Exploratory Data Mining for a Single Outcome in Clinical Research

Ross Jacobucci

University of Notre Dame

Overview

Definition

Terminology

Exploratory Data Mining

Terminology is attributed to McArdle & Ritschard (2013)

Confirmatory Data Mining?

Not really. The explicit reason for "exploratory" is to make sure researchers understand context.

Other Buzzwords

1. Data Mining
2. Statistical Learning
3. Machine Learning
4. Deep Learning – Generally refers to Neural Networks

Data Mining vs. Statistics

So, what is different?

Traditional Statistics is mostly concerned with:

1. High power in small samples
2. The use of distributions
3. Low bias under certain conditions
4. Efficient use of a small number of variables

Data Mining:

1. Ratio of predictors to sample size is smaller (P can be $\geq$ N)
2. Capture nonlinearity and interactions
3. All with more focus on out of sample generalizability (low variance)

Principles

The goal of EDM isn't to throw data into an algorithm for it to "uncover" something. This is a dangerous idea.

Instead, we use EDM to pose new types of questions. Our questions act as guides to use various algorithms to maximize the information we can glean from the dataset.

Understanding what methods are available and what they can do $\rightarrow$ Researchers thinking of new types of questions

Algorithms

When I conduct analyses, I usually include algorithms of varying degrees of complexity/flexibility. These are three that do a pretty good job of spanning the continuum.

1. Regularized Regression
2. Decision Trees
3. Ensembles – Boosting & Random Forests

Regularized Regression

Regularized Regression

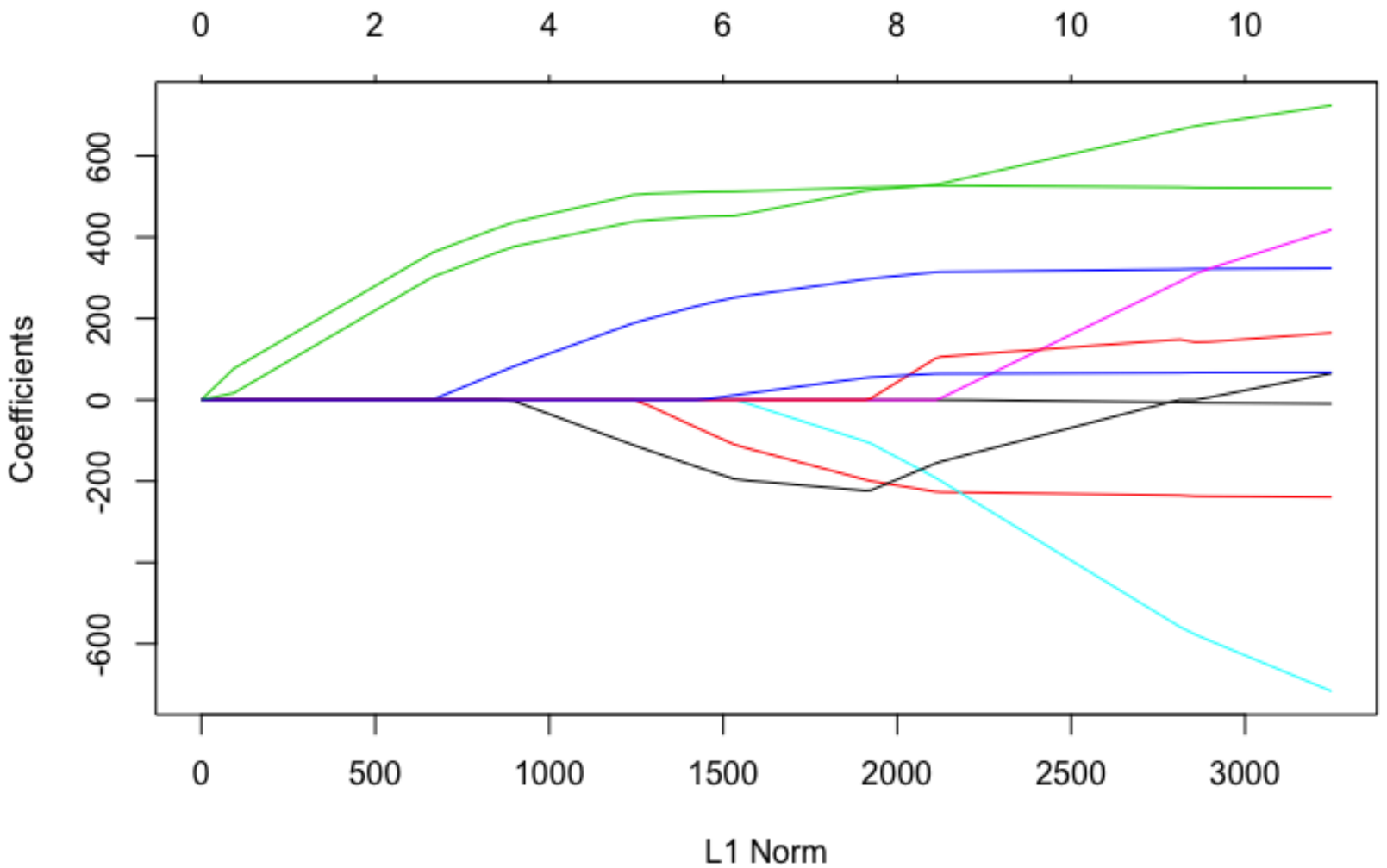
$$Ridge: \hat{\beta}^{ridge} = argmin\Big\{ \overbrace{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2}^{OLS} + \overbrace{\lambda \sum_{j=1}^{p} \beta_j^2}^{penalty} \Big\} \quad (1)$$

