Predictive Learning via Rule Ensembles

Jerome H. Friedman

Bogdan E. Popescu

Stanford University

PREDICTION (Regression/Classification)

$y =$ outcome/response variable

$\mathbf{x} = \{x_1, \cdots, x_n\}$ predictors

Goal: $\hat{y} = F(\mathbf{x})$

Want good $F(\mathbf{x})$

ACCURACY

Cost for error: $L(y, F)$

$$L(y,F) = (y-F)^2,\ |\,y-F\,| \qquad y \in R$$

$y \in \{-1, 1\}$:

$$L(y,F) = \log(1 + e^{-yF}) \quad \text{logistic reg.}$$

$$L(y,F) = (1 - yF)_+ \qquad \text{SVM}$$

any – log (likelihood)

many many more

Lack of accuracy ("risk"):

$$R(F) = E_{\mathbf{x}y} L(y, F(\mathbf{x}))$$

Optimal ("target") function:

$$F^* = \arg\min_F R(F)$$

Don't know $p(\mathbf{x}, y)$

Learning: $T = \{\mathbf{x}_i, y_i\}_1^N$ "training" sample

$$F(\mathbf{x}) = \text{learning procedure } (T) \simeq F^*(\mathbf{x})$$

ENSEMBLE LEARNING

$$F(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m f_m(\mathbf{x})$$

$\{f_m(\mathbf{x})\}_1^M$ = basis functions ("base learners")

Base learner: $f_m(\mathbf{x}) = f(\mathbf{x}; \mathbf{p}_m)$

$\{f(\mathbf{x}; \mathbf{p})\}_{\mathbf{p} \in P}$ = function class

Methods differ: choice $f(\mathbf{x}; \mathbf{p})$

select: $\{f_m(\mathbf{x})\}_1^M \subset \{f(\mathbf{x}; \mathbf{p})\}_{\mathbf{p} \in P}$,

determine: $\{a_m\}_0^M$

GENERIC ENSEMBLE GENERATION PROC. (EGP)

$F_0(\mathbf{x}) = 0$

For $m = 1$ to M $\{$

$\quad \mathbf{p}_m = \arg\min_{\mathbf{p}} \sum_{i \in S_m(\eta)} L(y_i, F_{m-1}(\mathbf{x}_i) + f(\mathbf{x}_i; \mathbf{p}))$

$\quad f_m(\mathbf{x}) = f(\mathbf{x}; \mathbf{p}_m)$

$\quad F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot f_m(\mathbf{x})$

$\}$

ensemble $= \{f_m(\mathbf{x})\}_1^M$

EGP CONTROL PARAMETERS (FP 2003)

$S_m(\eta) =$ random subsample of size $\eta \leq N$

$\eta \downarrow \Rightarrow$ ensemble diversity $\uparrow$ and comp. $\downarrow$

Auxiliary "memory" function: step m

$F_{m-1}(\mathbf{x}) = \nu \cdot \sum_{k=1}^{m-1} f_k(\mathbf{x})$

retains info $\{f_k(\mathbf{x})\}_1^{m-1}$

$0 \leq \nu \leq 1 =$ "memory control" parameter

POPULAR ENSEMBLE METHODS

Bagging: $L(y,\hat{y}) = (y-\hat{y})^2$, $\nu = 0$, $\eta = N/2$

$a_0 = 0$, $\{a_m = 1/M\}_1^M \Rightarrow$ simple average

Random forests: bagging with randomized trees

AdaBoost: $y \in \{-1, 1\}$; $L(y,\hat{y}) = \exp(-y \cdot \hat{y})$

$\nu = 1$ and $\eta = N$, $\hat{y} = sign(F_M(\mathbf{x}))$

MART (TreeNet): arbitrary y and $L(y,\hat{y})$

Defaults: $\nu = 0.1$, $\eta = N/2$, $\hat{y} = F_M(\mathbf{x})$

ISLE (FP 2003): $F(\mathbf{x}) = \hat{a}_0 + \sum_{m=1}^{M} \hat{a}_m f_m(\mathbf{x})$

