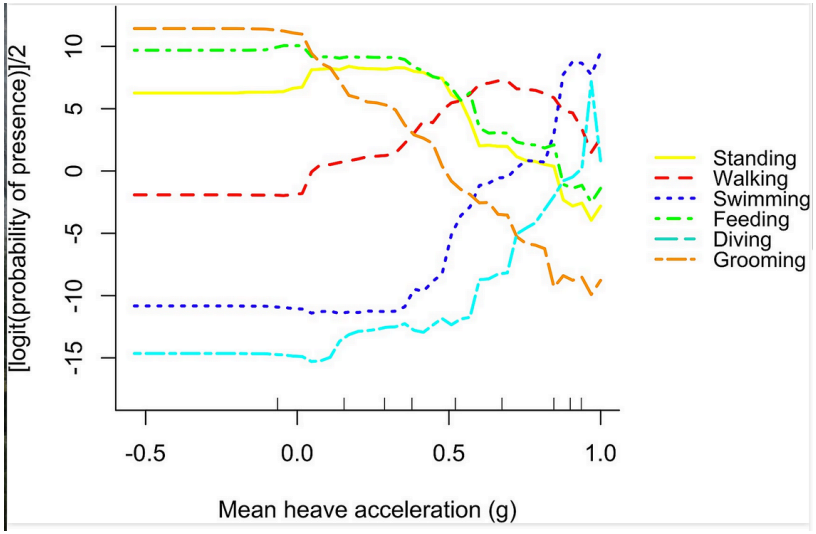
Variable Importance and Partial Dependence



Statistical Data Mining I
Rachael Hageman Blair

Outline

- Variable Importance
- Partial Dependence Plots

Recap:

- CART
  - Advantages: excellent for model interpretation, decision making, variable selection.
  - Disadvantages: unstable, typically underperforms compared to other methods.

- Ensemble Methods:
  - Advantages: stabilizes the tree.
  - Disadvantages: loss of model interpretation.

Variable Importance - Regression

- For a single decision tree, T, a measure of importance for X_l:

$$\mathcal{I}_\ell^2(T) = \sum_{t=1}^{J-1} \hat{\imath}_t^2 \, I(v(t) = \ell)$$

over all J-1 internal nodes.

- For multiple trees

$$\mathcal{I}_\ell^2 = \frac{1}{M} \sum_{m=1}^{M} \mathcal{I}_\ell^2(T_m).$$

Variable Importance - Classification

For K-class classification, K separate models $f_k(x), k = 1, 2, \ldots, K$ are induced, each consisting of a sum of trees

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

In this case (10.43) generalizes to

$$\mathcal{I}_{\ell k}^2 = \frac{1}{M} \sum_{m=1}^{M} \mathcal{I}_\ell^2(T_{km}). \tag{10.45}$$

Here $\mathcal{I}_{\ell k}$ is the relevance of X_ℓ in separating the class k observations from the other classes. The overall relevance of X_ℓ is obtained by averaging over all of the classes

$$\mathcal{I}_\ell^2 = \frac{1}{K} \sum_{k=1}^{K} \mathcal{I}_{\ell k}^2. \tag{10.46}$$

Partial Dependence Plots

Partial dependence functions can be used to interpret the results of any "black box" learning method. They can be estimated by

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

where $\{x_{1C}, x_{2C}, \ldots, x_{NC}\}$ are the values of X_C occurring in the training data. This requires a pass over the data for each set of joint values of X_S for which $\bar{f}_S(X_S)$ is to be evaluated. This can be computationally intensive,

Partial Dependence Plots

pdp: An R Package for Constructing Partial Dependence Plots

by Brandon M. Greenwell

Abstract Complex nonparametric models—like neural networks, random forests, and support vector machines—are more common than ever in predictive analytics, especially when dealing with large observational databases that don't adhere to the strict assumptions imposed by traditional statistical techniques (e.g., multiple linear regression which assumes linearity, homoscedasticity, and normality). Unfortunately, it can be challenging to understand the results of such models and explain them to management. Partial dependence plots offer a simple solution. Partial dependence plots are low-dimensional graphical renderings of the prediction function so that the relationship between the outcome and predictors of interest can be more easily understood. These plots are especially useful in explaining the output from black box models. In this paper, we introduce **pdp**, a general R package for constructing partial dependence plots.

Not specific to tree-based models !