$$Lasso: \hat{\beta}^{lasso} = argmin\Big\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\}. \quad (2)$$

λ is the shrinkage parameter

- λ controls the size of the coefficients
- λ controls the amount of **regularization**
- As $\lambda \downarrow 0$ least squares solution
- As $\lambda \uparrow \infty$ least squares solution

Example Lasso Paths



Benefits of Regularized Regression

Lasso Regression

- Performs variable selection (Coefficients $\rightarrow 0$)
- More principled than stepwise regression
- Estimate models with $N > P$

Ridge Regression

- Handles multicollinearity
- Estimate models with $N > P$
- Less bias with large coefficients

Elastic Net Regression

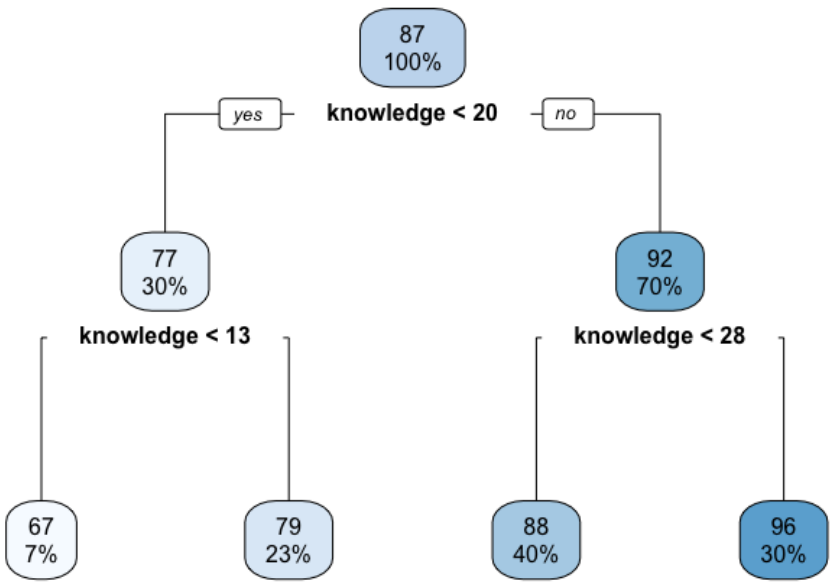
- Combines benefits of both

All can be run using the *glmnet* package in R for both continuous and categorical outcomes.