Lasso regression y on $\{f_m(\mathbf{x})\}_1^M$:

$$\{\hat{a}_m\}_0^M = \arg\min_{\{a_m\}_0^M}$$

$$\sum_{i=1}^{N} L\left(y_i\,, a_0 + \sum_{m=1}^{M} a_m f_m(\mathbf{x}_i)\right)$$

$$+\lambda \cdot \sum_{m=1}^{M} |\, a_m \,|$$

$\lambda \uparrow \Rightarrow$ more shrinkage and *diversity* of $\{|\,\hat{a}_m\,|\}_1^M$

with many $\hat{a}_m = 0$ (selection effect)

estimated by cross–validation

Almost all ensemble learning implementations:

Base learners: $f(\mathbf{x};\mathbf{p}) =$ decision trees

$\mathbf{p} =$ splitting variables and value subsets

defining branches

Reasons:

Desirable data mining properties

Accuracy helped the most

Fast (approximate) algorithms

Here base learners = RULES

$J(m) \subseteq \{x_1, x_2, \cdots, x_n\}$

s_{jm} = subset of values of $x_j \in J(m)$

$$f_m(\mathbf{x}) = r_m(\mathbf{x}) = \prod_{j \in J(m)} I(x_j \in s_{jm}) \in \{0, 1\}$$

$\{x_j\}_{j \in J(m)}$ "define" $r_m(\mathbf{x})$

EXAMPLE

$$r_m(\mathbf{x}) = \begin{cases} I(18 \leq \text{age} < 34) \\ \cdot I(\text{marital status} \in \{\text{single, living together} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{–not married}\}) \\ \cdot I(\text{householder status} \ = \ \text{rent}) \end{cases}$$

$= 1 \Rightarrow$ greater odds of visiting bars & night clubs

RULE GENERATION

$$f(\mathbf{x};\mathbf{p}_m) = \prod_{j \in J(m)} I(x_j \in s_{jm})$$ in EGP too slow

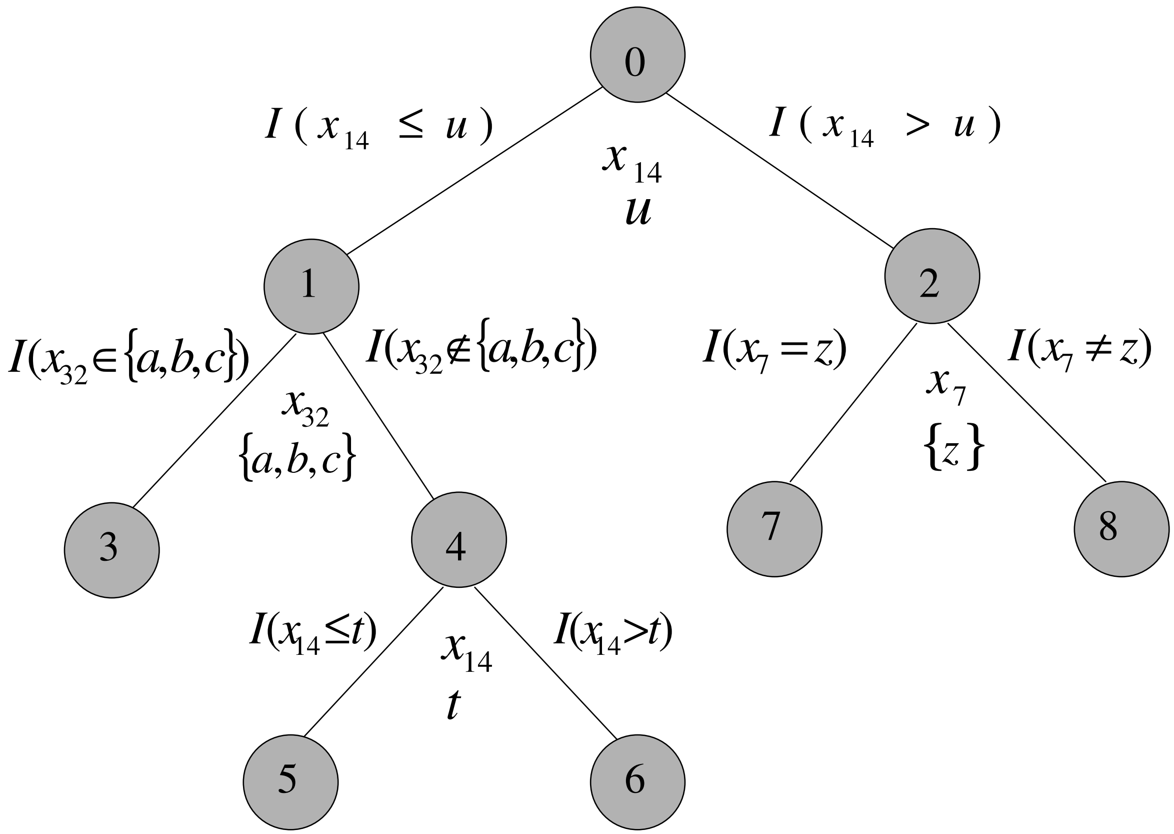
(combinatorial optimization at each step)

