Data Mining for the Development of Public Policy

Greg Ridgeway

RAND Statistics Group
Santa Monica, CA
http://gregridgeway.homestead.com

Outline

- Boosting and the Generalized Boosted Model
- Public policy examples
  - Causal modeling: Effectiveness of drug treatment programs. Adjust treatment effect estimates for selection bias
  - Least squares: Cost of rehabilitation. Predict rehabilitation cost from patient functional status and diagnosis
  - Classification: High school dropouts. Estimate risk of high school dropout at age 13

Gradient Boosting: Least squares regression

$$
\begin{aligned}
J(f) &= \mathrm{E}\left(y - f(\mathbf{x})\right)^2 \\
&\approx \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i)\right)^2
\end{aligned}
$$

Find a g that offers a decrease in squared error.

$$
\hat{J}(\hat{f} + g) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \left(\hat{f}(\mathbf{x}_i) + g(\mathbf{x}_i)\right)\right)^2
$$

Clearly this shows that $g(\mathbf{x}_i)$ should be a least squares predictor of the residual $z_i = y_i - \hat{f}(\mathbf{x}_i)$.

Gradient Boosting: Least squares regression

1. Initialize $\hat{f}(\mathbf{x}) = \bar{y}$.
2. Let $z_i = y_i - \hat{f}(\mathbf{x}_i)$.
3. Construct a least squares predictor of the residuals, $g(\mathbf{x})$ using a tree-structured regressor
4. Update our guess as

$$\hat{f}(\mathbf{x}) \leftarrow \hat{f}(\mathbf{x}) + g(\mathbf{x})$$

5. Return to step (2) for T iterations

LogitBoost: Logistic regression

1. Initialize $\hat{f}(\mathbf{x}) = \log \frac{\bar{y}}{1-\bar{y}}$
2. Let $z_i = y_i - \frac{1}{1-\exp\left(-\hat{f}(\mathbf{x}_i)\right)}$
3. Construct a tree structured predictor of z_i
4. The tree assigns each observation to a terminal node $g(T_k) = \arg\max_\lambda \sum_{i \in T_k} L(y_i, \hat{f}(\mathbf{x}_i) + \lambda)$
5. Update our guess as

$$\hat{f}(\mathbf{x}) \leftarrow \hat{f}(\mathbf{x}) + g(\mathbf{x})$$

6. Return to step (2) for T iterations

LogitBoost: Logistic regression

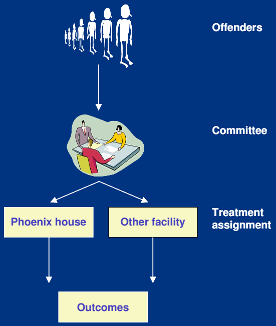
1. Initialize $\hat{f}(\mathbf{x}) = \log \frac{\bar{y}}{1-\bar{y}}$
2. Let $z_i = y_i - \frac{1}{1-\exp\left(-\hat{f}(\mathbf{x}_i)\right)}$
3. Construct a tree structured predictor of z_i
4. The tree assigns each observation to a terminal node $g(T_k) = \arg\max_\lambda \sum_{i \in T_k} y_i(\hat{f}(\mathbf{x}_i) + \lambda) - \log\left(1 + \exp(\hat{f}(\mathbf{x}_i) + \lambda)\right)$
5. Update our guess as
$$\hat{f}(\mathbf{x}) \leftarrow \hat{f}(\mathbf{x}) + g(\mathbf{x})$$
6. Return to step (2) for T iterations

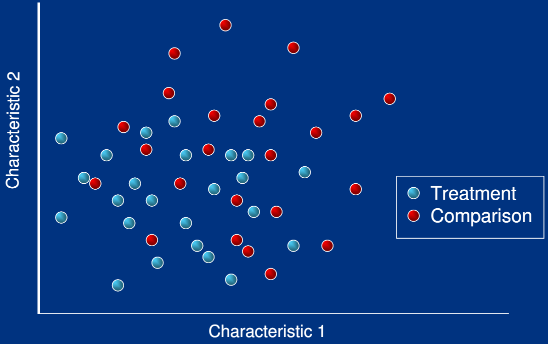
Advantages

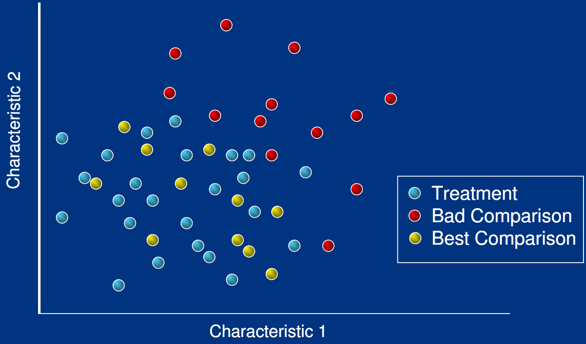
1. Boosting has a straightforward application to most prediction problems and loss functions
2. Trees handle continuous, nominal, ordinal, and missing x's
3. Invariant to one-to-one transformations of the x's
4. Model higher interaction terms with more complex regression trees
5. Use low variance models on each iteration: shrinkage, subsampling, bagging
6. Automate the selection of the number of iterations: out-of-bag estimation