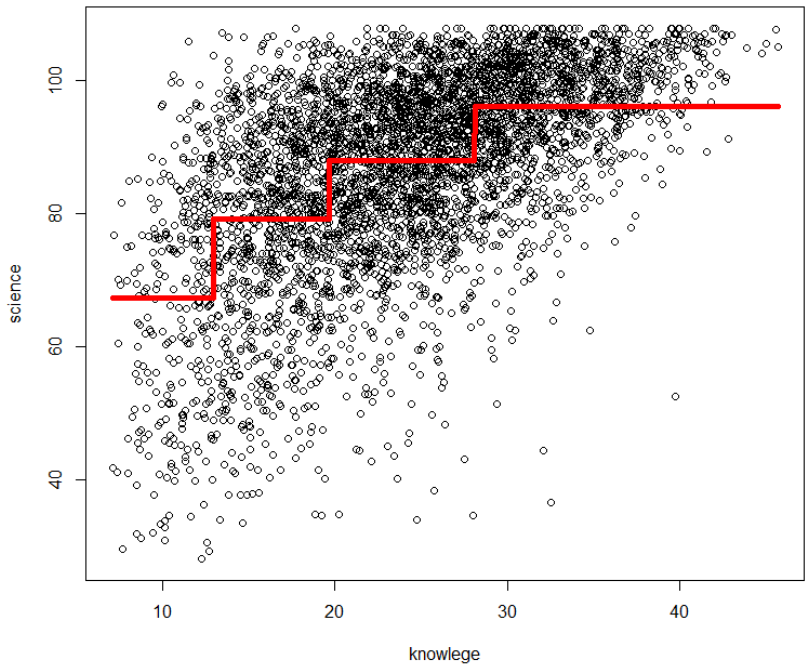
Decision Trees

Example

With one variable, Decision Trees uses a step-wise function to create predictions.



