    - 各項8自由度の平滑化スプラインによるGAM
  - GAMに比べ性能が改善

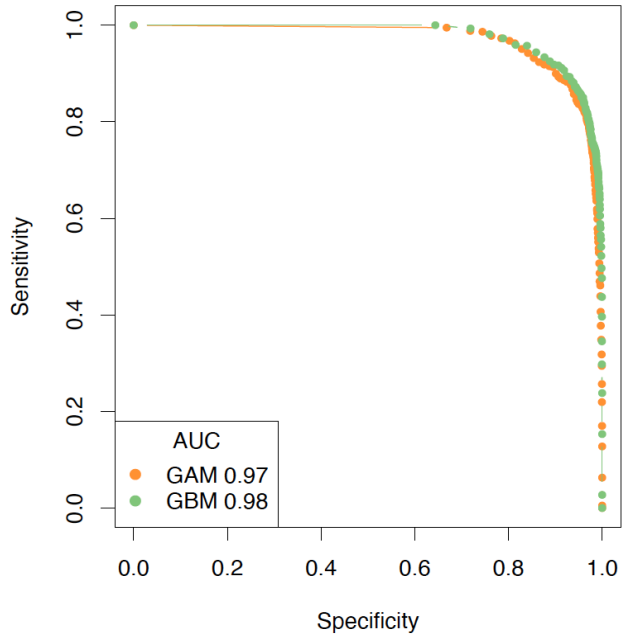




予測性能

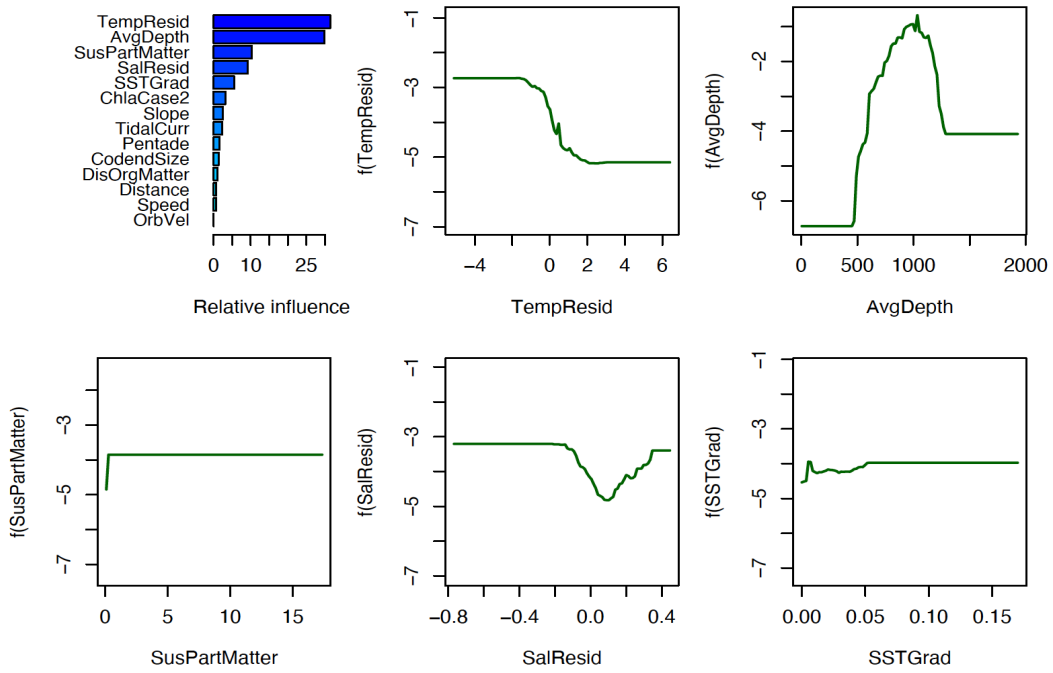
- 右側：
  - GAMとGBMのROC曲線
  - 非常によく似た性能だが、GBMがわずかに優れている。





