Data Mining Methods: Applications for Institutional Research

Nora Galambos, PhD
Office of Institutional Research, Planning & Effectiveness
Stony Brook University

NEAIR Annual Conference
Philadelphia 2014

Data Mining

- The methods discussed in this presentation were covered in workshops and presentations given at the Joint Statistical Meetings in Boston in August 2014.
  - Most of the methods are available as part of data mining packages, so discussing them will help users understand how to put them into practice.
- Data mining: overview
  - The beginnings of what we now think of data mining had roots in machine learning as far back as the 1960s.
  - In 1989 the Association of Computing Machinery Knowledge Discovery in Databases conferences began informally. Starting in 1995 the international conferences were held formally.
  - Features of data mining
    - Few assumptions to satisfy relative to traditional hypothesis driven methods
    - A variety of different methods for different types of data and predictive needs
    - Able to handle a great volume of data with hundreds of predictors

Data Mining:
How Do Data Scientists Spend Their Time?

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By Steve Lohr: The New York Times, August 17, 2014

Photo: The New York Times, August 15, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.

Stony Brook University

Data Wrangling

- According to the NY Times article, data scientists spend 50 to 80 percent of their time “collecting and preparing unruly data, before it can be explored for useful nuggets.”[1]
- Although CART and CHAID, for example, are able to incorporate missing data without listwise deletion, it still remains important to examine the data and be cognizant of the missing data mechanisms.
- There is a wide variety of formats for data, and it takes time and effort to configure data from numerous sources so it can be combined.
- Companies are starting up to provide data cleaning and configuring services.

[1] Lohr, Steve. *For big-data scientists ‘janitor work’ is key hurdle to insights*.
The New York Times, August 17, 2014.

Data Mining: Initial Steps

- Some of the initial steps are similar to traditional data analysis.
  - Study the problem and select the appropriate analysis method.
  - Study the data and examine for missingness.
    - Though there are data mining methods that are capable of including missing values in the results rather than listwise deleting the observations, one must still examine the data to understand the missing data mechanisms.
  - Study distributions of the continuous variables.
    - Examine for outliers.
  - Recode and combine groups of categorical variables.

Data Mining: Training, Validation, and Test Partitions

- The purpose of the analysis is both explanatory and predictive.
- Need to find the correct level of model complexity.
  - A model that is not complex enough may lack the flexibility to represent the data, under-fitting.
  - When the model is too complex it can be influenced by random noise, over-fitting.
  - For example, if there are outliers, an overly complex model will be fit to them. Then when the model is run on new data, it may be a poor fit.

Data Mining: Training, Validation, and Test Partitions

- Partitioning is used to avoid over- or under-fitting. Divide the data into three parts: training, validation, and testing.
- The ***training*** partition is used to build the model.
- The ***validation*** partition is set aside and is used to test the accuracy and fine tune the model.
  - The prediction error is calculated using the validation data.
  - An increase in the error in the validation set may be caused by over-fitting. The model may need modification.
- The ***test*** partition is used for evaluating how the model will work on new data.

CART: Classification and Regression Trees

- Developed by statisticians at Stanford and Berkley in 1984, but was not used widely until after the turn of the century with the expanded use of data mining.
- Able to handle missing values: does not listwise delete them.
- Easier to use and often more accurate than logistic regression or other parametric methods.
- Data transformations, such as those that are sometimes needed for linear regression to satisfy the assumptions, are unnecessary.

CART: Classification and Regression Trees

- Performs binary splits of the measures in the data.
- CART handles both categorical and continuous measures.
- The MSE is used to determine the best split for regression trees and a measure of the smallest impurity, such as the Gini Index, for categorical data.
- The CART algorithm is robust to outliers, which sometimes are isolated in single nodes.
- When the variable is categorical, classification trees are used, and regression trees are used for continuous variables.
- For categorical variables, indicator, ordinal, and non-ordinal data can be used.

CART: Algorithm

- Creates a set of decision rules to predict an outcome.
- Splits categorical predictors into a smaller number of groups or finds the optimal split in numerical measures.
- Uses recursive partitioning to determine splits with the greatest "purity," i.e., the greatest number of correct values in each split.
- **Recursive Partitioning**
  - Start with a dependent variable, e.g., did the student graduate?
  - All variables will be searched at every value to find the optimal split into two parts.
  - The search continues to find the optimal split in the new region, continuing until all values have been exhausted.

