

Predictive Learning via Rule Ensembles

PREDICTIVE LEARNING

y = “response”, “output” attribute (unknown)

$\mathbf{x} = (x_1, x_2, \dots, x_n)$ = “input”, “predictor” attributes

Prediction: $\hat{y} = F(\mathbf{x})$; $L(y, \hat{y})$ = loss criterion

Lack of accuracy (“risk”): $R(F) = E_{\mathbf{x}y}L(y, F(\mathbf{x}))$

Optimal (“target”) function: $F^* = \arg \min_F R(F)$

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

Goal: approximate $F^*(\mathbf{x})$ by

$F(\mathbf{x})$ = learning procedure (T)

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

$\mathbf{p} = (p_1, p_2, \dots) =$ parameters

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

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

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 {

$$\mathbf{p}_m = \arg \min_{\mathbf{p}}$$

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

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

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

}

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

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

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

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

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

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

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

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

estimated by cross-validation

EGP: $(\eta, \nu) = \text{small}$; $\nu \simeq 0.01$, $\eta \sim \sqrt{N}$

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

Let $x_j \in S_j$ and $s_{jm} \subseteq S_j$

$$f(\mathbf{x}; \mathbf{p}_m) = r_m(\mathbf{x}) = \prod_{j=1}^n I(x_j \in s_{jm}) \in \{0, 1\}$$

Numeric variables: $s_{jm}: t_{jm} < x_j \leq u_{jm}$

Categorical variables: s_{jm} explicitly enumerated

If $s_{jm} = S_j \Rightarrow$ omit x_j factor from rule:

$$r_m(\mathbf{x}) = \prod_{s_{jm} \neq S_j} I(x_j \in s_{jm})$$

$\{x_j \mid s_{jm} \neq S_j\}$ “define” $r_m(\mathbf{x})$

EXAMPLE

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

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

RULE GENERATION

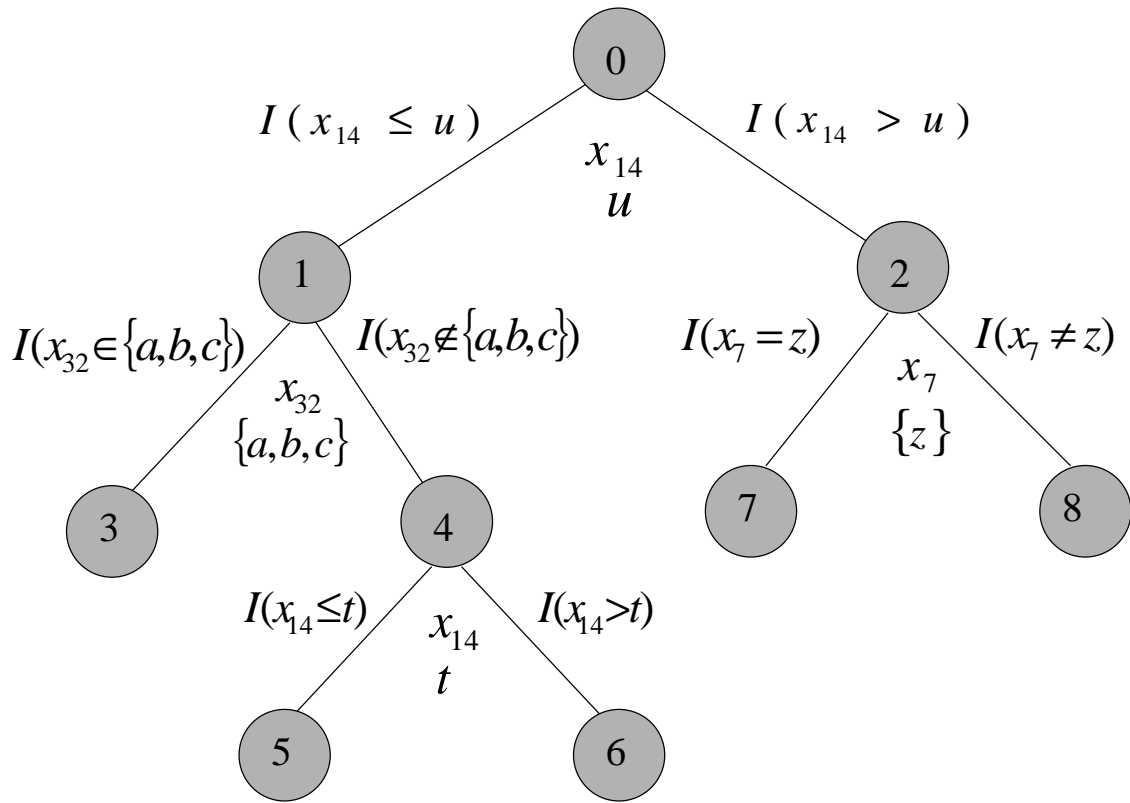
$$f(\mathbf{x}; \mathbf{p}_m) = \prod_{j=1}^n I(x_j \in s_{jm}) \text{ in EGP difficult}$$

Fast algorithms for decision trees \Rightarrow

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

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

All tree nodes (interior and terminal) represent rules



$$r_1(\mathbf{x}) = I(x_{14} \leq u)$$

$$r_6(\mathbf{x}) = I(t < x_{14} \leq u) \cdot I(x_{32} \notin \{a, b, c\})$$

$$r_7(\mathbf{x}) = I(x_{14} > u) \cdot I(x_7 = z).$$

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

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

$$K = \sum_{m=1}^M 2(t_m - 1)$$

$t_m = \#$ terminal nodes of T_m

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

TREE SIZE

Controls maximum complexity of rules

$t_m \uparrow \Rightarrow \# \text{ factors in rules } \uparrow$

$\# \text{ factors } \uparrow \Rightarrow \text{ higher order interactions}$

$F^*(\mathbf{x}) \sim \text{ high order interactions } \Rightarrow \text{ larger trees}$

$F^*(\mathbf{x}) \sim \text{ main effects and/or low order}$

$\text{interaction } \Rightarrow \text{ smaller trees}$

Our strategy: tree size $t_m = \text{random}$

$$t_m \sim \exp(-t; \bar{L})$$

$\bar{L} = 2 \Rightarrow$ one factor rules $\Rightarrow F(\mathbf{x}) =$ main effects only

$\bar{L} > 2 \Rightarrow$ distribution of tree sizes $\{t_m\}_1^M$

Exp. dist. $\Rightarrow \sim$ uniform ensemble rule complexity

Lasso chooses among them

ACCURACY

FP 2003: compared tree ensembles

Bag, RF, MART, ISLEs: $\text{EGP}(\nu, \eta) \rightarrow$ lasso

large Monte Carlo + real data

Results: $\text{EGP}(\nu \simeq 0.01, \eta \sim \sqrt{N}) \rightarrow$ lasso best

FP (2005): same data sets (regression & classification)

RuleFit using same EGP

slightly more accurate $\sim 5\% - 10\%$ than ISLE

Considerably better than others

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

Accuracy increase (empirical):

sometimes dramatic

usually modest but significant

never < 0

RULE BASED INTERPRETATION

$F(\mathbf{x}) =$ linear model in $\{r_k(\mathbf{x})\}_1^K$ & $\{x_j\}_1^n$

Both rules and linear terms easy to interpret

$K + n \sim 10^3$, $\#\{\hat{a}_k, \hat{b}_j \neq 0\} \sim 10^2$

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

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 k th rule (containing x_j)

m_k = # variables defining k th rule

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

PARTIAL DEPENDENCE FUNCTIONS

\mathbf{x}_s = selected subset of input variables

indexed by $s \subset \{1, 2, \dots, n\}$

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

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

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

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

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

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

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 :

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

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

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

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

If x_j interacts with NO other variable:

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

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

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

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

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

$F(\mathbf{x})$ 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 :

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

$$F_{jkl}(x_j, x_k, x_l) = F_{jk}(x_j, x_k) + F_{jl}(x_j, x_l) + F_{kl}(x_k, x_l) \\ - F_j(x_j) - F_k(x_k) - F_l(x_l)$$

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

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) if $\#\{x_k\} > 1$ use H_{jkl} to check for
three-variable interactions
- (5) view relevant partial dependence plots

SPURIOUS INTERACTIONS

Empirical: $H = H(F(\mathbf{x}))$; $F(\mathbf{x}) \leftarrow$ data

Truth: $H^* = H(F^*(\mathbf{x}))$; $F^*(\mathbf{x}) \leftarrow$ population

Hope: $F(\mathbf{x}) \simeq F^*(\mathbf{x})$ (accurate estimation)

Assume: $F(\mathbf{x}) \simeq F^*(\mathbf{x}) \Rightarrow H \simeq H^*$

NOT necessarily so!

$$\frac{1}{N} \sum_{i=1}^N [F^*(\mathbf{x}_i) - F(\mathbf{x}_i)]^2 = \text{small} \not\Rightarrow H \simeq H^*$$

WHY?