各変数の貢献度

- ロジスティックGBM（第2項部分）における各変数の貢献度
  - クロマトウダイが捕獲される深さの範囲は明確で、冷水領域で高い頻度で捕獲される
- 捕獲高（第1項部分）でも重要な変数は同様



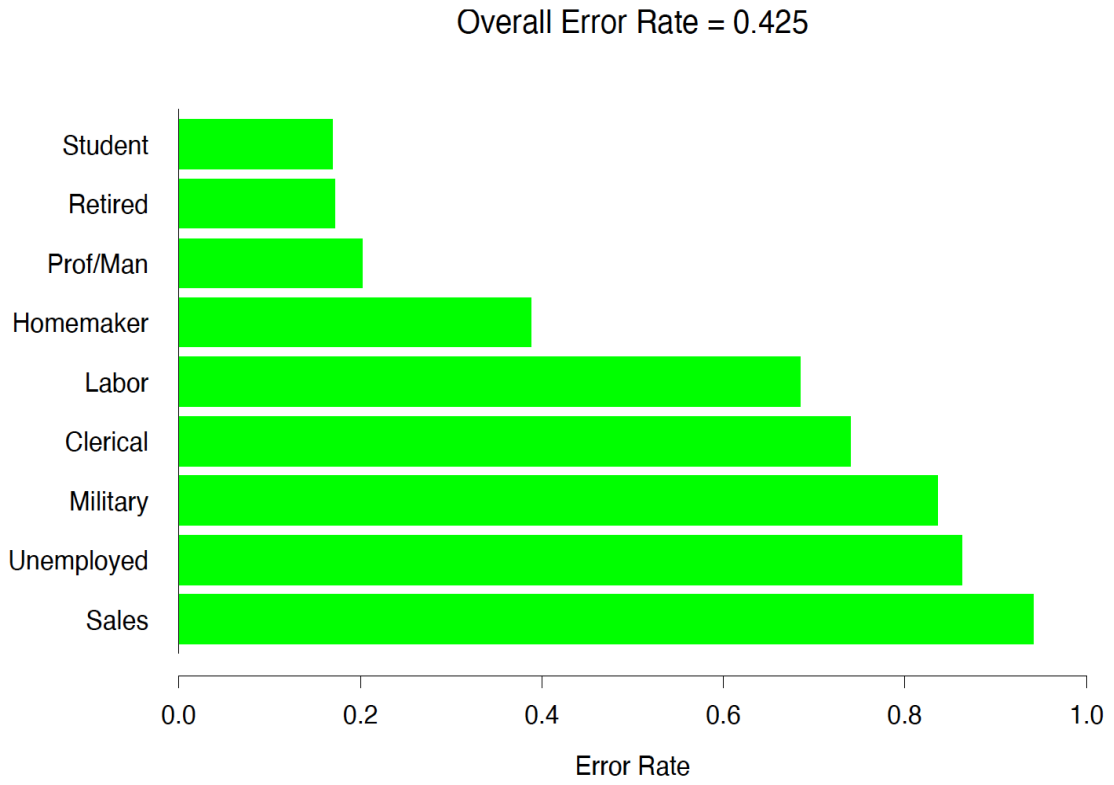
GBMが、相互作用のモデル化と変数の自動選択が可能であり、外れ値や欠損データに対するロバスト性があることで、一般的手法として普及した例

10.14.3 個人属性情報データ

- サンフランシスコ・ベイエリアのショッピングモールの顧客によるアンケート（9243件）
  - 応答変数：職業
    - `学生`，`退職者`，`専門職 / 管理職`，`主婦`，`労働者`，`聖職者`，`軍人`，`無職`，`営業職`
  - 予測変数：
    - `子供`，`言語`，`民族`，`性別`，`教育`，`収入`，`年齢` etc
  - 分析の目的
    - 職業の予測
    - 職業の違いをもたらしている変数の特定

