3 Ways to Improve Your Regression

February 10, 2015

Kaitlin Onthank

Salford Systems

- Salford Systems is an advanced analytics software development company founded in 1983
- Salford is known as the company that brought the world the CART decision tree and TreeNet gradient boosting
- Tools are based on original research of scientists that invented the most important algorithms in machine learning
- Documented advantages in computational speed and accuracy

Outline

- Issues in a Standard Linear Regression
- 3 Solutions
  - MARS
  - TreeNet
  - RandomForests
- Demonstration
- Results and Other Applications

Issues in Regression

- Missing values
  - Results in record deletion OR
  - Requires imputation
- Nonlinearities
  - Ignores local effects
  - Requires manual transformations
- Interactions
  - Requires manual detection
- Variable selection
  - Could be thousands available

Solutions

- Automatic variable selection
- Automatic handling of missing values
- Allows for nonlinearity
- Allows for interactions

Concrete Strength

- Target:
  - STRENGTH
  - Compressive strength of concrete ranging from 2-82 megapascals
- Predictors:
  - CEMENT
  - BLAST_FURNACE_SLAG
  - FLY_ASH
  - WATER
  - SUPERPLASTICIZER
  - COARSE_AGGREGATE
  - FINE_AGGREGATE
  - AGE

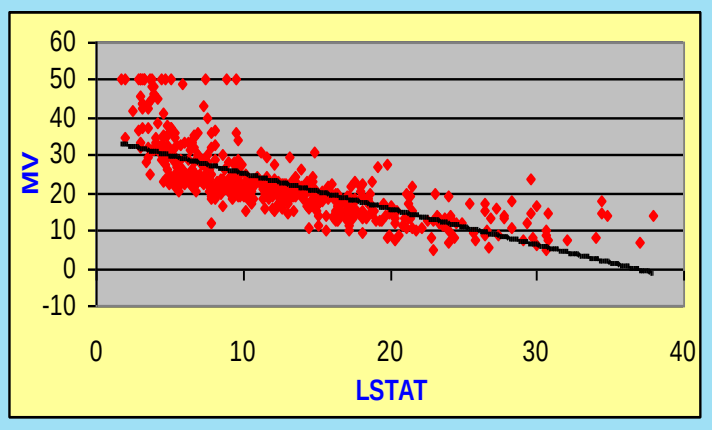
I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)

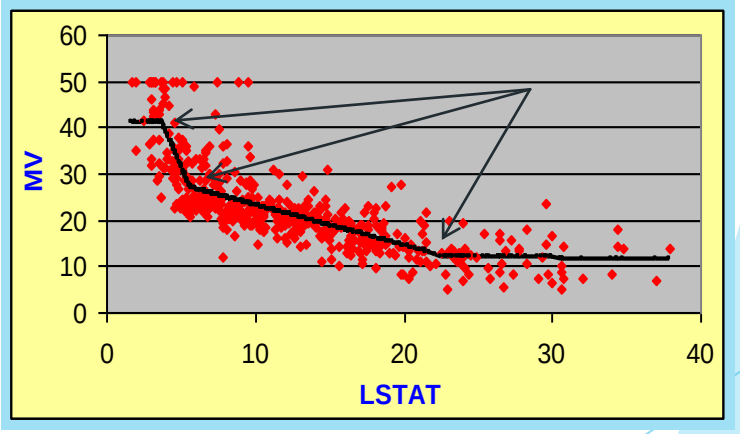
Results - 1

Method	MSE	R^2
Standard Linear Regression	**107.21**	**61.55%**
MARS	-	-
TreeNet	-	-
RandomForests	-	-

MARS

- Multivariate Adaptive Regression Splines
- Uses “knots” to impose local linearities
- These knots create “basis functions” to decompose the information in each variable individually





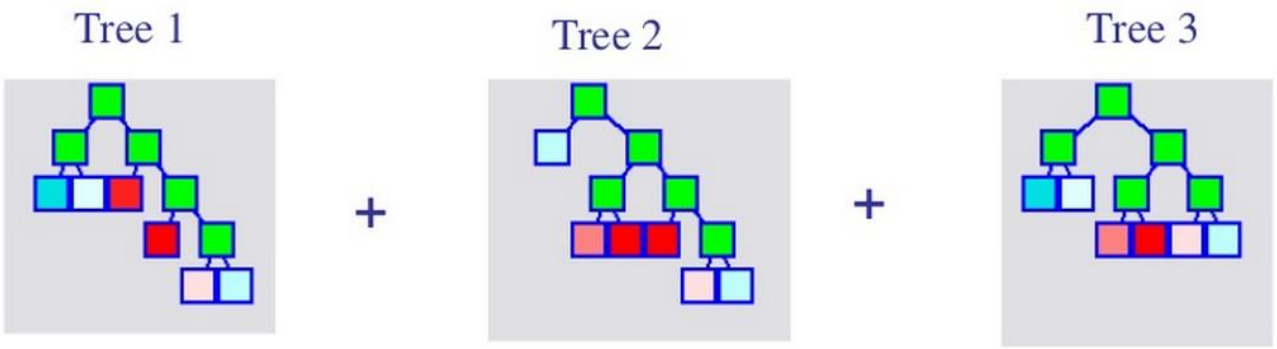
Results - 2

Method	MSE	R^2
Standard Linear Regression	107.21	61.55%
MARS	**34.20**	**87.73%**
TreeNet	-	-
RandomForests	-	-

TreeNet

- Stochastic Gradient Boosting
- Small decision trees built in an error-correcting sequence

1. Begin with small tree as initial model
2. Compute residuals from this model for all records
3. Grow a second small tree to predict these residuals
4. And so on…



Results - 3

Method	MSE	R^2
Standard Linear Regression	107.21	61.55%
MARS	34.20	87.73%
TreeNet	**31.82**	**88.59%**
RandomForests	-	-

Random Forests

- Ensemble of trees built on bootstrap samples
- Algorithm:
  - Each tree is grown on a bootstrap sample from the learning data
  - During tree growing, only *P* predictors are selected and tried at each node
  - By default, *P* is the square root of total predictors
- The overall prediction is determined by averaging
- Law of Large Numbers ensures convergence
- The key to accuracy is low correlation and bias
- To keep bias low, trees are grown to maximum depth

Final Results*

Method	MSE	R^2
Standard Linear Regression	107.21	61.55%
MARS	34.20	87.73%
TreeNet	31.82	88.59%
RandomForests	**25.54**	**90.84%**

*Results can vary depending on parameters, testing methods, random seeds, etc.

Other Applications

- Epidemiology
  - Prostate cancer diagnosis by protein sequencing
- Real Estate
  - Housing values by location, number of rooms, etc.
- Ecology
- Public Health
- Marketing
- Finance

- Upcoming webinars: send us your data!

Questions?

- support@salford-systems.com

- Follow-up email:
  - Recording of webinar
  - SPM 30-day trial download instructions
  - Tutorial with concrete dataset
  - Comprehensive Data Mining Training in NYC: 3/18-3/20