Fast algorithms for decision trees $\Rightarrow$

$f(\mathbf{x};\mathbf{p}) = T(\mathbf{x};\mathbf{p}) =$ decision tree in EGP

harvest rules from resulting $\{T_m(\mathbf{x})\}_1^M$

All tree nodes (interior and terminal) represent rules



$$
\begin{aligned}
r_1(\mathbf{x}) &= I(x_{14} \leq u) \\
r_6(\mathbf{x}) &= I(t < x_{14} \leq u) \cdot I(x_{32} \notin \{a, b, c\}) \\
r_7(\mathbf{x}) &= I(x_{14} > u) \cdot I(x_7 = z).
\end{aligned}
$$

All such rules derived from all trees $\{T_m(\mathbf{x})\}_1^M$

constitute the rule ensemble $\{r_k(\mathbf{x})\}_1^K$

$M =$ large $\Rightarrow$ $K =$ much larger

Model: $F(\mathbf{x}) = \hat{a}_0 + \sum_{k=1}^K \hat{a}_k r_k(\mathbf{x})$

$\{\hat{a}_k\}_0^K =$ lasso regression (y on $\{r_k(\mathbf{x})\}_1^K$)

Lasso selection effect $\Rightarrow$

most ($\sim$80% – 90%) $\hat{a}_k = 0$

LINEAR BASIS fUNCTIONS

Linear targets $F^*(\mathbf{x}) = b_0 + \sum_{j=1}^{n} b_j x_j$

most difficult for rules (and trees)

$\Rightarrow$ include $\{x_j\}_1^n$ in ensemble

RULE BASED INTERPRETATION

$F(\mathbf{x}) =$ linear model in $\{r_k(\mathbf{x})\}$ & $\{x_j\}$

Both rules and linear terms easy to interpret

Examine most important terms for interpretation

Linear model:

Rule importance: $I_k = |\,\hat{a}_k\,| \cdot \sqrt{s_k(1 - s_k)}$

$s_k =$ support

Linear importance: $I_j = |\,\hat{b}_j\,| \cdot std(x_j)$

LOCAL IMPORTANCE

$\mathbf{x}$ = prediction point $\in X$

Rules: $I_k(\mathbf{x}) = |\,\hat{a}_k\,| \cdot |\,r_k(\mathbf{x}) - s_k\,|$

Linear: $I_j(x_j) = |\,\hat{b}_j\,| \cdot |\,x_j - \bar{x}_j\,|$

Change in $|F(\mathbf{x})\,|$ when coefficient $\rightarrow 0$

Note: ave. (rms) over $\mathbf{x}$ = standard global measures

Average over $S \subset X$: $I_k(S) = \frac{1}{|\,S\,|} \sum_{\mathbf{x}_i \in S} I_k(\mathbf{x}_i)$;

INPUT VARIABLE IMPORTANCE

Most important variables are those that define

most important terms (rules or linear)

Importance of x_j at $\mathbf{x}$:

$J_j(\mathbf{x}) = I_j(x_j) + \sum_{x_j \in r_k} I_k(\mathbf{x})/m_k$

$I_j(x_j) =$ importance of x_j linear term

$I_k(\mathbf{x}) =$ importance of kth rule (containing x_j)