職業の予測

- まず、データを訓練集合(80%)とテスト集合(20%)に分割
- 以下のパラメータで、$K = 9$ 各職業クラスを予測する木を構築
  - 頂点数：$J = 6$
  - 学習率：$\nu = 0.1$
- 結果：
  - 全体誤差率：42.5%
  - 最多の `専門職 / 管理職` のみを予測した場合は69%
  - `学生`，`退職者`，`専門職 / 管理職`，`主婦` を最もよく予測

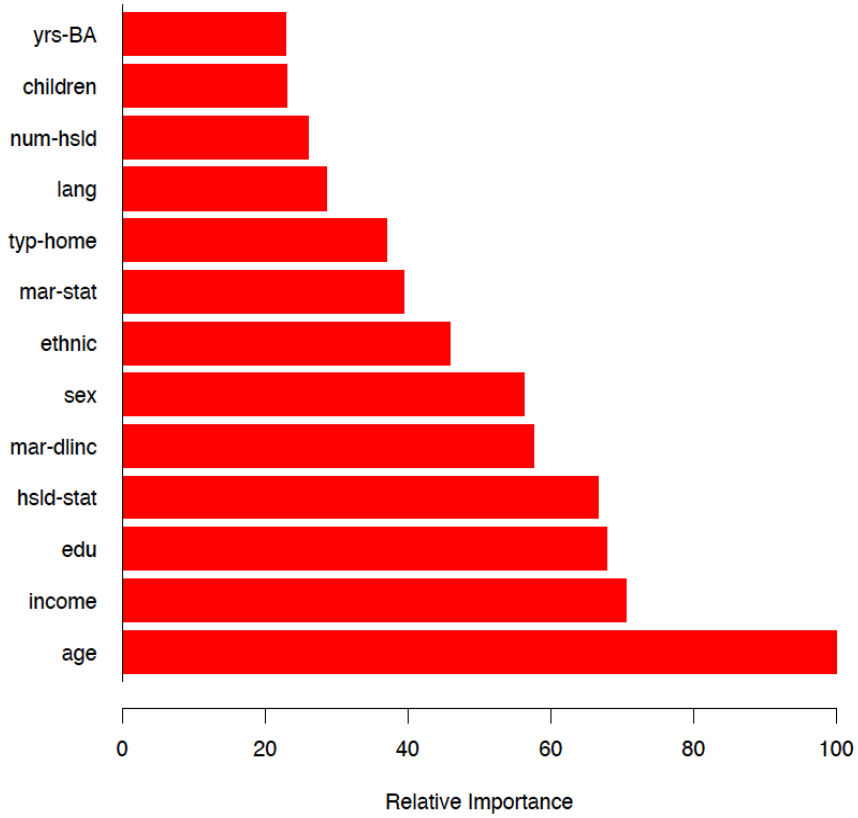


職業の違いをもたらしている変数の特定

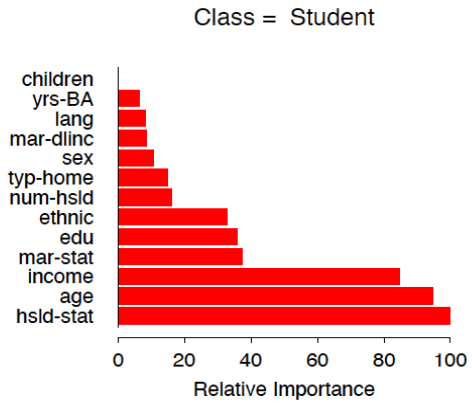
結果

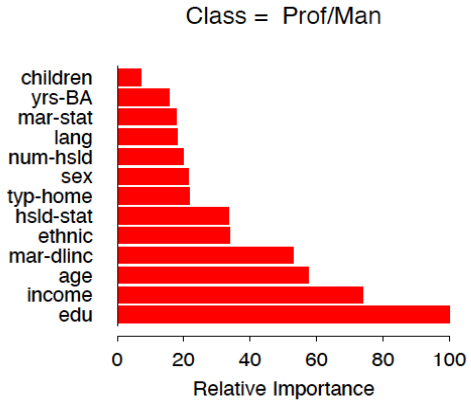
- 退職者 が高年齢層において高く、学生では逆
- 専門職 / 管理職 が中年層に対して最高