Example: Effect of drug treatment

- The treatment assignments are non-random
- We want to estimate treatment effect
- We can reweight the individuals from the other facility to look like those from the Phoenix house




Characteristic 2
Characteristic 1
Treatment
Comparison


Characteristic 2
Characteristic 1
Treatment
Bad Comparison
Best Comparison

Drug treatment: Propensity scores

- Each individual has a y_c and a y_t, the outcome that would happen if they went to the control or treatment facility

$$\begin{aligned} \mathrm{E}(y_t|t) &\approx \frac{\sum_{i\in T} y_{ti}}{N_T} \\ \mathrm{E}(y_c|t) &= \iint y_c f(y_c, x|t)\, dx\, dy_c \end{aligned}$$

Drug treatment: Propensity scores

- Each individual has a y_c and a y_t, the outcome that would happen if they went to the control or treatment facility

$$
\begin{aligned}
\mathrm{E}(y_c|t) &= \iint y_c f(y_c, x|t)\, dx\, dy_c \\
&= \iint y_c \frac{f(y_c, x|t)}{f(y_c, x|c)} f(y_c, x|c)\, dx\, dy_c
\end{aligned}
$$

Drug treatment: Propensity scores

- Each individual has a y_c and a y_t, the outcome that would happen if they went to the control or treatment facility

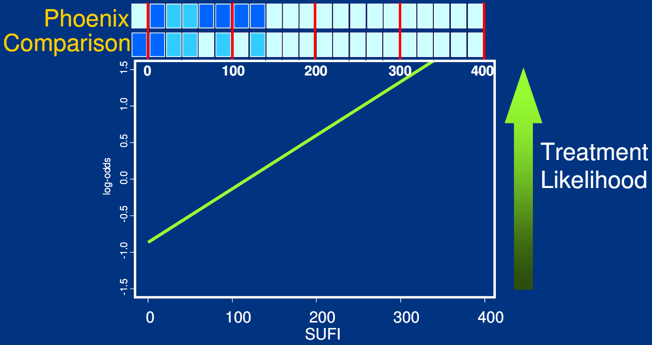
$$\begin{aligned} \mathrm{E}(y_c|t) &= \frac{f(c)}{f(t)} \iint y_c \frac{f(t|x)}{1 - f(t|x)} f(y_c, x|c)\, dx\, dy_c \\ &\approx \frac{\sum_{i \in C} w_i y_{ci}}{\sum_{i \in C} w_i} \end{aligned}$$

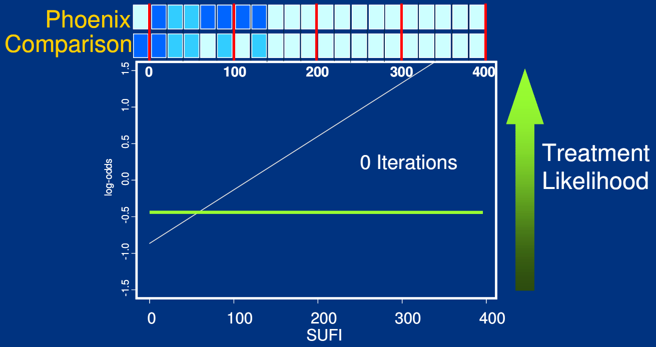
Summary of the method

$$\mathrm{E}(y_t|t) \approx \frac{\sum_{i \in T} y_{ti}}{N_T}, \mathrm{E}(y_c|t) \approx \frac{\sum_{i \in C} w_i y_{ci}}{\sum_{i \in C} w_i}$$