Partial Dependence Plots

Type of model	R package	Object class
Decision tree	**C50** (Kuhn et al., 2015)	`"C5.0"`
	party	`"BinaryTree"`
	partykit	`"party"`
	rpart (Therneau et al., 2017)	`"rpart"`
Bagged decision trees	**adabag** (Alfaro et al., 2013)	`"bagging"`
	ipred (Peters and Hothorn, 2017)	`"classbagg"`, `"regbagg"`
Boosted decision trees	**adabag** (Alfaro et al., 2013)	`"boosting"`
	gbm	`"gbm"`
	xgboost	`"xgb.Booster"`
Cubist	**Cubist** (Kuhn et al., 2016)	`"cubist"`
Discriminant analysis	**MASS** (Venables and Ripley, 2002)	`"lda"`, `"qda"`
Generalized linear model	**stats**	`"glm"`, `"lm"`
Linear model	**stats**	`"lm"`
Nonlinear least squares	**stats**	`"nls"`
Multivariate adaptive regression splines (MARS)	**earth** (Milborrow, 2017a)	`"earth"`
	mda (Leisch et al., 2016)	`"mars"`
Projection pursuit regression	**stats**	`"ppr"`
Random forest	**randomForest**	`"randomForest"`
	party	`"RandomForest"`
	partykit	`"cforest"`
	ranger (Wright, 2017)	`"ranger"`
Support vector machine	**e1071** (Meyer et al., 2017)	`"svm"`
	kernlab (Karatzoglou et al., 2004)	`"ksvm"`

Partial Dependence Plots

Constructing a PDP (3) in practice is rather straightforward. To simplify, let $z_s = x_1$ be the predictor variable of interest with unique values $\{x_{11}, x_{12}, \ldots, x_{1k}\}$. The partial dependence of the response on x_1 can be constructed as follows:

1. For $i \in \{1, 2, \ldots, k\}$:
   (a) Copy the training data and replace the original values of x_1 with the constant x_{1i}.
   (b) Compute the vector of predicted values from the modified copy of the training data.
   (c) Compute the average prediction to obtain $\bar{f}_1(x_{1i})$.
2. Plot the pairs $\{x_{1i}, \bar{f}_1(x_{1i})\}$ for $i = 1, 2, \ldots, k$.