$m_k = \#$ variables defining kth rule

Average over S using $I_j(S)$ & $I_k(S)$

PARTIAL DEPENDENCE FUNCTIONS

$\mathbf{x}_s =$ selected subset of input variables, $s \subset \{1, 2, \cdots, n\}$

$\mathbf{x} = (\mathbf{x}_s, \mathbf{x}_{\backslash s})$

Partial dep. on $\mathbf{x}_s$: $\quad F_s(\mathbf{x}_s) = E_{\mathbf{x}_{\backslash s}}[F(\mathbf{x}_s, \mathbf{x}_{\backslash s})]$

Estimate: $\quad \hat{F}_s(\mathbf{x}_s) = \frac{1}{N} \sum_{i=1}^{N} F(\mathbf{x}_s, \mathbf{x}_{i \backslash s})$

$\{\mathbf{x}_{i \backslash s}\}_1^N =$ data values of $\mathbf{x}_{\backslash s}$

Used (Friedman 2001) to view dep. of $F(\mathbf{x})$

on $\mathbf{x}_s$ *accounting* for ave. effects of $\mathbf{x}_{\backslash s}$

INTERACTION EFFECTS

$F(\mathbf{x})$ has *interaction* between x_j & x_k

$\Rightarrow F(x_j \,|\, \mathbf{x}_{\setminus j}) - F(x'_j \,|\, \mathbf{x}_{\setminus j})$ depends on x_k

$$E_{\mathbf{x}} \left[\frac{\partial^2 F(\mathbf{x})}{\partial x_j \, \partial x_k} \right]^2 > 0 \quad \text{(cat. } \Rightarrow \text{ finite diff.)}$$

If no interaction between x_j & x_k:

$$F(\mathbf{x}) = f_{\backslash j}(\mathbf{x}_{\backslash j}) + f_{\backslash k}(\mathbf{x}_{\backslash k})$$

Partial dep.: $F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k)$

$$H_{jk}^2 = \sum_{i=1}^{N} [\hat{F}_{jk}(x_{ij}, x_{ik}) - \hat{F}_j(x_{ij}) - \hat{F}_k(x_{ik})]^2$$

$$\Big/ \sum_{i=1}^{N} \hat{F}_{jk}^2(x_{ij}, x_{ik})$$

If x_j interacts with NO other variable:

$$F(\mathbf{x}) = f_j(x_j) + f_{\backslash j}(\mathbf{x}_{\backslash j}) \quad \text{(additive)}$$

$$F(\mathbf{x}) = F_j(x_j) + F_{\backslash j}(\mathbf{x}_{\backslash j})$$

$$F_j(x_j) = \text{partial dep. on } x_j$$

$$F_{\backslash j}(\mathbf{x}_{\backslash j}) = \text{partial dep. on } \mathbf{x}_{\backslash j}$$

$$H_j^2 = \sum_{i=1}^N [F(\mathbf{x}_i) - \hat{F}_j(x_{ij}) - \hat{F}_{\backslash j}(\mathbf{x}_{i \backslash j})]^2 \Big/ \sum_{i=1}^N F^2(\mathbf{x}_i)$$

$F(\mathbf{x})$ has three–variable interaction among x_j, x_k, & x_l

$$\text{if} \;\; E_{\mathbf{x}} \left[\frac{\partial^3 F(\mathbf{x})}{\partial x_j \, \partial x_k \, \partial x_l} \right]^2 > 0 \;\; \text{(cat. } \Rightarrow \text{ finite diff.)}$$

If no three–variable interaction among x_j, x_k, & x_l:

$$F(\mathbf{x}) = f_{\backslash j}(\mathbf{x}_{\backslash j}) + f_{\backslash k}(\mathbf{x}_{\backslash k}) + f_{\backslash l}(\mathbf{x}_{\backslash l})$$

$$F_{jkl}(x_j, x_k, x_l) = F_{jk}(x_j, x_k) + F_{jl}(x_j, x_l) + F_{kl}(x_k, x_l)$$

$$-F_j(x_j) - F_k(x_k) - F_l(x_l)$$

$$H^2_{jkl} = \hat{E}[LHS - RHS]^2 / \hat{E}[LHS^2]$$

STRATEGY

(1) identify important input variables x_j

(2) among these use H_j to identify which

are interacting with others

(3) for each interacting x_j use $\{H_{jk}\}_{k \neq j}$ to

identify $\{x_k\}$ with which it interacts

(4) use H_{jkl} to check for three–variable interactions

(5) view relevant partial dependence plots

ILLUSTRATION

Defaults:

$$\nu = 0.01, \quad \eta = \min(N/2, 100 + 6\sqrt{N})$$

Ave. tree size $\bar{L} = 4$ terminal nodes

$M = 333$ trees $\Rightarrow K \simeq 2000$ rules

$+$ linear terms

BOSTON HOUSING DATA

$N = 506$ neighborhoods in the Boston metropolitan area

14 summary statistics were collected in each

$y =$ median house value, $\mathbf{x} = 13$ other (predictor) variables

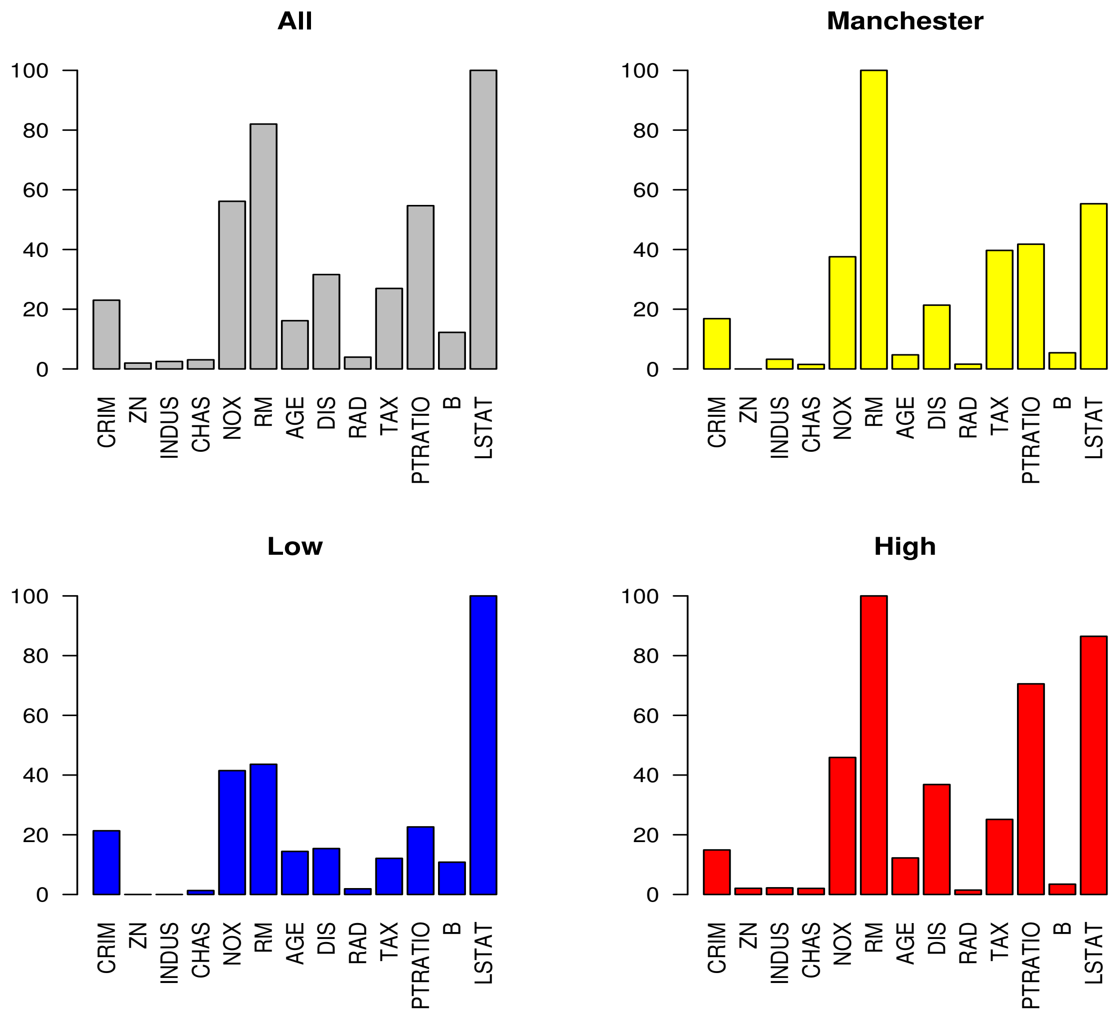
RuleFit model: 215 terms (rules+ linear)