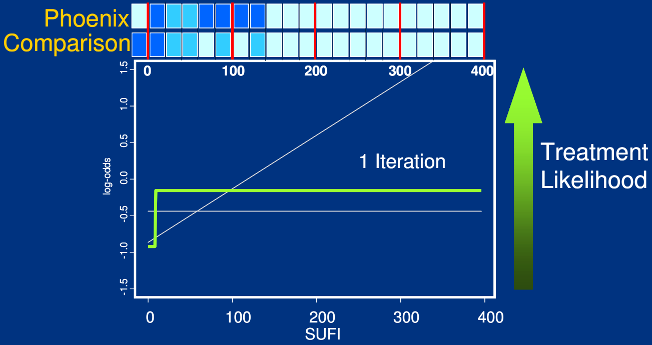
- $w_i = \frac{p_i}{1-p_i}$, and p_i is the probability that subject i goes to the treatment group
- Derivation requires that treatment assignments depend only on x
- x is high-dimensional (56) and we use the boosted logistic regression method to estimate the probabilities

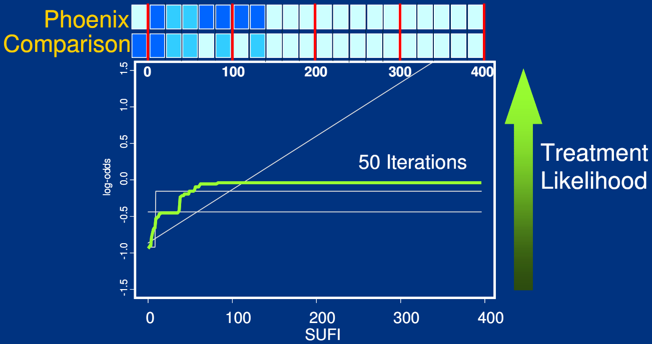
Algorithm progression

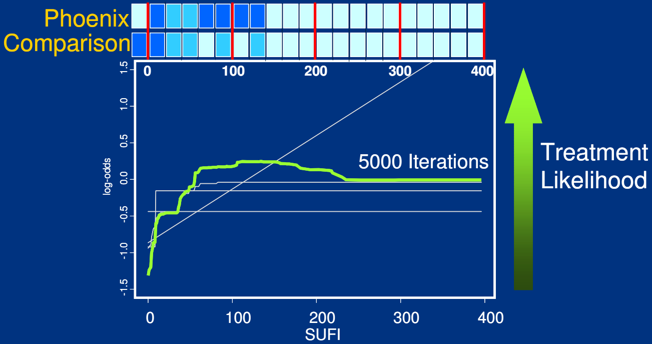




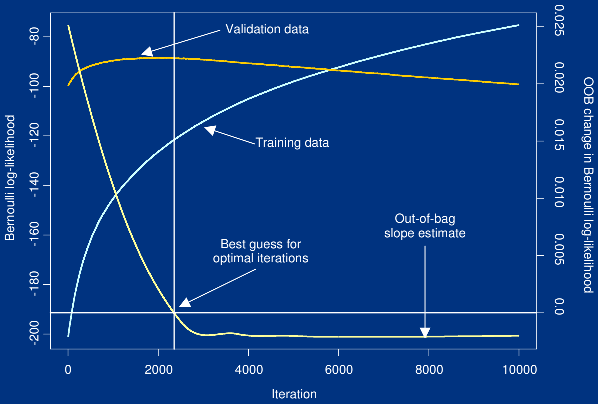
Phoenix
Comparison
0
100
200
300
400
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
log-odds
0 Iterations
0
100
200
300
400
SUFI
Treatment
Likelihood


Phoenix
Comparison
0
100
200
300
400
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
log-odds
1 Iteration
0
100
200
300
400
SUFI
Treatment
Likelihood

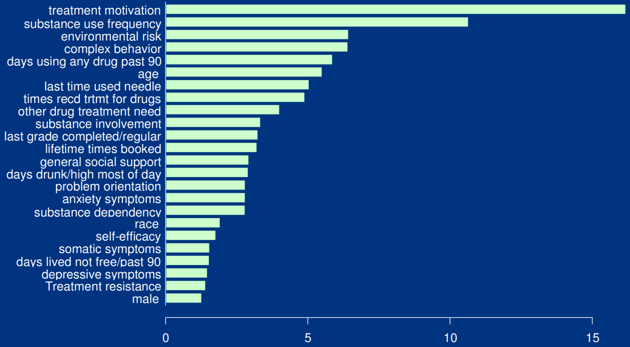

Phoenix
Comparison
0
100
200
300
400
50 Iterations
log-odds
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
SUFI
Treatment
Likelihood


Phoenix
Comparison
0
100
200
300
400
log-odds
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
5000 Iterations
SUFI
Treatment
Likelihood