Associations among $\{x_1, x_2, \dots, x_n\}$

Example: $|corr(x_j, x_k)| = \text{big}$

$$\Rightarrow f_j(x_j) \simeq g_{jk}(x_j, x_k) \quad (\text{e.g. } x_j^2 \simeq x_j \cdot x_k)$$

$$\text{or } f_{jl}(x_j, x_l) \simeq g_{jkl}(x_j, x_k, x_l)$$

What to do?

(1) suppress spurious interactions in $F(\mathbf{x})$

(2) null dist. of H when *no* interactions

SURPRESS SPURIOUS INTERACTIONS

TreeNet model: $F(\mathbf{x}) = \sum_{m=1}^M a_m T_m(\mathbf{x})$

Decision tree: $T_m(\mathbf{x}) = \sum_{l=1}^L b_{lm} t_{lm}(\mathbf{x})$

Terminal node :

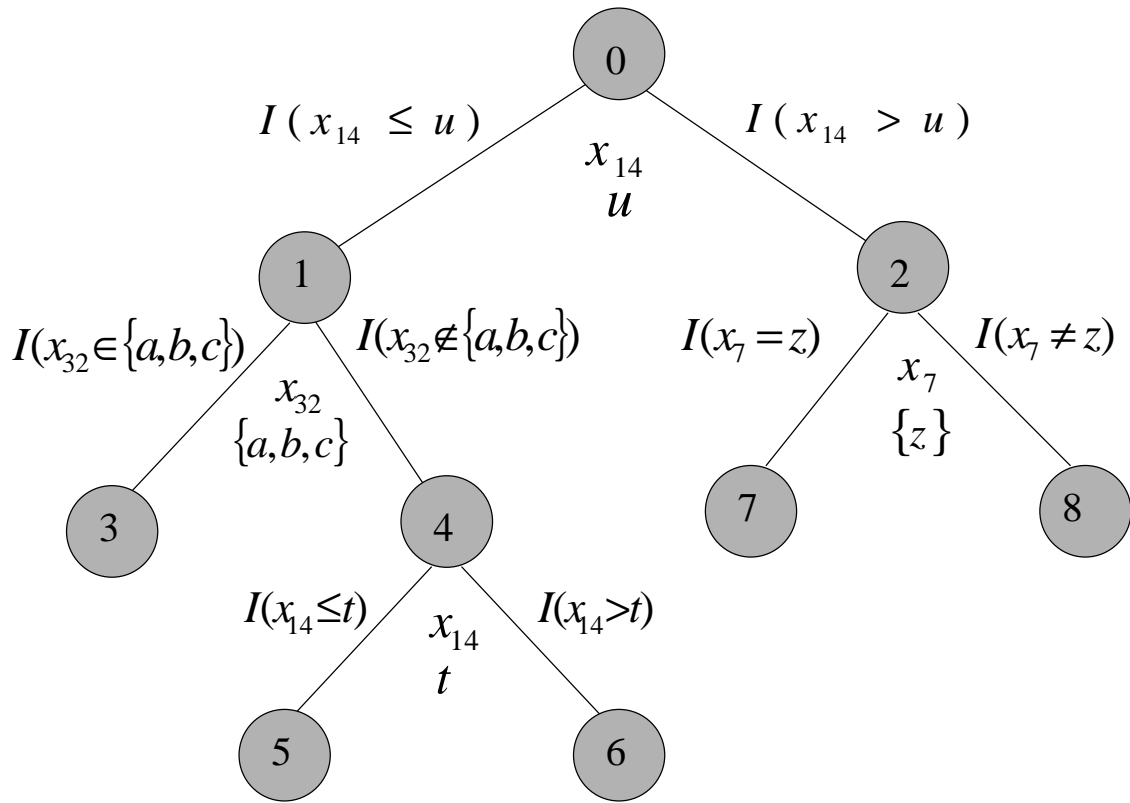
$$t_{lm}(\mathbf{x}) = \prod_{k \in \text{path}(l,m)} I(x_{j(k)} \in s_k)$$

$\text{path}(l, m) =$ path from root to term. node (l, m)

Interactions:

$t_{lm}(\mathbf{x}) \sim$ *different* variables in $\text{path}(l, m)$

$T_m(\mathbf{x}) \sim \{t_{lm}(\mathbf{x})\}; F(\mathbf{x}) \sim \{T_m(\mathbf{x})\}$



$$t_5(\mathbf{x}) = I(x_{14} \leq t) \cdot I(x_{32} \notin \{a, b, c\})$$

$$t_6(\mathbf{x}) = I(t < x_{14} \leq u) \cdot I(x_{32} \notin \{a, b, c\})$$

BUILDING TREES

Splitting at interior node l :

$$j^*(l) = \arg \max_{1 \leq j \leq n} I_j(l)$$

$I_j(l)$ = improvement splitting node l on x_j

Problem:

$$| \text{corr}(x_j, x_k) | = \text{big (locally)}$$

$$\Rightarrow I_j(l) \simeq I_k(l)$$

\Rightarrow spurious interactions

Solution:

incentive for splitting on previously used variables

$$j^*(l) = \arg \max_{1 \leq j \leq n} \kappa_j(l) \cdot I_j(l)$$

$$\kappa_j(l) = 1 \quad \text{if } j \notin \text{path}(\text{root} \rightarrow l)$$

$$\kappa_j(l) = \kappa \quad (\kappa > 1) \quad \text{otherwise}$$

κ = incentive parameter

set to largest value that does not degrade

prediction accuracy

NOTE: incentive strategy does not inhibit highly

correlated variables from jointly entering model:

(a) different paths in same tree

(b) different trees

Only inhibits them from defining interaction effects

NULL DISTRIBUTION

Distributions of H_j , H_{jk} , H_{jkl} when NO interactions

Parametric bootstrap:

Compute statistics H on series of artificial

data sets $\{T_b\}_1^B$ derived from $T = \{\mathbf{x}_i, y_i\}_1^N$

$\{H_b = H(F(T_b))\}_1^B$

= reference dist. for $H = H(F(T))$

Goal: $T_b \approx T$ with $F^*(\mathbf{x}) = \text{no interactions}$

$$T_b = \{\mathbf{x}_i, \tilde{y}_i^{(b)}\}_1^N$$

Let $F_A(\mathbf{x}) =$ closest additive approx to $F^*(\mathbf{x})$

Estimate by setting $\{t_m = \bar{L} = 2\}_1^M$ in EGP

Regression:

$$\tilde{y}_i^{(b)} = F_A(\mathbf{x}_i) + (y_{p_b(i)} - F_A(\mathbf{x}_{p_b(i)}))$$

$$\{p_b(i)\}_1^N = \text{random permutation } \{i\}_1^N$$

Classification: $\tilde{y}_i^{(b)} \in \{-1, 1\}$

$$\Pr(y_i = 1) = (1 + F_A(\mathbf{x}_i))/2$$

ILLUSTRATIONS

All examples:

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

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

$$M = 333 \text{ trees} \Rightarrow K \simeq 2000 \text{ rules}$$

+ linear terms

ARTIFICIAL DATA

$$N = 5000, n = 100$$

$$y_i = F^*(\mathbf{x}_i) + \varepsilon_i$$

$$x_{ij} \in \{k/10\}_0^9, k \sim U(\{0, \dots, 9\})$$

$$\varepsilon_i \sim N(0, \sigma^2); \text{ 2/1 signal/noise}$$

$$F^*(\mathbf{x}) = 9 \prod_{j=1}^3 \exp(-3(1 - x_j)^2) \quad [3\text{-var. int.}]$$

$$-0.8 \exp(-2(x_4 - x_5)) \quad [2\text{-var. int.}]$$

$$+2 \sin^2(\pi \cdot x_6) - 2.5(x_7 - x_8) \quad [\text{additive}]$$

Note: $\{x_j\}_9^{100} = \text{pure noise variables}$

RuleFit model: 351 terms (rules+ linear)

Relative average absolute error

Prediction (accuracy):

$$E[|y - F(\mathbf{x})|] / E[|y - \text{med}(y)|]$$

Target estimation (interpretation):

$$E[|F^*(\mathbf{x}) - F(\mathbf{x})|] / E[|F^*(\mathbf{x}) - \text{med}(F^*(\mathbf{x}))|]$$

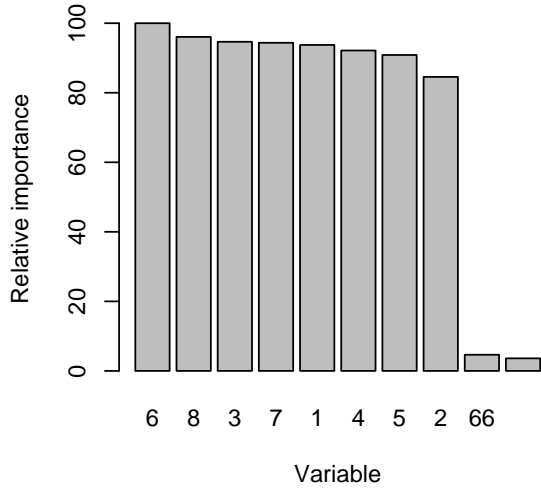
	Full	Additive	Linear
Prediction	0.49	0.61	0.69
Target estimation	0.18	0.43	0.58

Simulated example: six most important rules

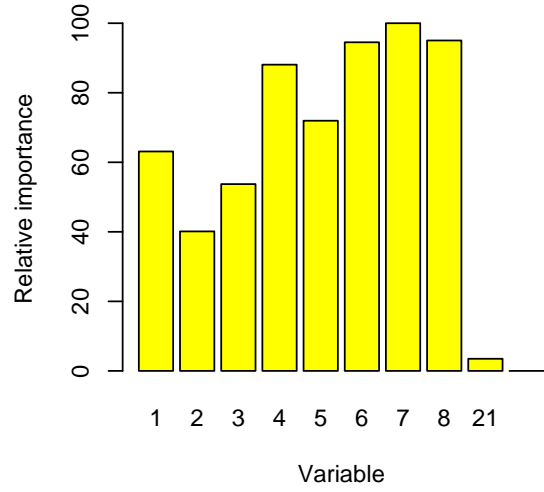
all predictions

Imp.	Coeff.	sup.	Rule
100	0.57	0.49	$0.25 \leq x_6 < 0.75$
99	0.79	0.15	$x_1 \geq 0.35 \ \& \ x_2 \geq 0.45 \ \& \ x_3 \geq 0.45$
83	-0.81		linear: x_7
63	0.61		linear: x_8
61	0.34	0.51	$0.35 \leq x_6 < 0.85$
58	-0.38	0.25	$x_4 < 0.35 \ \& \ x_5 \geq 0.45$