Relative average absolute error (50–fold X–val)

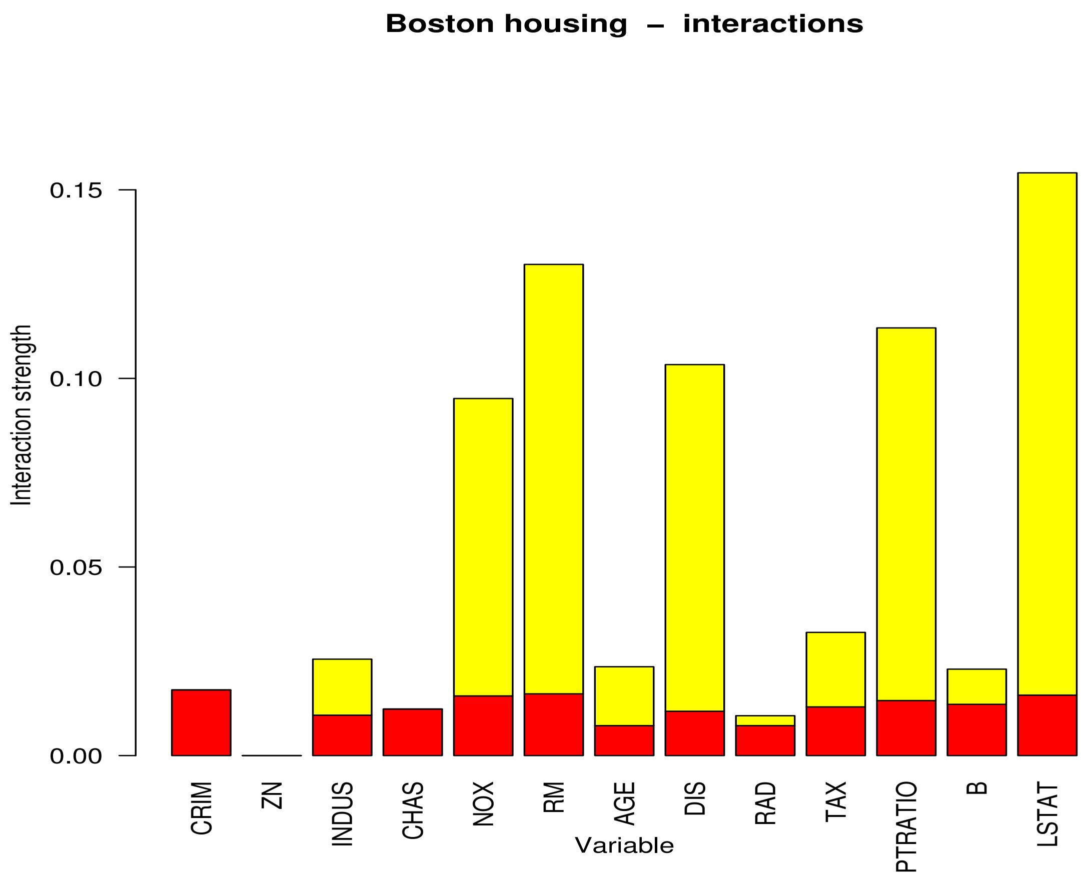
	Full	Additive	Linear
Prediction	0.33	0.37	0.49

Boston housing data: most important rules

Imp.	Coeff	Sup.	Rule
100	-0.40		linear: $LSTAT$
37	-0.036		linear: AGE
36	10.1	0.01	$DIS < 1.4$ & $PTRATIO > 17.9$ & $LSTAT < 10.5$
35	2.26	0.23	$RM > 6.62$ & $NOX < 0.67$
26	-2.27	0.88	$RM < 7.45$ & $DIS > 1.37$
20	2.58	0.05	$RM > 7.44$ & $PTRATIO < 17.9$