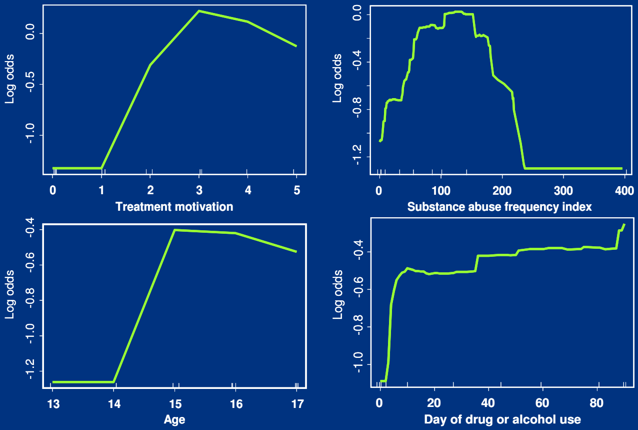
Estimating the optimal number of iterations



Relative influence



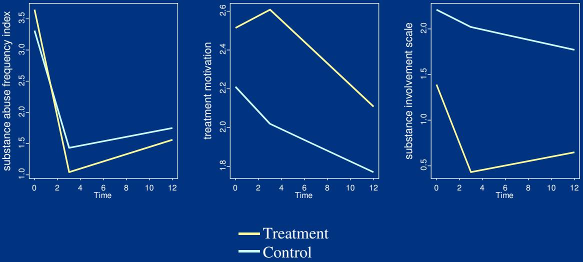
Marginal effects



Balance of subject features

Variable	treatment mean	weighted control mean	unweighted control mean	t
treatment motivation	2.52	2.22	1.35	1.84
environmental risk	30.61	30.68	28.94	-0.07
substance abuse	76.85	67.59	43.34	1.16
complex behavior	12.84	12.77	12.11	0.07
age	15.82	15.77	15.31	0.45
l5a124	0.62	0.55	0.38	1.13
withdrawal index	2.42	2.34	2.27	0.75
days in detention	44.37	52.37	54.11	-0.74
substance problem	9.91	9.26	6.64	1.27
age of first use	12.55	12.27	11.97	1.04
ESS	175	106	274	

Results



substance abuse frequency index
1.0
1.5
2.0
2.5
3.0
3.5
0
2
4
6
8
10
12
Time
treatment motivation
1.8
2.0
2.2
2.4
2.6
0
2
4
6
8
10
12
Time
substance involvement scale
0.5
1.0
1.5
2.0
0
2
4
6
8
10
12
Time
Treatment
Control

Example: Prospective payment system

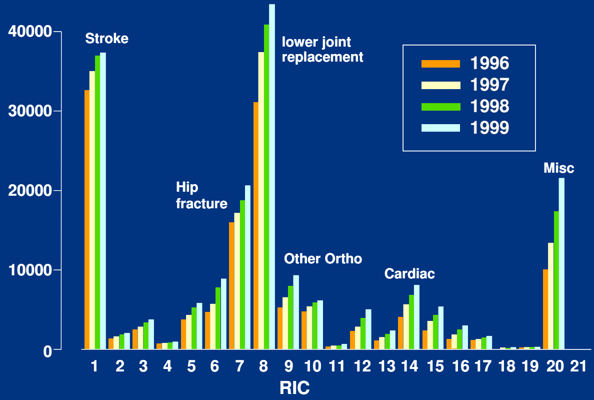
Balanced Budget Act of 1997

- Centers for Medicare & Medicaid Services (CMS) must implement a Prospective Payment System for inpatient rehabilitation
- The system should be based on a new severity-level classification of cases

Medicare data from 1996-1999

- hospital reported costs,
- patient disease and functional status data,
- hospital level data,
- and modeled the cost to rehabilitate patients.