Partial Dependence Plots

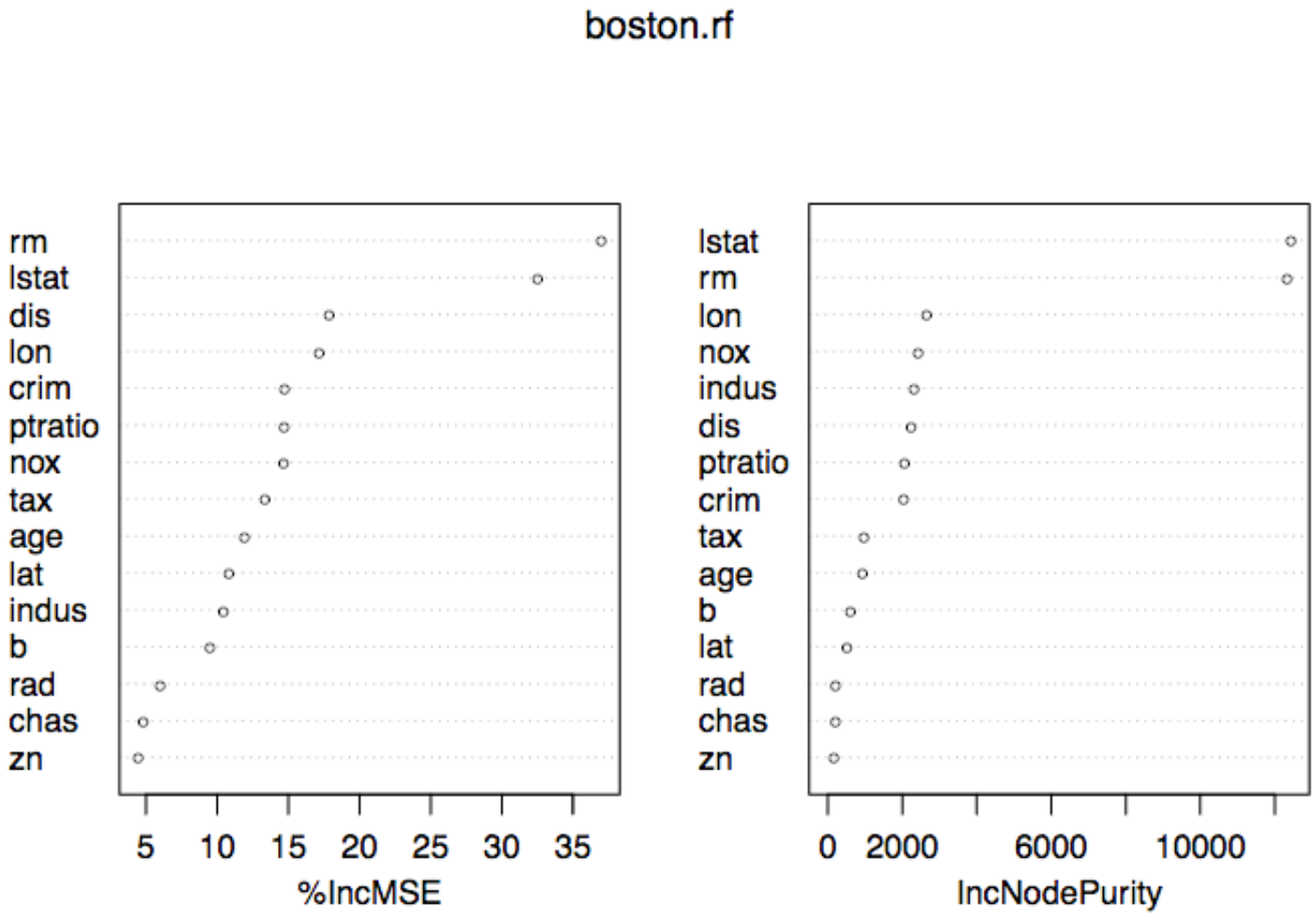


Figure 1: Dotchart of variable importance scores for the Boston housing data based on a random forest with 500 trees.

Partial Dependence Plots

```
# Figure 2 (right)
boston.rf %>%  # the %>% operator is read as "and then"
  partial(pred.var = "lstat") %>%
  plotPartial(smooth = TRUE, lwd = 2, ylab = expression(f(lstat)))
```

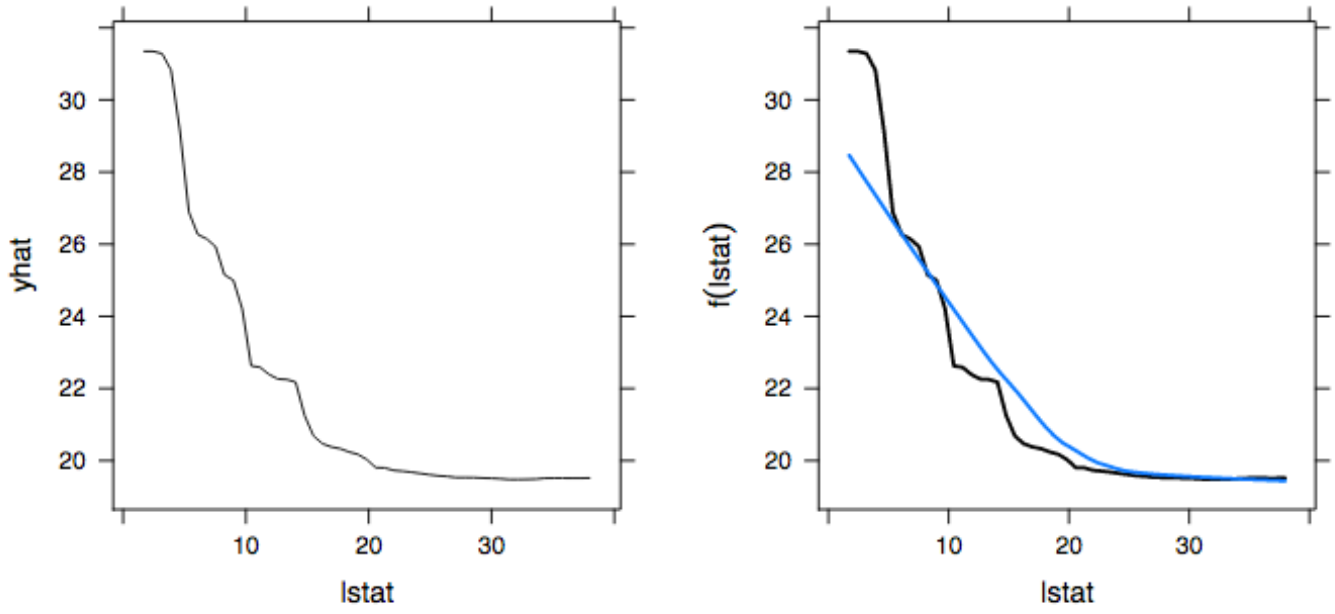


Figure 2: Partial dependence of `cmedv` on `lstat` based on a random forest. *Left*: Default plot. *Right*: Customized plot obtained using the `plotPartial` function.