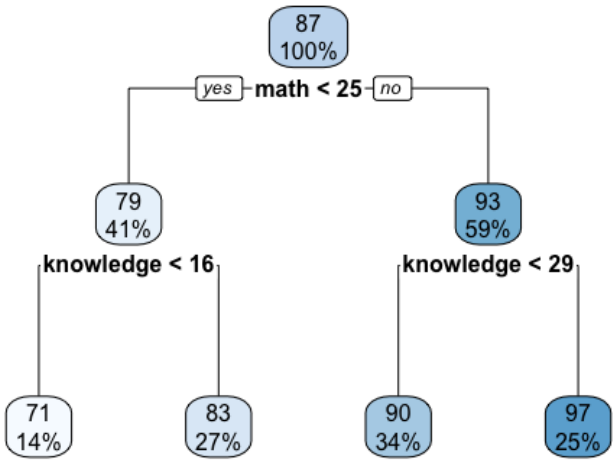
(a) Single Predictor Tree



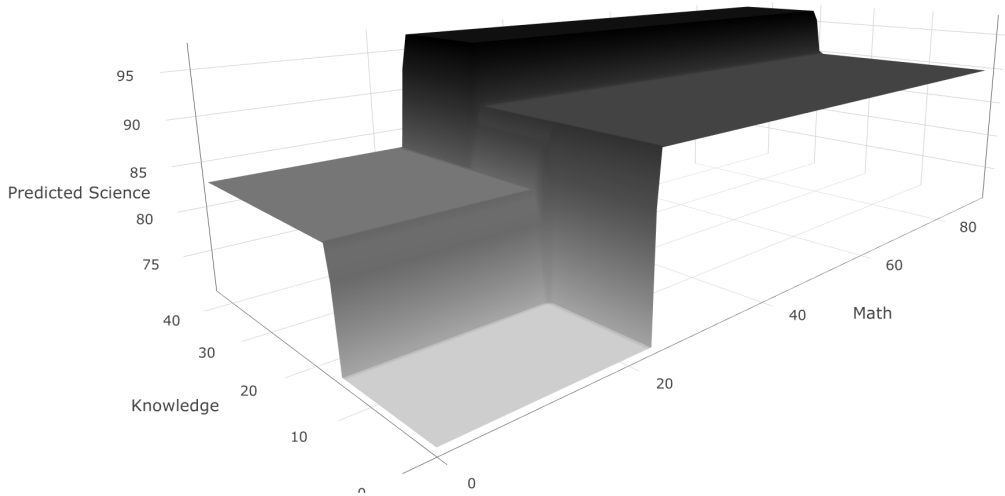
(b) Prediction Function

Example Continued

With 2+ variables, the function becomes more complex.

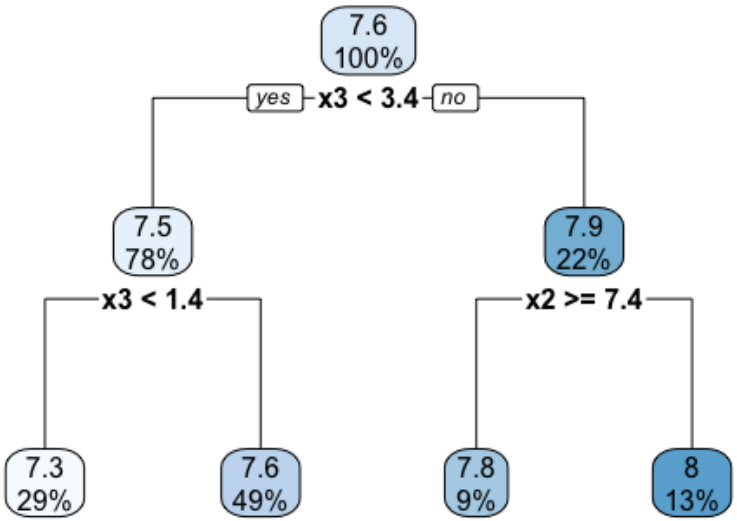


(a) Two Predictor Tree w/ Interaction

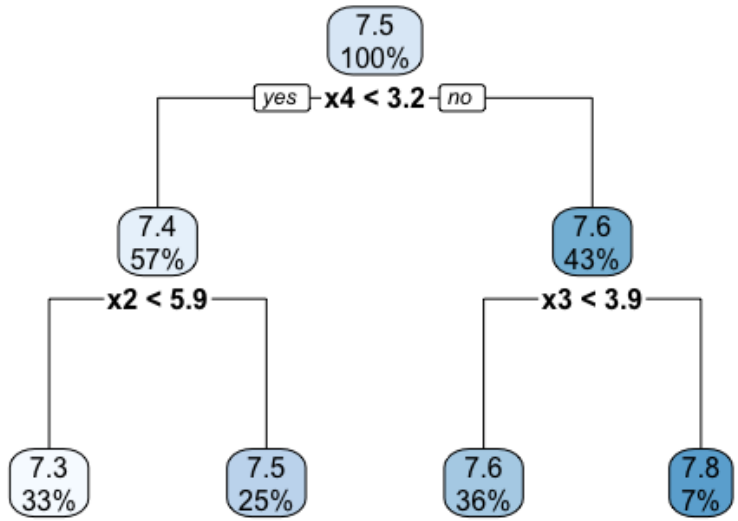


(b) Prediction Function

Problem with Stability



(a) First Half of Sample



(b) Second Half of Sample

Ensembles

Ensembles

Ensemble Definition

- Refers to any algorithm that combines multiple models
- Random Forests
- Boosting
- And many extensions