Patients seek rehabilitation for an assortment of impairments



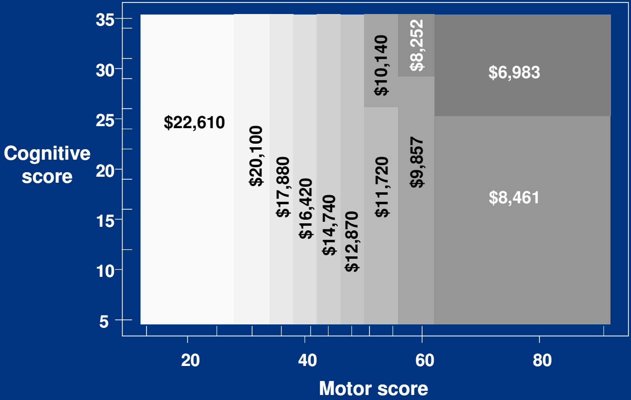
Payment system model

The prospective payment system has the form

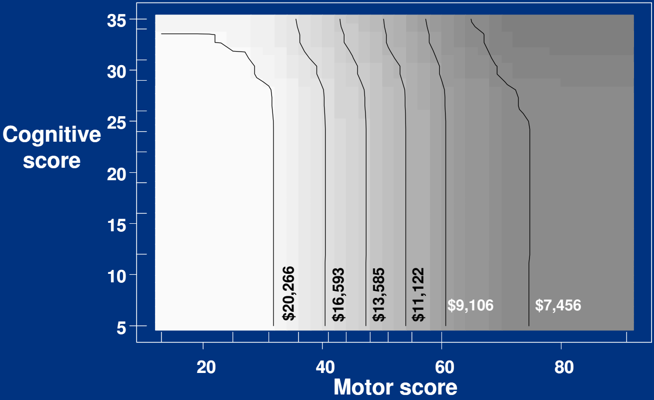
$$\text{payment}_{ij} = M \times F_j \times w(\text{age}_i, \text{motor}_i, \text{cognitive}_i) \times c_i \times a_i$$

- w is the main focus of this discussion
- M is a fixed budget normalizing constant
- F_j is a facility level adjustment for wages
- c_i is an adjustment for comorbidities
- a_i is an adjustment for transfer and outlier cases

CART costs predictions



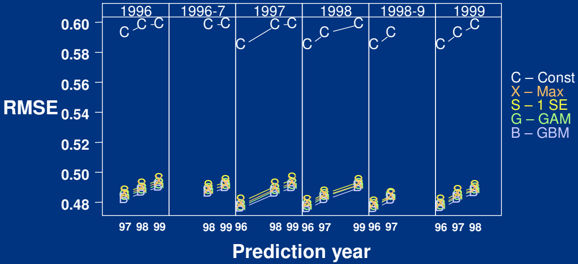
Boosting cost predictions



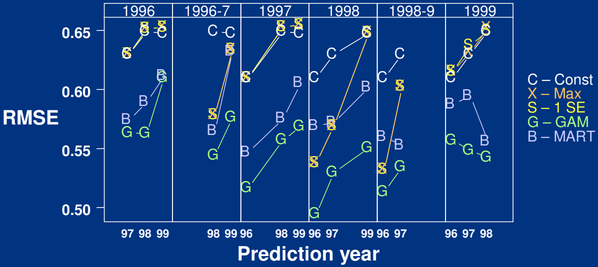
Aggregate Performance of the Various Methods, R^2

Fit Year	Evaluation Year	Const	CART Max	CART 1 SE	GAM	GBM
96	97	0.16	0.35	0.34	0.36	0.36
	98	0.15	0.33	0.33	0.35	0.35
	99	0.15	0.32	0.32	0.33	0.33
97	96	0.17	0.36	0.35	0.37	0.37
	98	0.15	0.34	0.33	0.35	0.35
	99	0.15	0.32	0.32	0.34	0.34
98	96	0.17	0.36	0.35	0.37	0.37
	97	0.16	0.35	0.34	0.36	0.36
	99	0.15	0.33	0.32	0.34	0.34
99	96	0.17	0.36	0.35	0.37	0.37
	97	0.16	0.35	0.34	0.36	0.36
	98	0.15	0.34	0.33	0.35	0.35
96-97	98	0.15	0.34	0.33	0.35	0.35
	99	0.15	0.33	0.32	0.34	0.34
98-99	96	0.17	0.36	0.36	0.37	0.37
	97	0.16	0.35	0.35	0.36	0.36

Explainable Variation for Stroke



Explainable Variation for Burns



CART and Boosting Pay Hospitals Similarly

Hospital Payment Ratio	Percent of Hospitals (Case Weighted) 1996	1997	1998	1999
90	0.0	0.0	0.0	0.0
94	0.1	0.0	0.0	0.0
95	0.0	0.0	0.2	0.1
96	0.4	0.3	0.3	1.1
97	2.5	2.5	2.4	2.1
98	11.6	9.8	11.9	8.0
99	21.6	25.8	21.8	24.6
100	28.9	28.4	30.9	29.7
101	22.3	22.0	21.5	24.3
102	9.7	8.3	8.3	7.0
103	1.9	2.7	2.2	2.4
104	0.8	0.2	0.5	0.6
105	0.1	0.1	0.1	0.2
106	0.2	0.0	0.0	0.0
107	0.0	0.0	0.0	0.0
Total	100.0	100.0	100.0	100.0

Summary

- Boosting methods offer flexible modeling strategies when faced with
  - many features,
  - features of different types,
  - redundant features
- Such situations are the norm in public policy applications
- Public policy is a ripe area for data mining applications