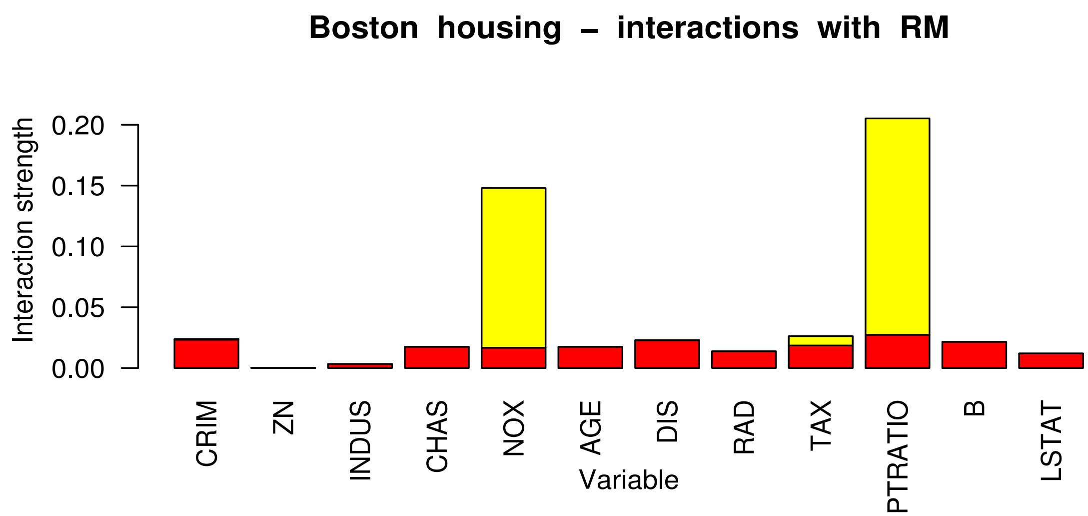


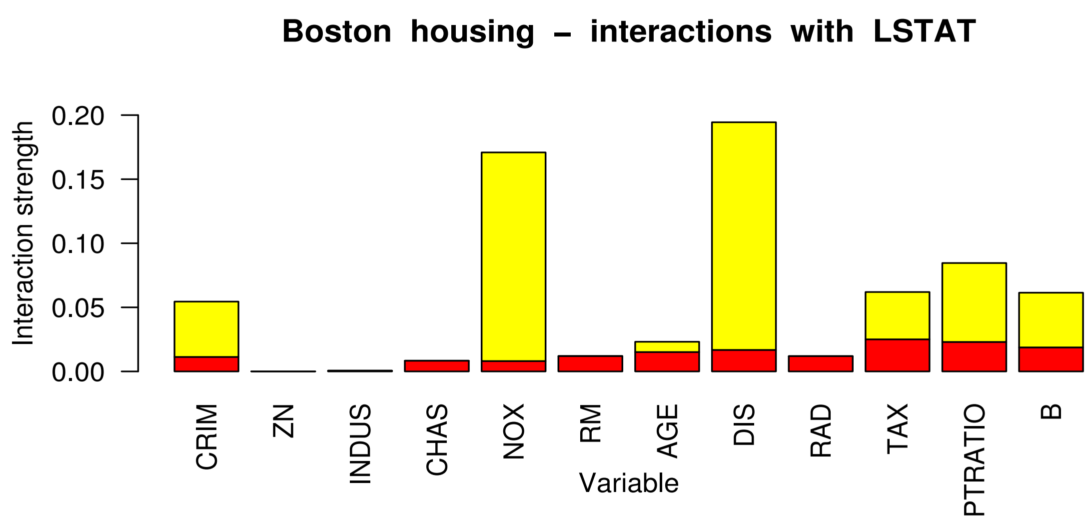
Boston housing – variable importance



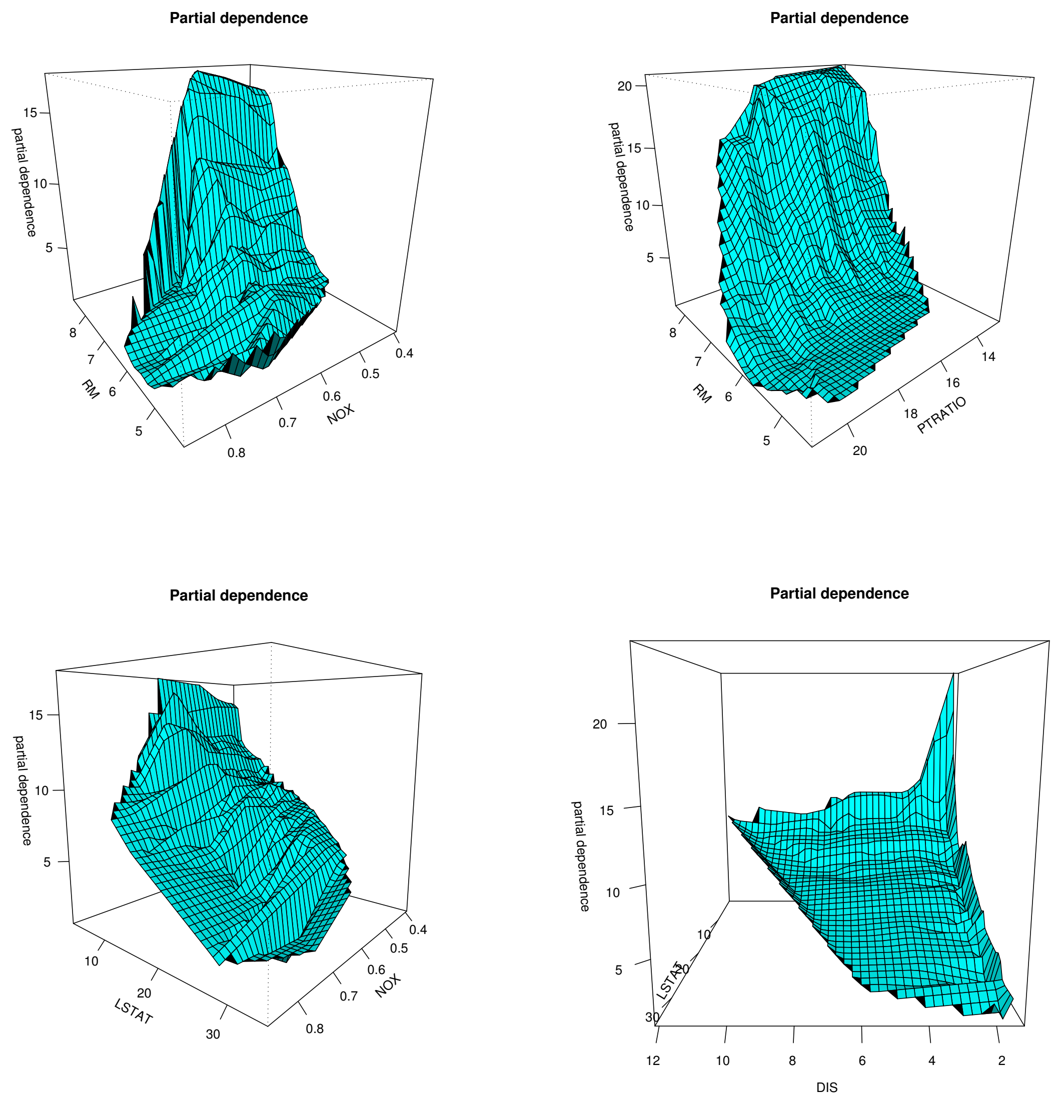
$$\tilde{H}_j = H_j - \bar{H}_j^{(0)} \text{ (yellow)}, \;\; \sigma_j^{(0)} \text{ (red)}$$

$$\bar{H}_j^{(0)} = \text{expected null}, \;\; \sigma_j^{(0)} = \text{std. dev. null}$$





$H_{jkl} \Rightarrow$ no 3–var. interactions involving *RM* or *LSTAT*



Boston housing – partial dependence plots

Future Work: rule summarization

Bibliography

Talk: http://www-stat.stanford.edu/~jhf/talks/toronto2.pdf

ISLE: FP (2003):

http://www-stat.stanford.edu/~jhf/ftp/isle.pdf

Fast lasso: FP (2004):

http://www-stat.stanford.edu/~jhf/ftp/path.pdf

LARS: Efron *et al*; Rosset & Zhu *et al*