Partial Dependence Plots

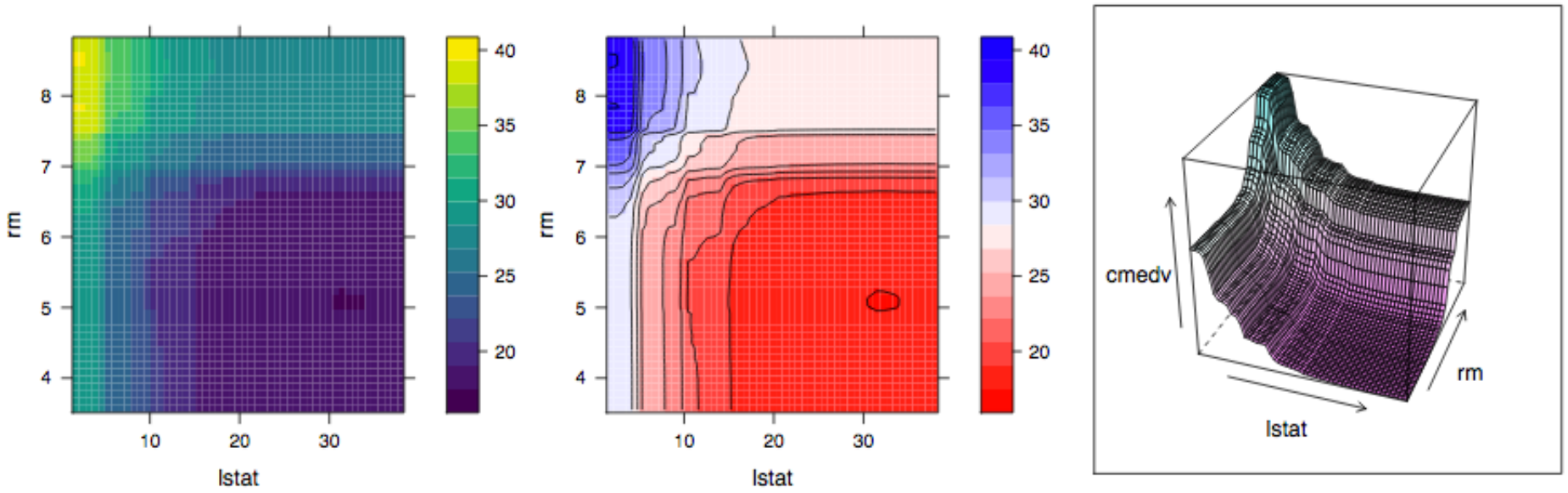


Figure 3: Partial dependence of `cmedv` on `lstat` and `rm` based on a random forest. *Left*: Default plot. *Middle*: With contour lines and a different color palette. *Right*: Using a 3-D surface.

Variable Importance - SPAM:

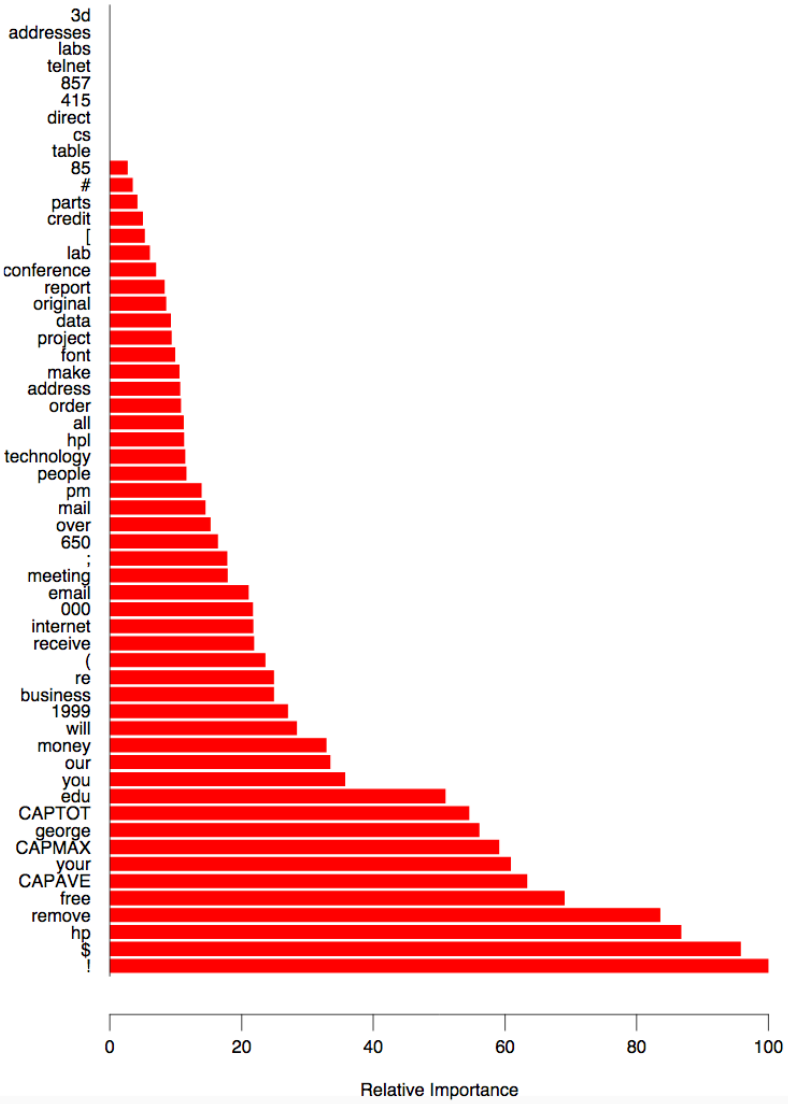


FIGURE 10.6. *Predictor variable importance spectrum for the* `spam` *data.*

Partial Dependence Plots - SPAM

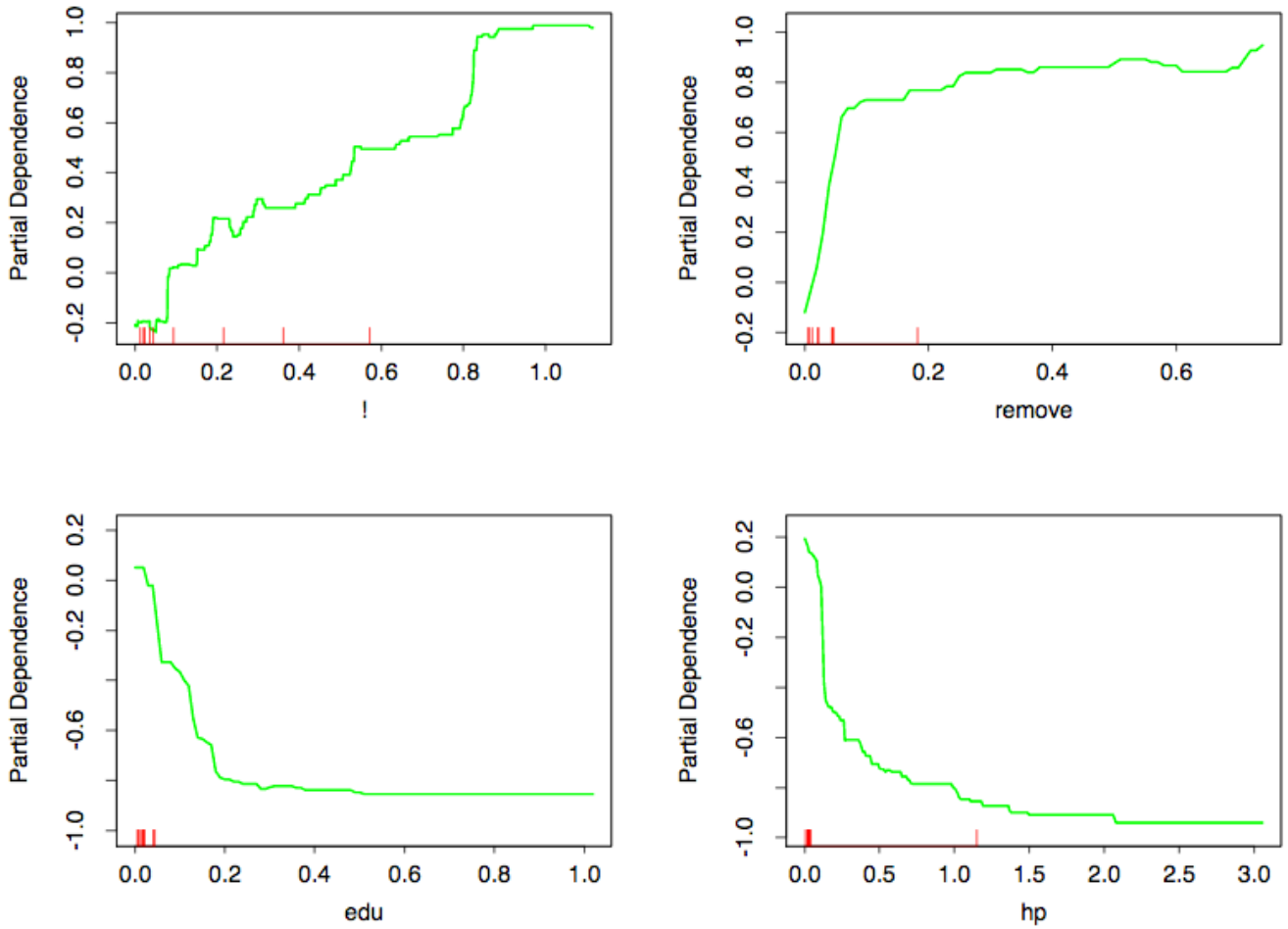


FIGURE 10.7. *Partial dependence of log-odds of* **spam** *on four important predictors. The