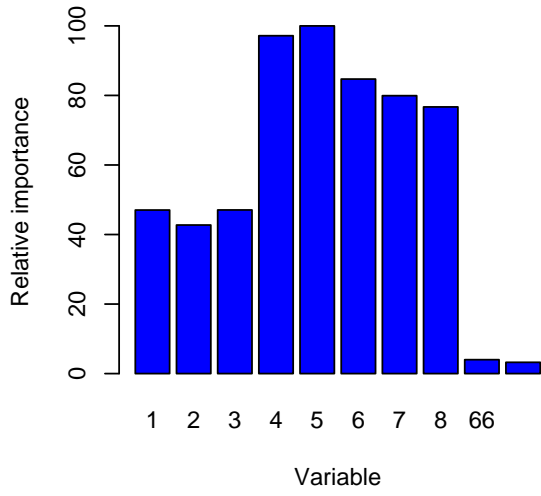
Variable importance – all



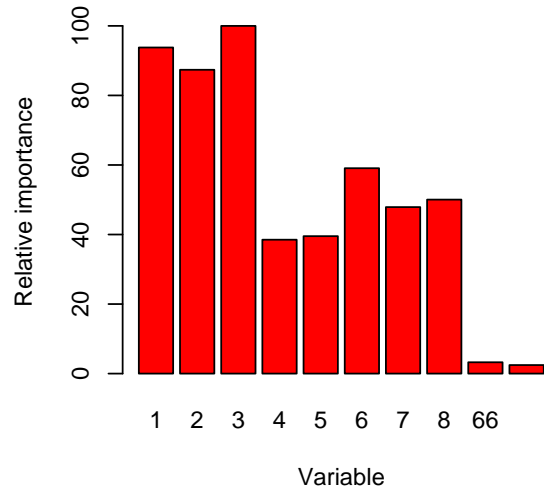
Variable importance – x = 0.5



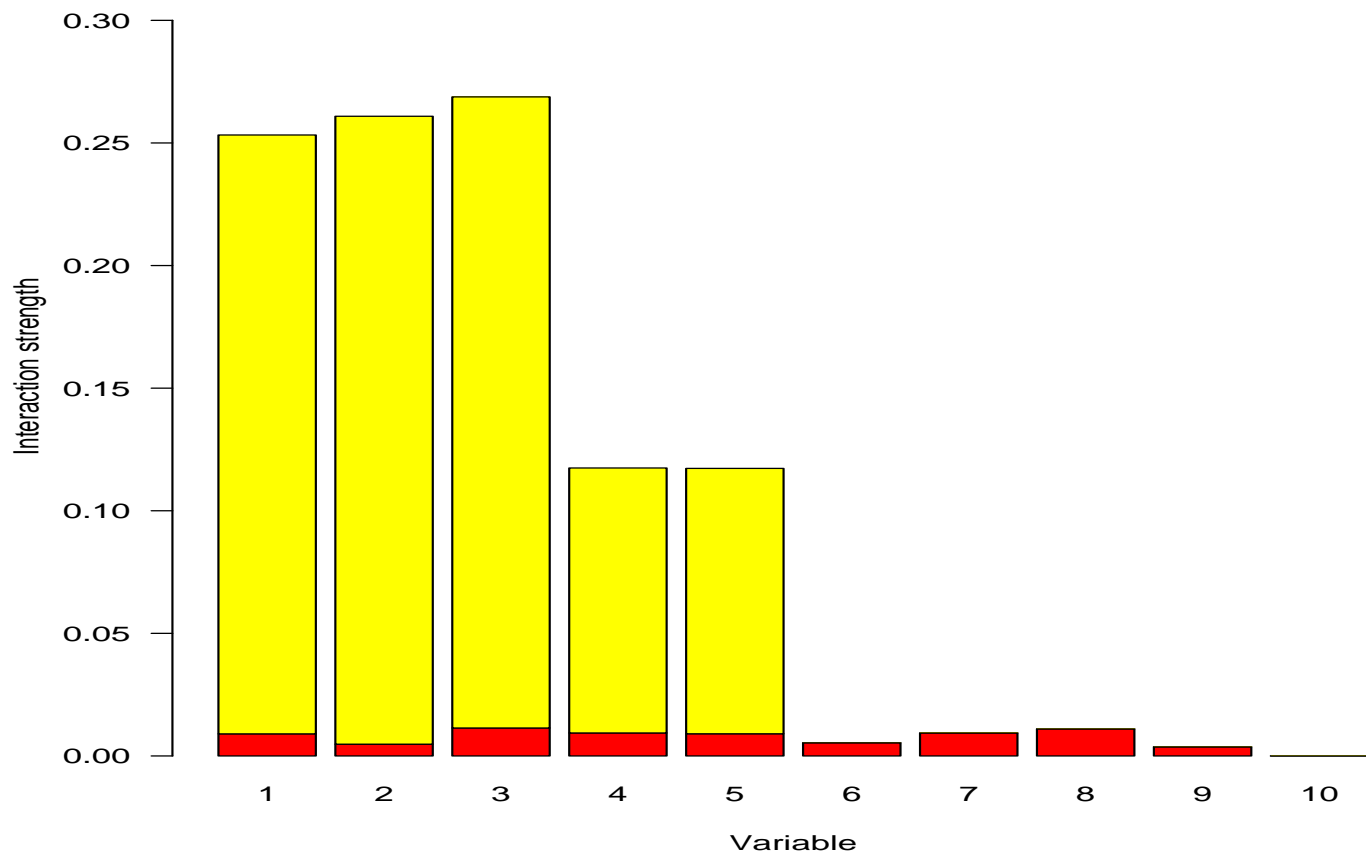
Variable importance – low



Variable importance – high



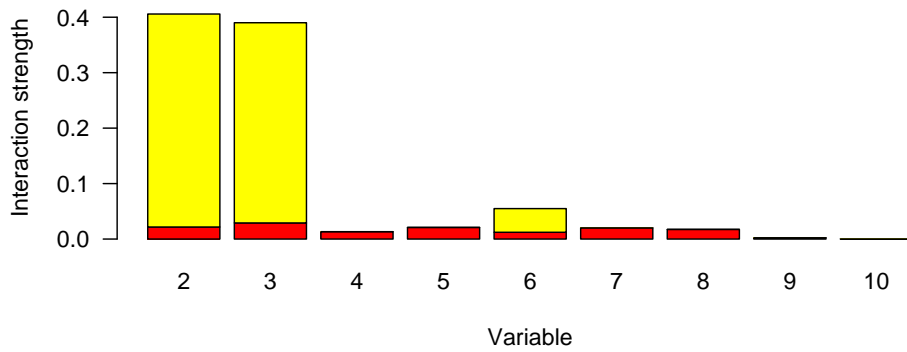
Simulated data – Interactions



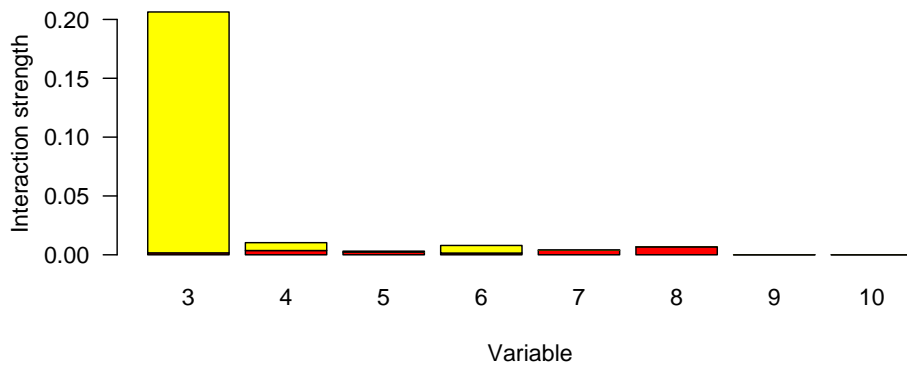
$$\tilde{H}_j = H_j - \bar{H}_j^{(0)} \text{ (yellow), } \sigma_j^{(0)} \text{ (red)}$$

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

Simulated data – interactions with x1



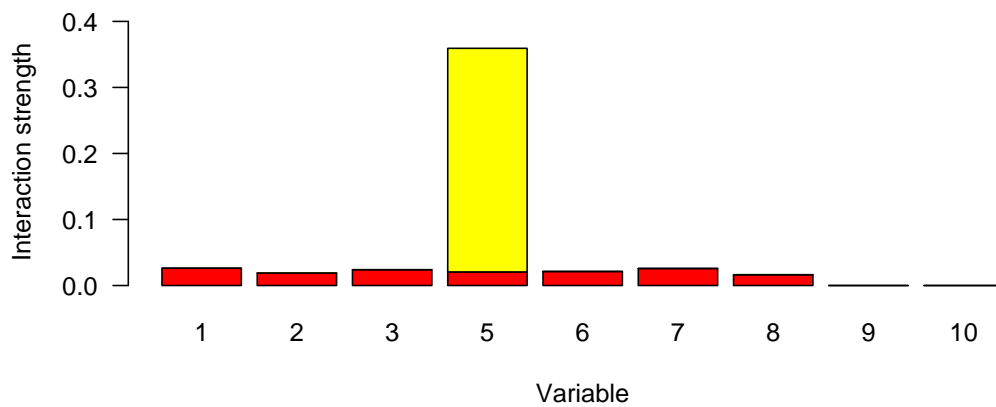
Three variable interactions with x1 & x2



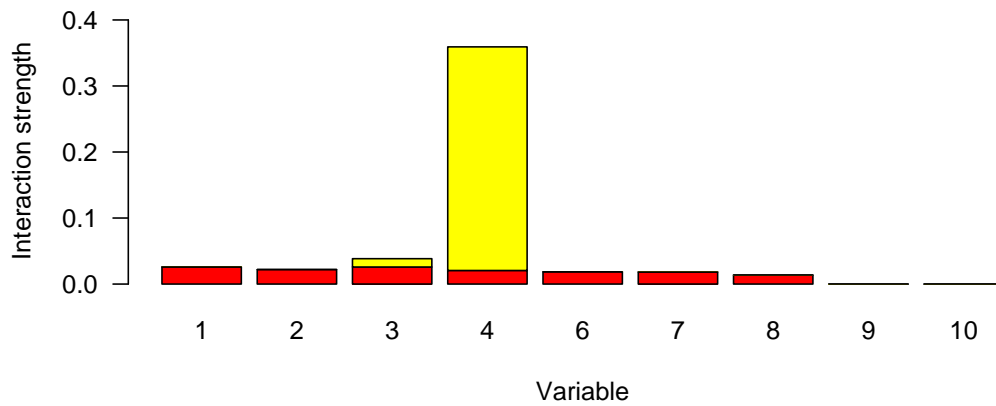
Top: $\tilde{H}_{j1} = H_{j1} - \bar{H}_{j1}^{(0)}$ (yellow), $\sigma_{j1}^{(0)}$ (red)

Bottom: $\tilde{H}_{j12} = H_{j12} - \bar{H}_{j12}^{(0)}$ (yellow), $\sigma_{j12}^{(0)}$ (red)