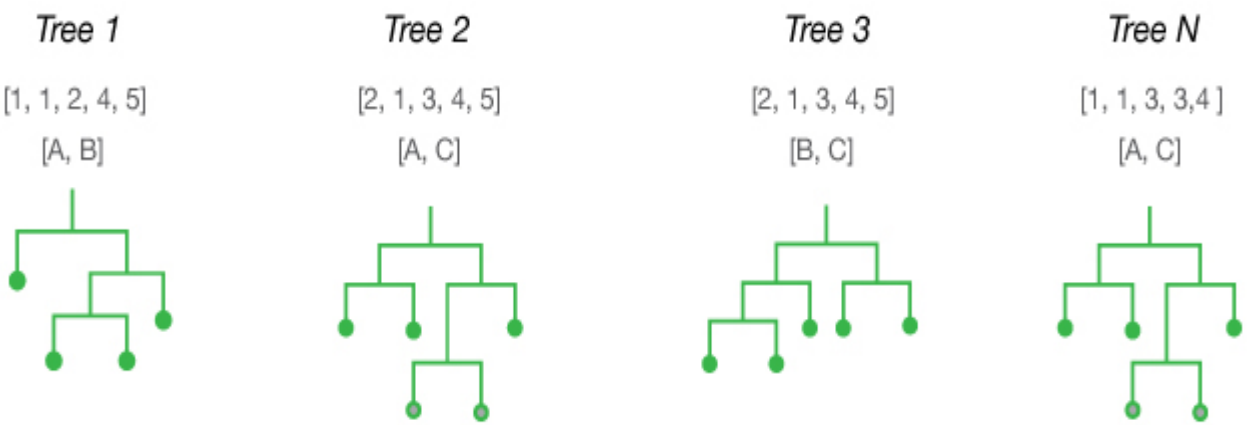
Rationale for Ensemble

- More stable than a single tree (smaller variance)
- Better prediction than a single tree (within and out of sample)

Random Forests

Steps

- Take random sample (subsample or bootstrap)
- Select a subset of predictors
- Create tree
- Repeat hundreds or thousands of times
- Create predictions by aggregating across trees



Boosting

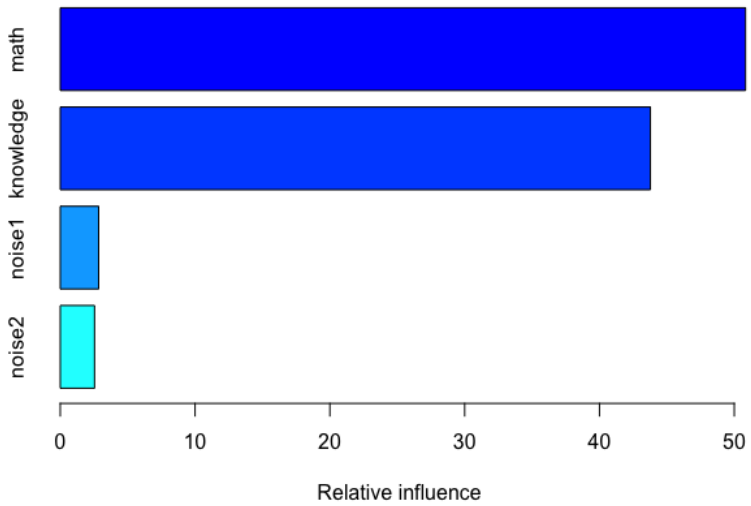
Steps

- Select amount of shrinkage and # of trees
- Fit a small tree to the current residuals
- Update residuals based on new fit * shrinkage (learns slowly)
- Repeat hundreds or thousands of times
- Create predictions by aggregating across trees