red ticks at the base of the plots are deciles of the input variable.*

Partial Dependence Plots - SPAM:

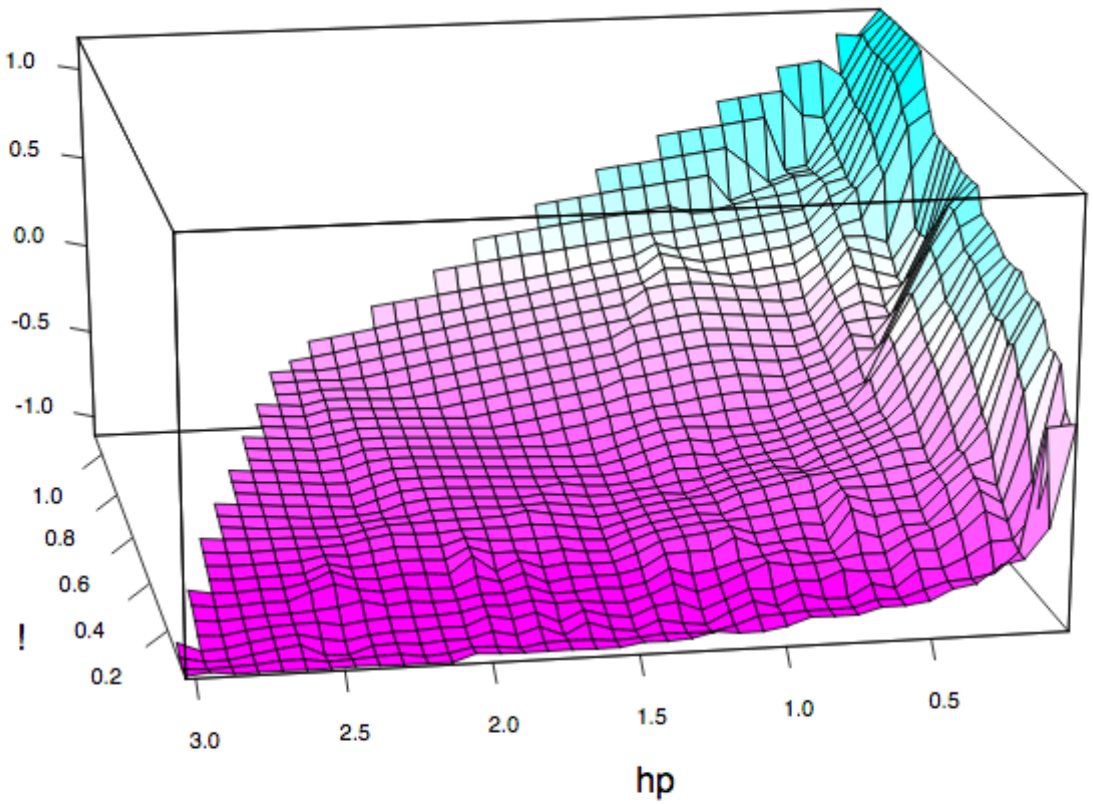


FIGURE 10.8. *Partial dependence of the log-odds of* `spam` *vs.* `email` *as a function of joint frequencies of* `hp` *and the character* `!`.