Simulated data – Interactions with x4



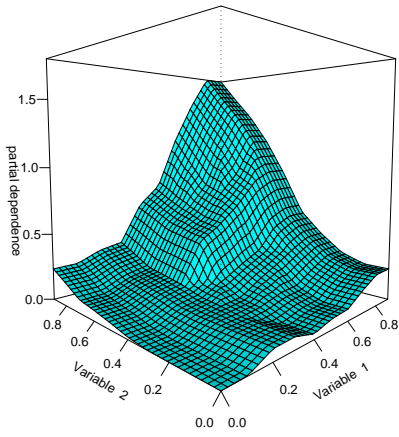
Simulated data – Interactions with x5



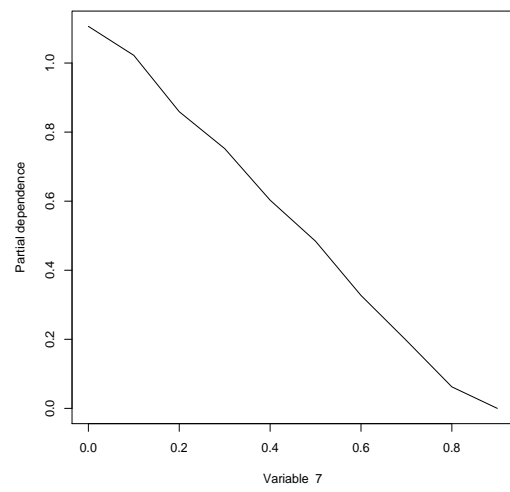
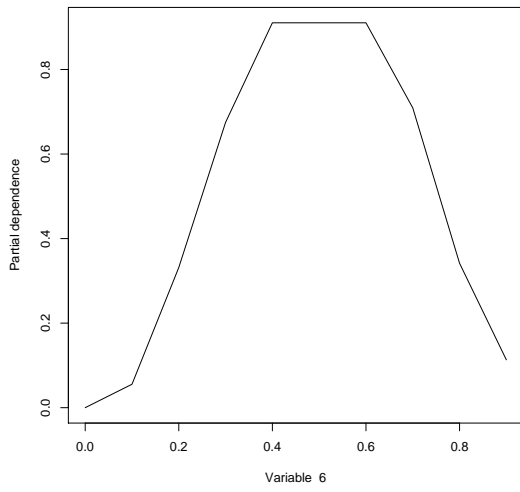
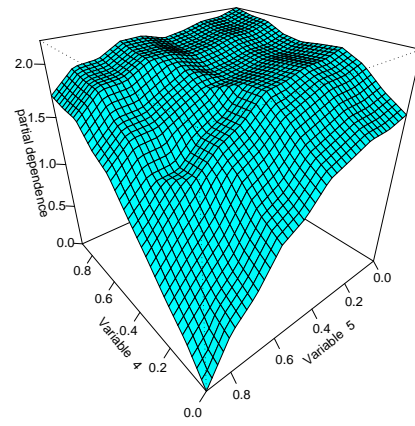
Top: $\tilde{H}_{j4} = H_{j4} - \bar{H}_{j4}^{(0)}$ (yellow), $\sigma_{j4}^{(0)}$ (red)

Bottom: $\tilde{H}_{j5} = H_{j5} - \bar{H}_{j5}^{(0)}$ (yellow), $\sigma_{j5}^{(0)}$ (red)

Partial dependence



Partial dependence



Simulated data - partial dependence plots

BOSTON HOUSING DATA

$N = 506$ neighborhoods in the Boston metropolitan area

14 summary statistics were collected in each

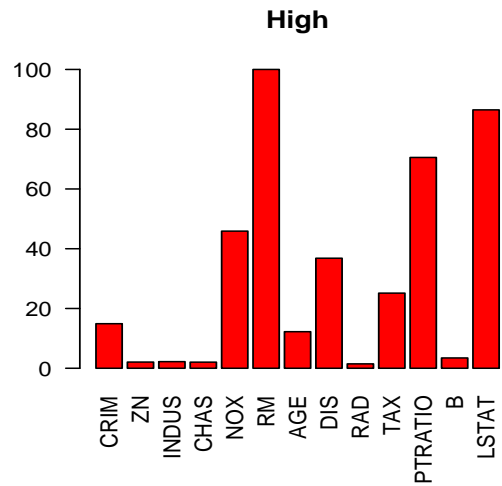
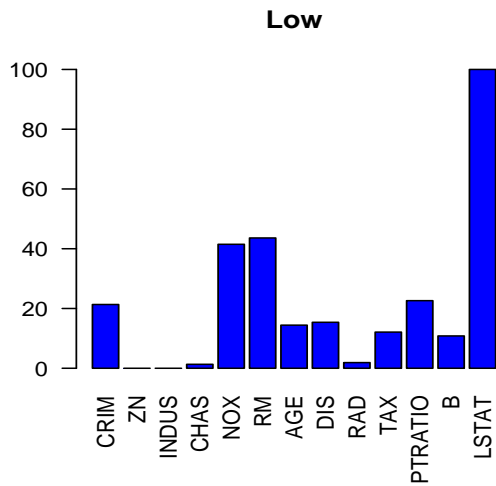
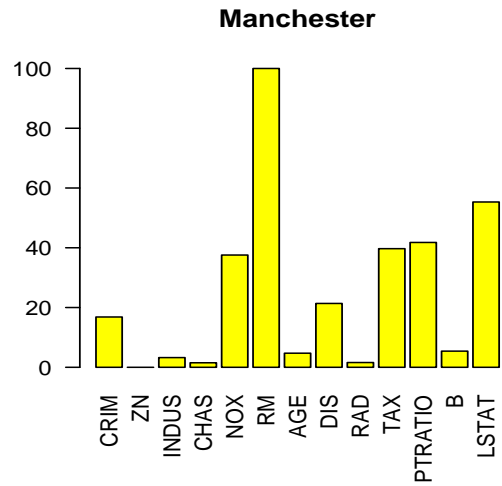
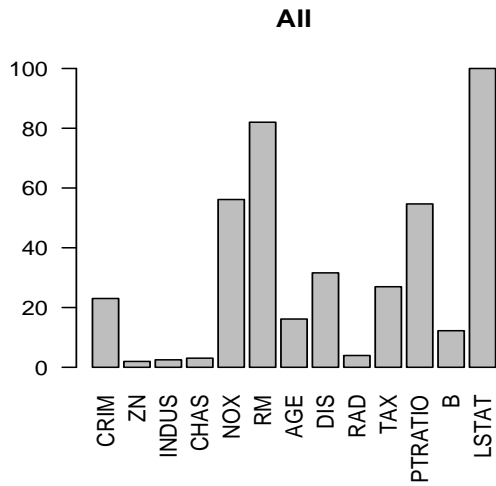
$y =$ median house value

$\mathbf{x} = 13$ other (predictor) variables

RuleFit model: 215 terms (rules+ linear)

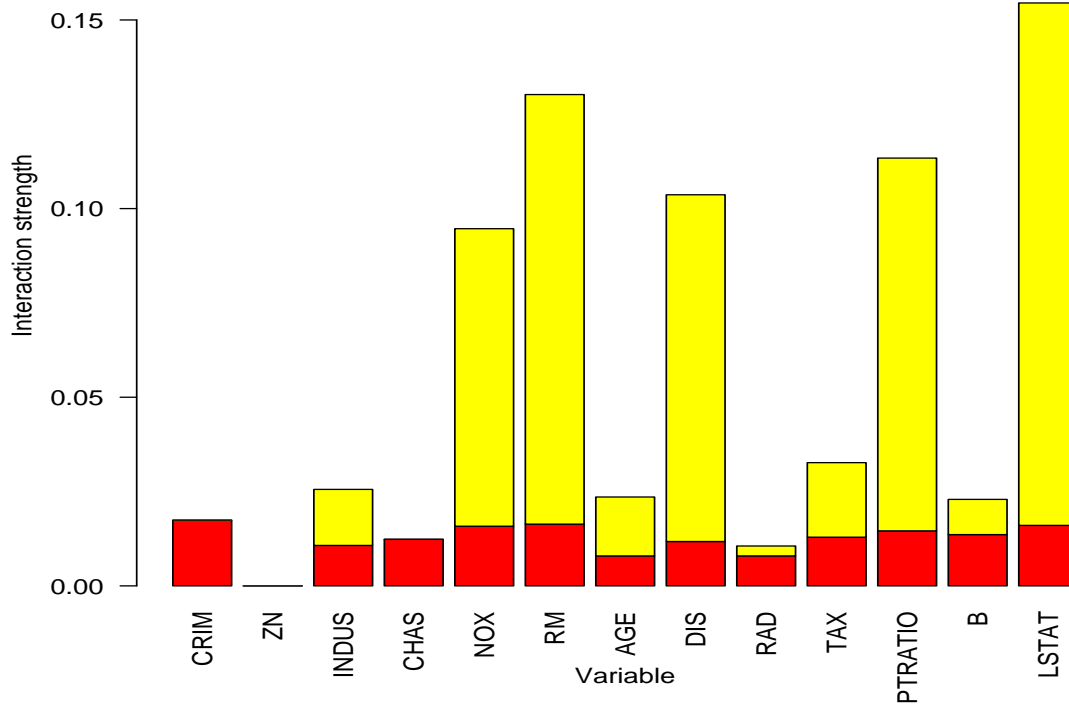
Relative average absolute error (50-fold X-val)