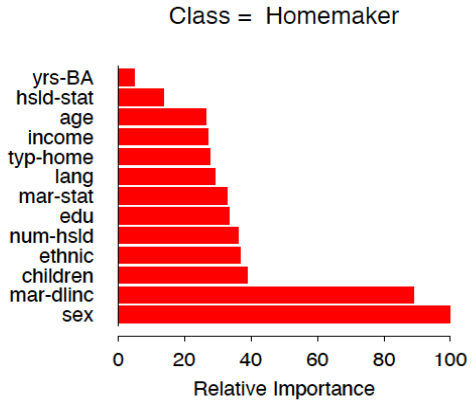
直感とも整合する理に適った結果が得られる

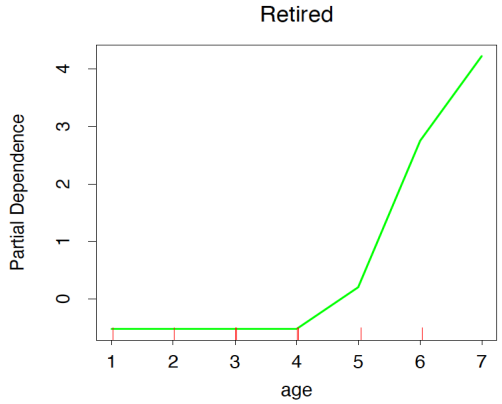


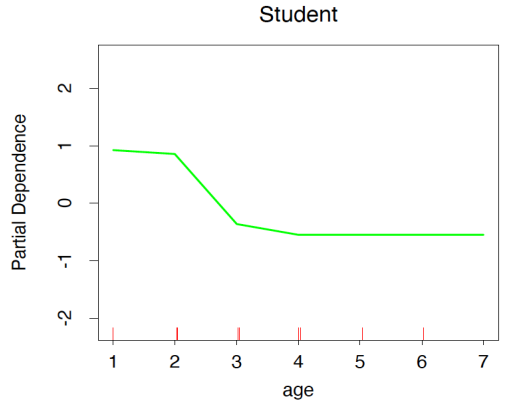

Class = Retired
yrs-BA
num-hsld
edu
children
typ-home
lang
mar-stat
hsld-stat
income
ethnic
sex
mar-dlinc
age
0 20 40 60 80 100
Relative Importance

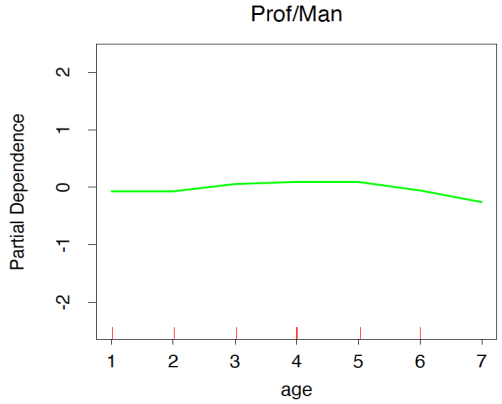

Class = Student
children
yrs-BA
lang
mar-dlinc
sex
typ-home
num-hsld
ethnic
edu
mar-stat
income
age
hsld-stat
0 20 40 60 80 100
Relative Importance


Class = Prof/Man
children
yrs-BA
mar-stat
lang
num-hsld
sex
typ-home
hsld-stat
ethnic
mar-dlinc
age
income
edu
0 20 40 60 80 100
Relative Importance


Class = Homemaker
yrs-BA
hsld-stat
age
income
typ-home
lang
mar-stat
edu
num-hsld
ethnic
children
mar-dlinc
sex
0 20 40 60 80 100
Relative Importance


Retired
Partial Dependence
0 1 2 3 4
1 2 3 4 5 6 7
age


Student
Partial Dependence
-2 -1 0 1 2
1 2 3 4 5 6 7
age


Prof/Man
Partial Dependence
-2 -1 0 1 2
1 2 3 4 5 6 7
age