California Housing Examples

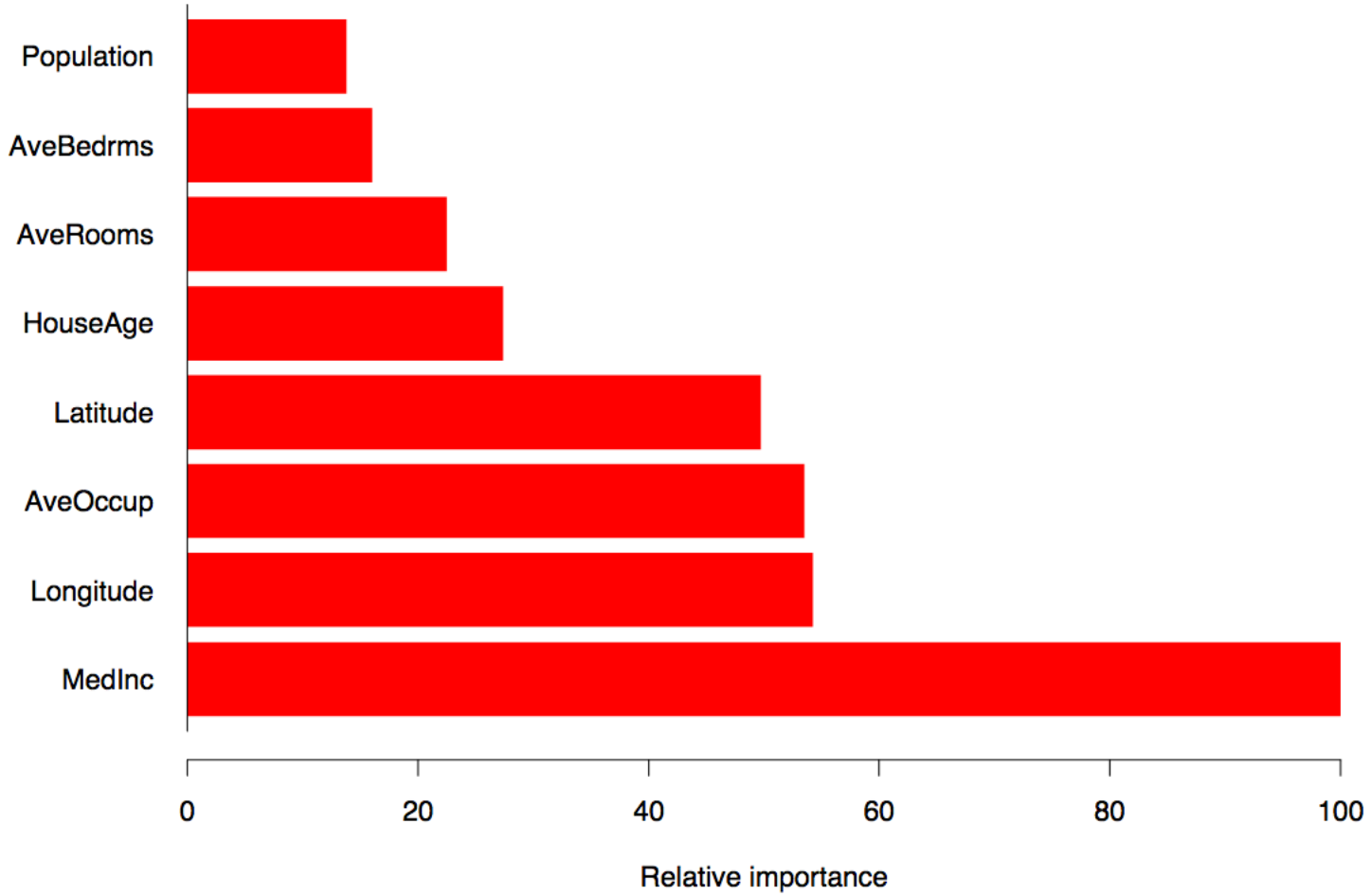


FIGURE 10.14. *Relative importance of the predictors for the California housing data.*