	Full	Additive	Linear
Prediction	0.33	0.37	0.49



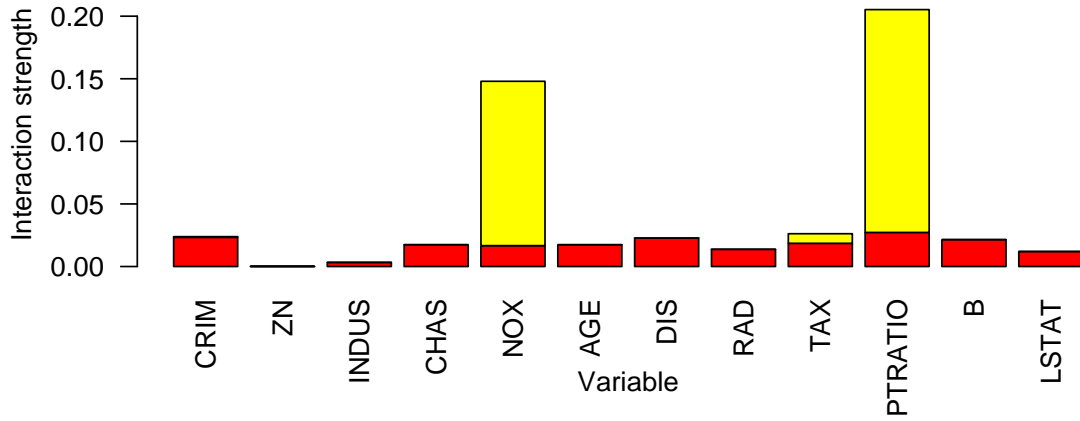
Boston housing – variable importance

Boston housing – interactions

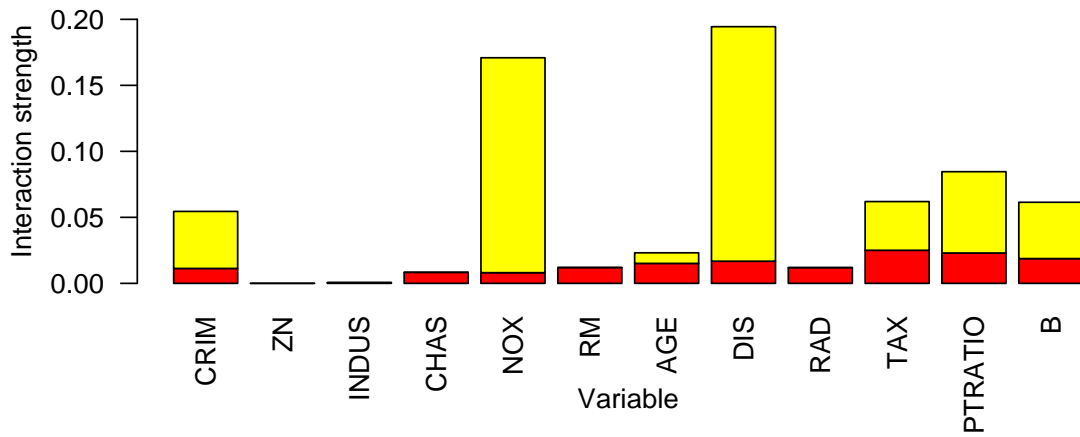


$$\tilde{H}_j = H_j - \bar{H}_j^{(0)} \text{ (yellow), } \sigma_j^{(0)} \text{ (red)}$$

Boston housing - interactions with RM

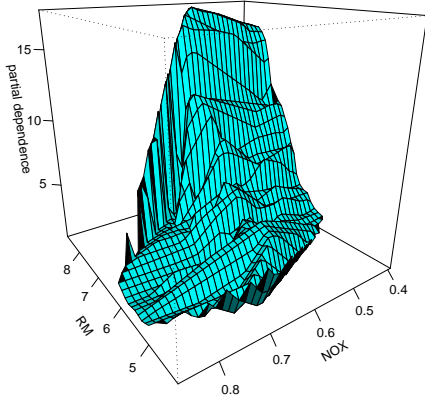


Boston housing - interactions with LSTAT

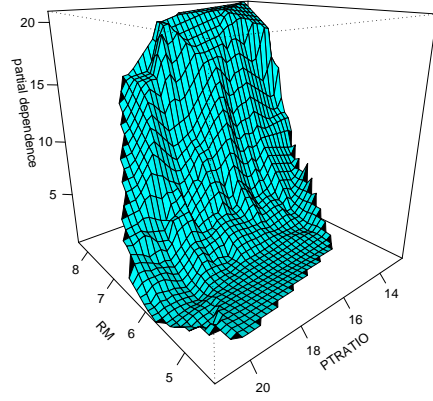


$H_{jkl} \Rightarrow$ no 3-var. interactions involving *RM* or *LSTAT*

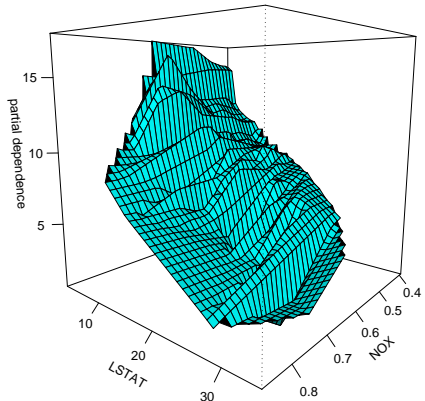
Partial dependence



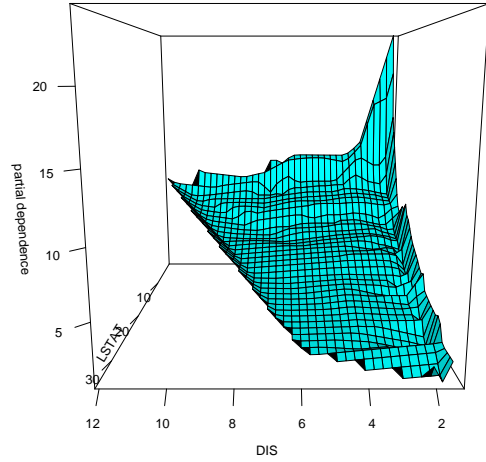
Partial dependence



Partial dependence



Partial dependence



Boston housing – partial dependence plots

ADULT CENSUS DATA

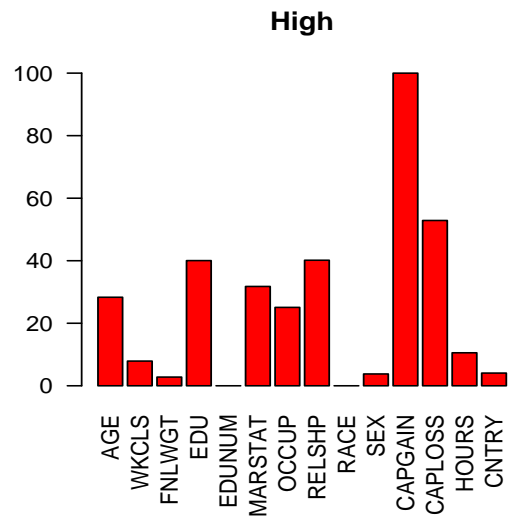
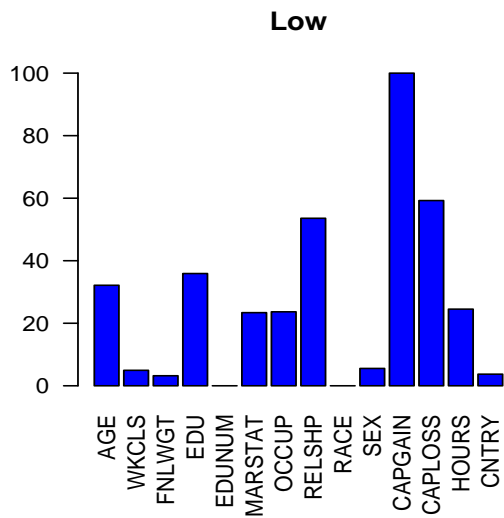
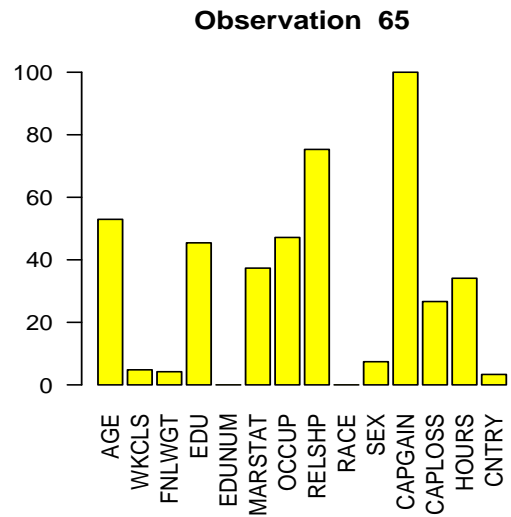
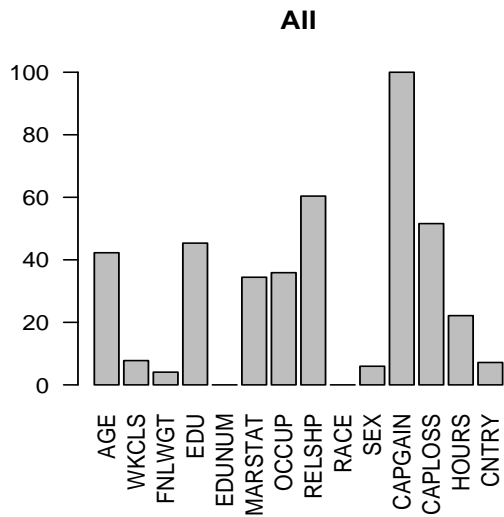
$N = 32561$ (training sample), $N_t = 16281$ (test)