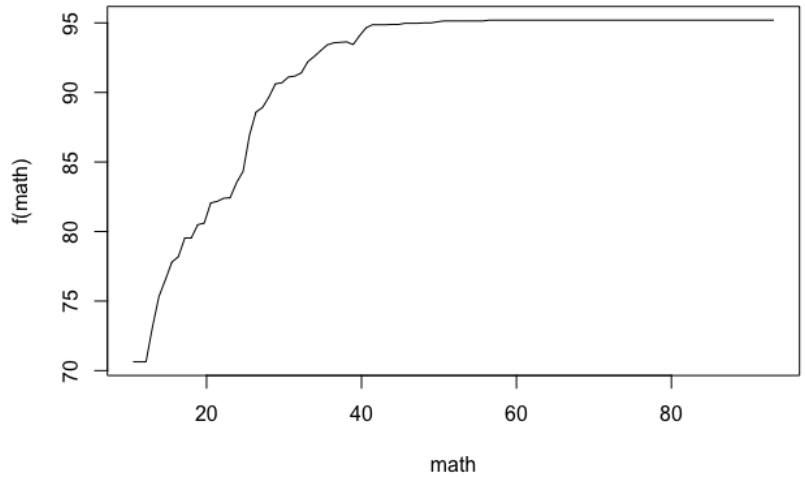
Usually produces similar results to random forests, however, requires slightly more tuning.

Ensemble Drawbacks

No single tree, so have to use variable importance metric and partial dependence plots.



(a) Variable Importance



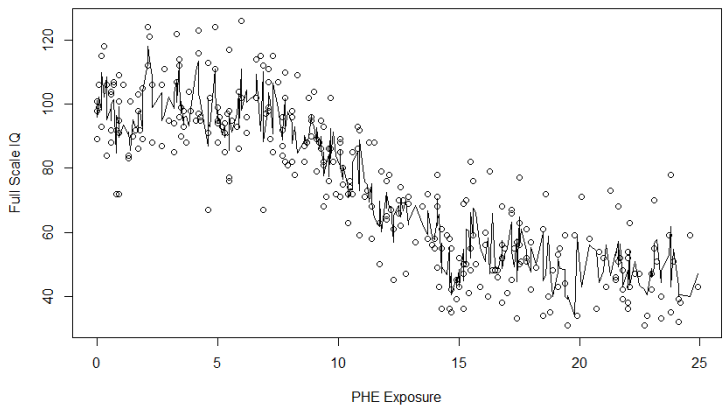
(b) Partial Dependence Plot for Math

Miscellaneous

The Fallacy of the Fancy Algorithm

Don't discount linear models. If you have a lot of variables, use regularized regression.

Example of using too much firepower: Fitting a random forest to 1 nonlinear relationship.



Have to include forms of cross-validation (see Kassraian-Fard et al. [2017]) to choose among models – e.g. Is random forests necessary above and beyond logistic regression.

Interactions

Decision Trees

- Automatically "detects"
- Just need to examine tree structure

Ensembles

- Automatically "detects" and includes across trees.
- See Elith, Leathwick, J.& Hastie (2008; `dismo` package) for understandable overview.

Lasso for Interactions

- If non-zero regression for two variables, puts lasso on interaction.
- See Bien, Taylor, & Tibshirani (2013; `hierNet` or `FAMILY` package in R)
- Multivariate Adaptive Regression Splines does something similar (`earth` package)

Further Info & Material

Slides are posted at rjacobucci.com/presentations and are at github.com/Rjacobucci/abct2017

For a bunch of data mining R code, see github.com/Rjacobucci/SearchWkshp_labs16

Also, Kevin Grimm leads a workshop through the ATI on data mining: http://www.apa.org/science/resources/ati/data-mining-schedule.aspx

References

References

- Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. Annals of Statistics, 41, 1111
- Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77, 802-813.
- Kassraian-Fard, P., Matthis, C., Balsters, J. H., Maathuis, M. H., & Wenderoth, N. (2016). Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. Frontiers in psychiatry, 7.
- McArdle, J. J., & Ritschard, G. (Eds.). (2013). Contemporary issues in exploratory data mining in the behavioral sciences. Routledge.

R packages

- dismo & gbm – boosting
- glmnet – regularized regression
- hierNet & FAMILY – regularized regression with interactions
- rpart & rpart.plot – decision trees
- randomForest – random forests