California Housing Examples

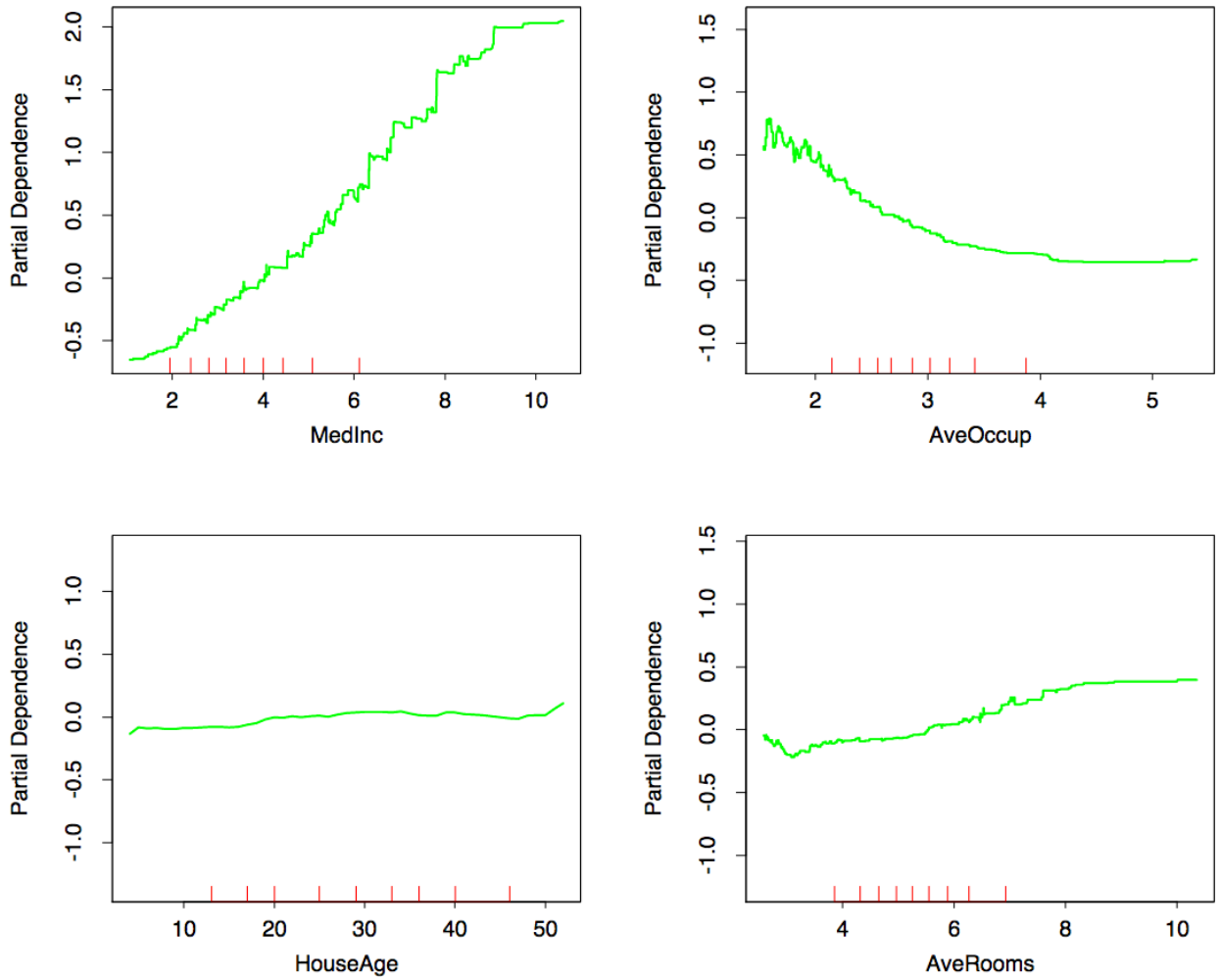


FIGURE 10.15. *Partial dependence of housing value on the nonlocation variables for the California housing data. The red ticks at the base of the plot are deciles of the input variables.*