x	description	Values
1	age	numeric
2	work class	cat: 8
3	final weight	numeric
4	education	numeric
5	education-num	not used
6	marital status	cat: 7
7	occupation	cat: 14
8	spouse relationship	cat: 6
9	race	cat: 5
10	sex	cat: 2
11	capital gain	numeric
12	capital loss	numeric
13	hours per week	numeric
14	native country	cat: 41
y	income > \$50000	cat: 2

RuleFit model: 214 terms (rules+ linear)

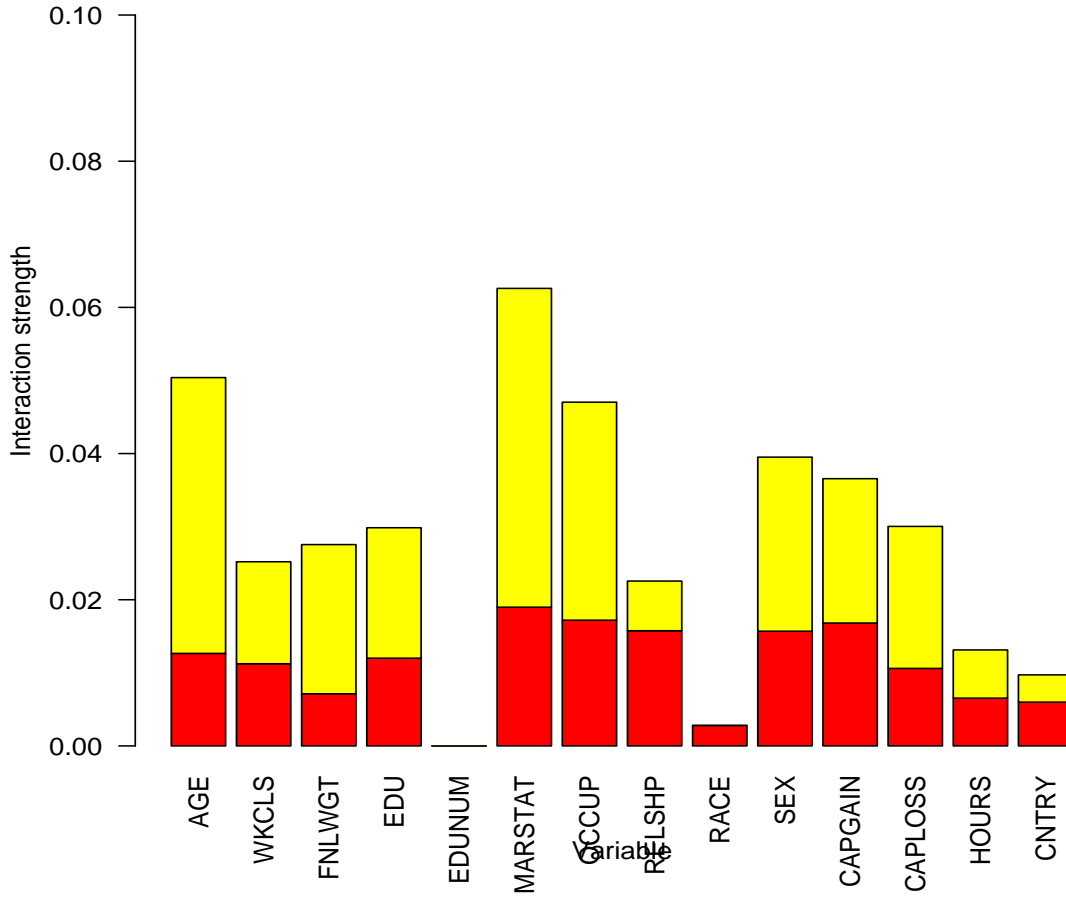
Test set error rate

	Full	Additive
Prediction	12.8%	13.6%



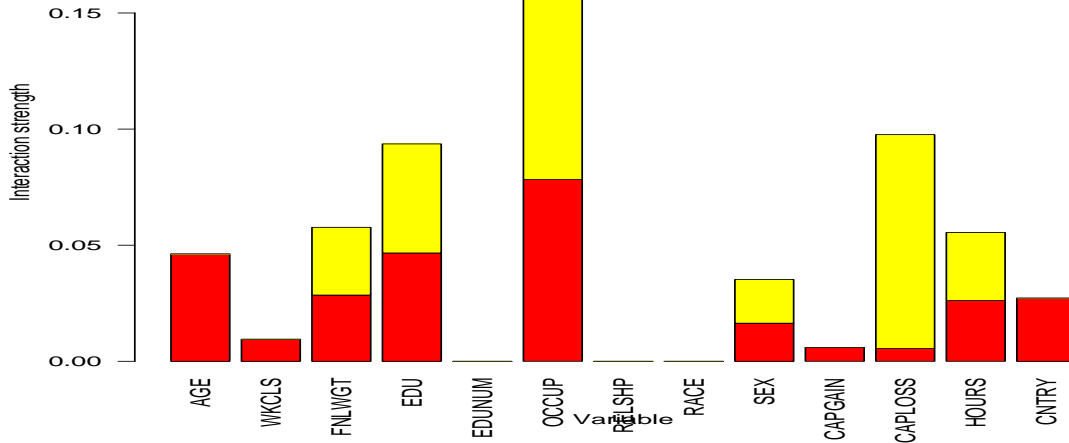
Adult census data – variable importance

Adult census – interactions

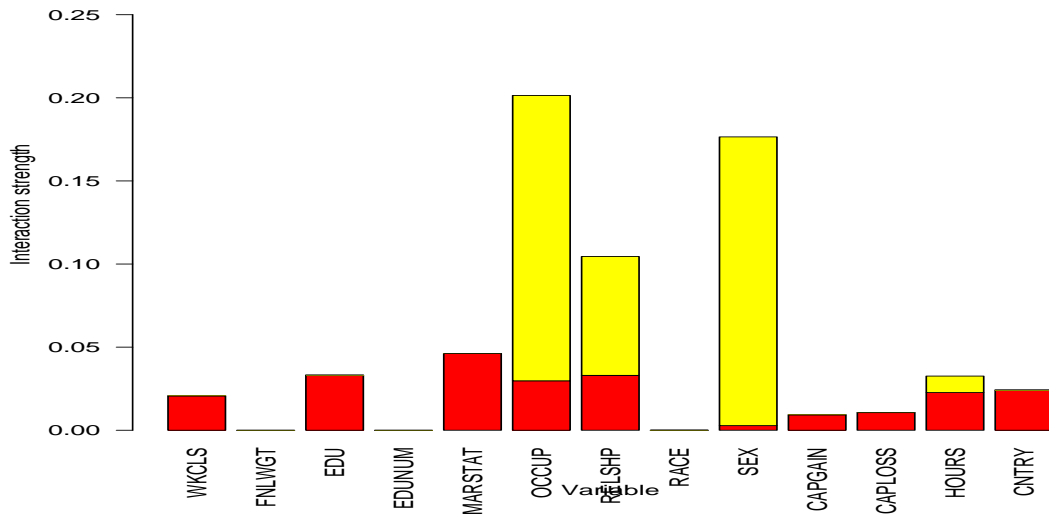


$$\tilde{H}_j = H_j - \bar{H}_j^{(0)} \text{ (yellow), } \sigma_j^{(0)} \text{ (red)}$$

Adult census – interactions with MARSTAT

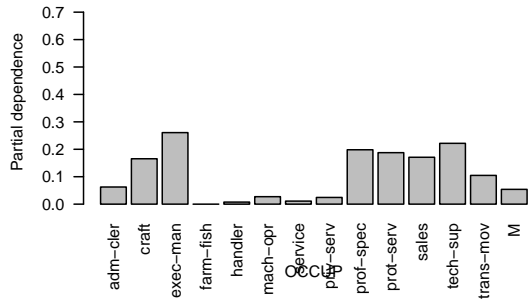


Adult census – interactions with AGE

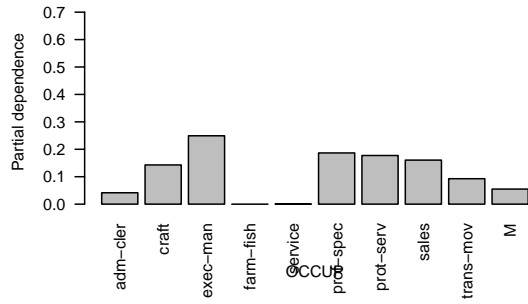


$H_{jkl} \Rightarrow$ no 3-var. inter's involving *MARSTAT* or *AGE*

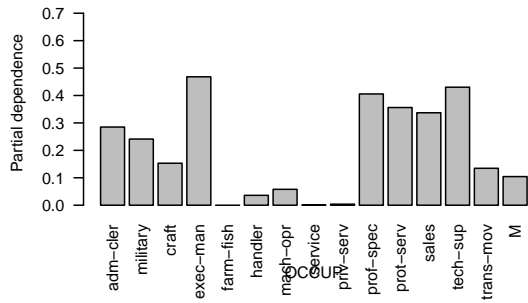
MARSTAT = divorced



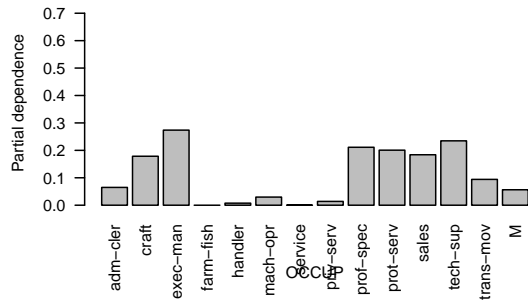
MARSTAT = military-spouse



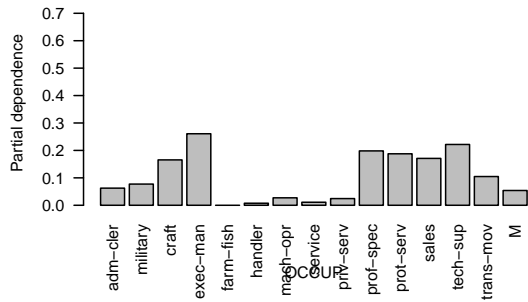
MARSTAT = civilian-spouse



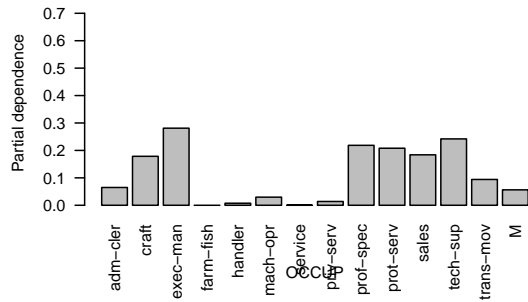
MARSTAT = absent-spouse



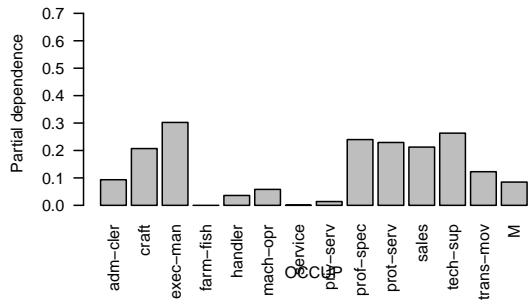
MARSTAT = never-married

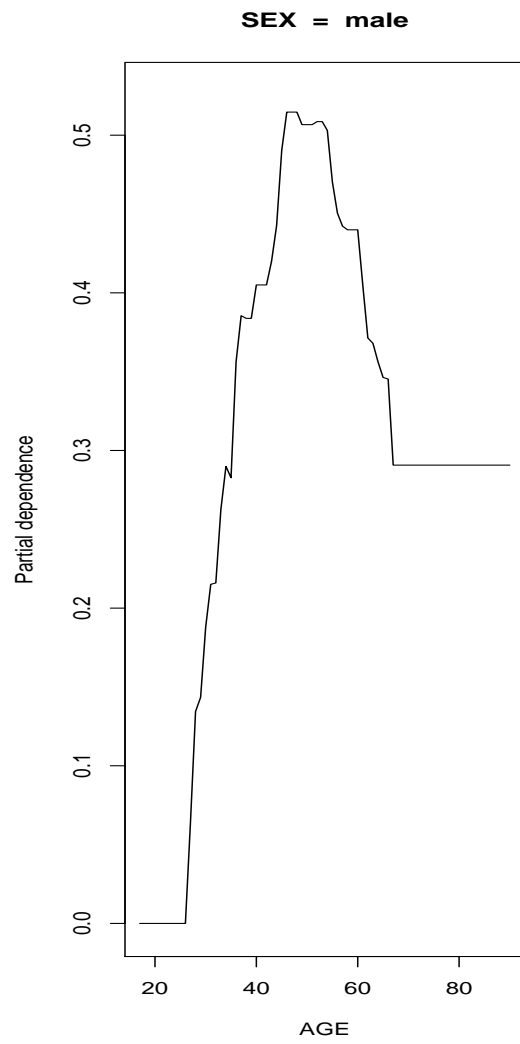
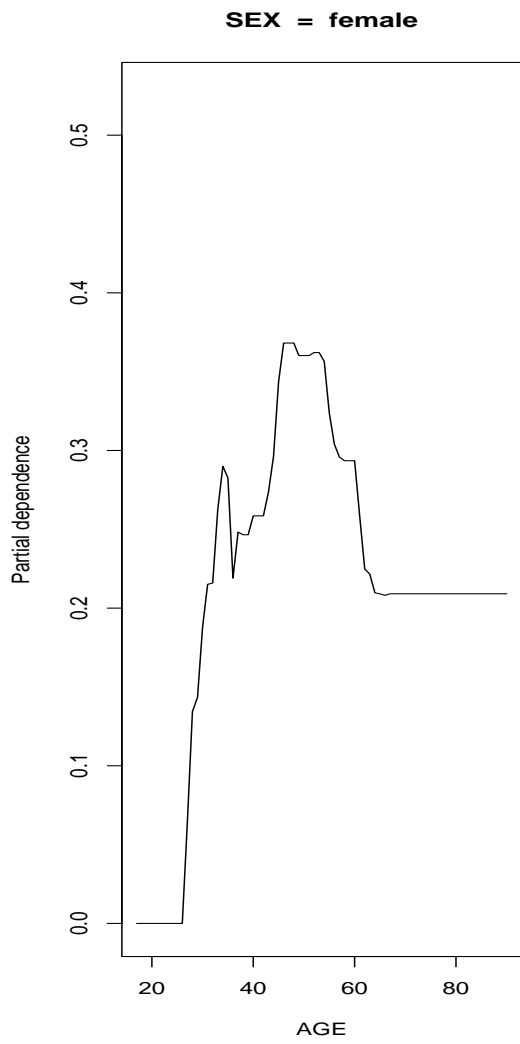


MARSTAT = separated



MARSTAT = widowed





Partial dependence on *AGE* & *SEX*

Future Work: rule summarization

Bibliography

Talk: <http://www-stat.stanford.edu/~jhf/talks/RuleFit.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*

Function ANOVA: Liu & Owen; Hooker