CART: Finding the Tree Size

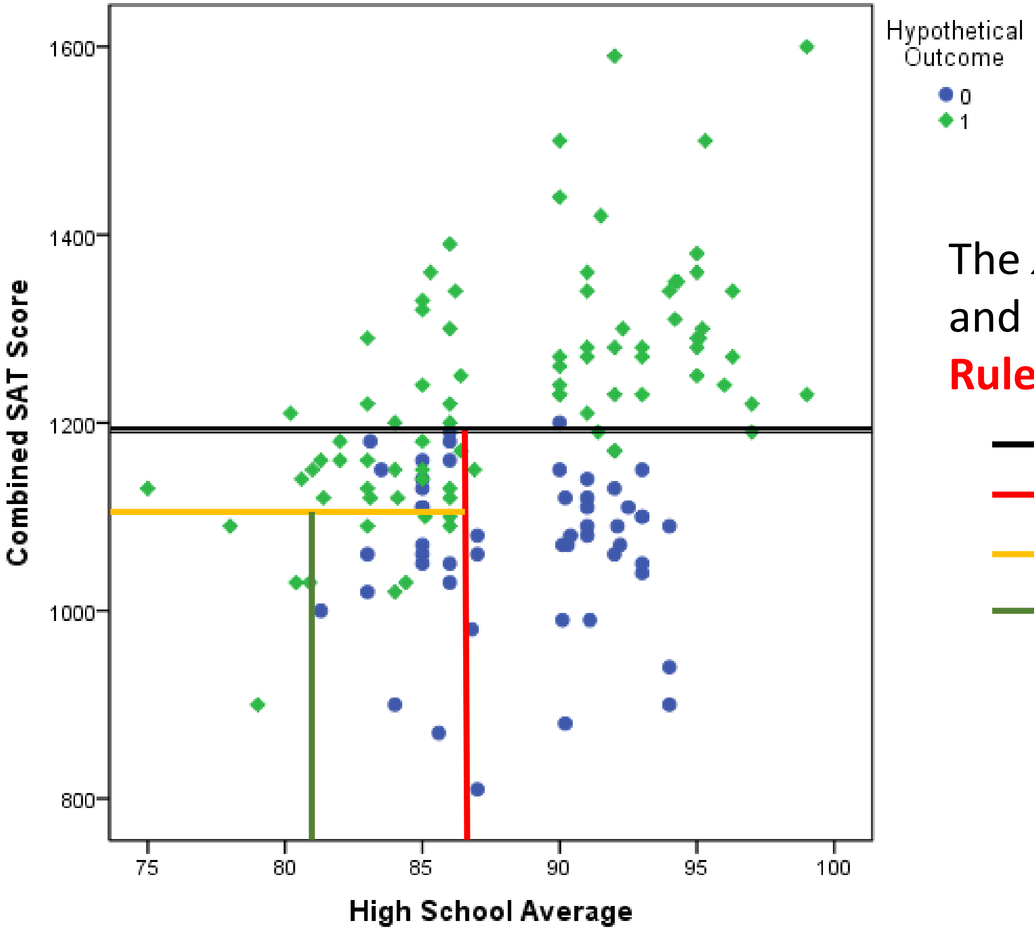
- When the tree grows to use all of the variables, which may be hundreds of levels for large complex datasets, the result may not be useful for making predictions with new data.
- Over-fitting will result in poor predictions when the decision rules are used on new data. The error rate will increase in the validation data.
- CHAID, a different tree-type algorithm, will halt when statistically significant splits are no longer found in the data.
- There are pruning algorithms to find the optimal tree size.
    - Select a minimum number of observations in a node
    - The complexity of the tree is balanced with the impurity. (The overall impurity is measured as the sum of terminal node classification errors.)
    - Limit the total number of nodes.

CART: Missing Value Handling

- Income is a common survey item that is used to illustrate the handling of missing data.
- The tails of the distribution may be biased because high and low income people are more likely to not report their income.
  - *Problem: Need to separate the low income missing from the high income missing*.
- Surrogates are used to fill in the decisions for missing observations.
- CART mathematically finds predictors (and ranks them by strength of association, if any exist) that match the decision split of the primary splitter. In that way missing values can be split into both sides of a decision.
- The output contains the percentage reduction in error for using each surrogate.

CART: Classification and Regression Trees

Hypothetical Example

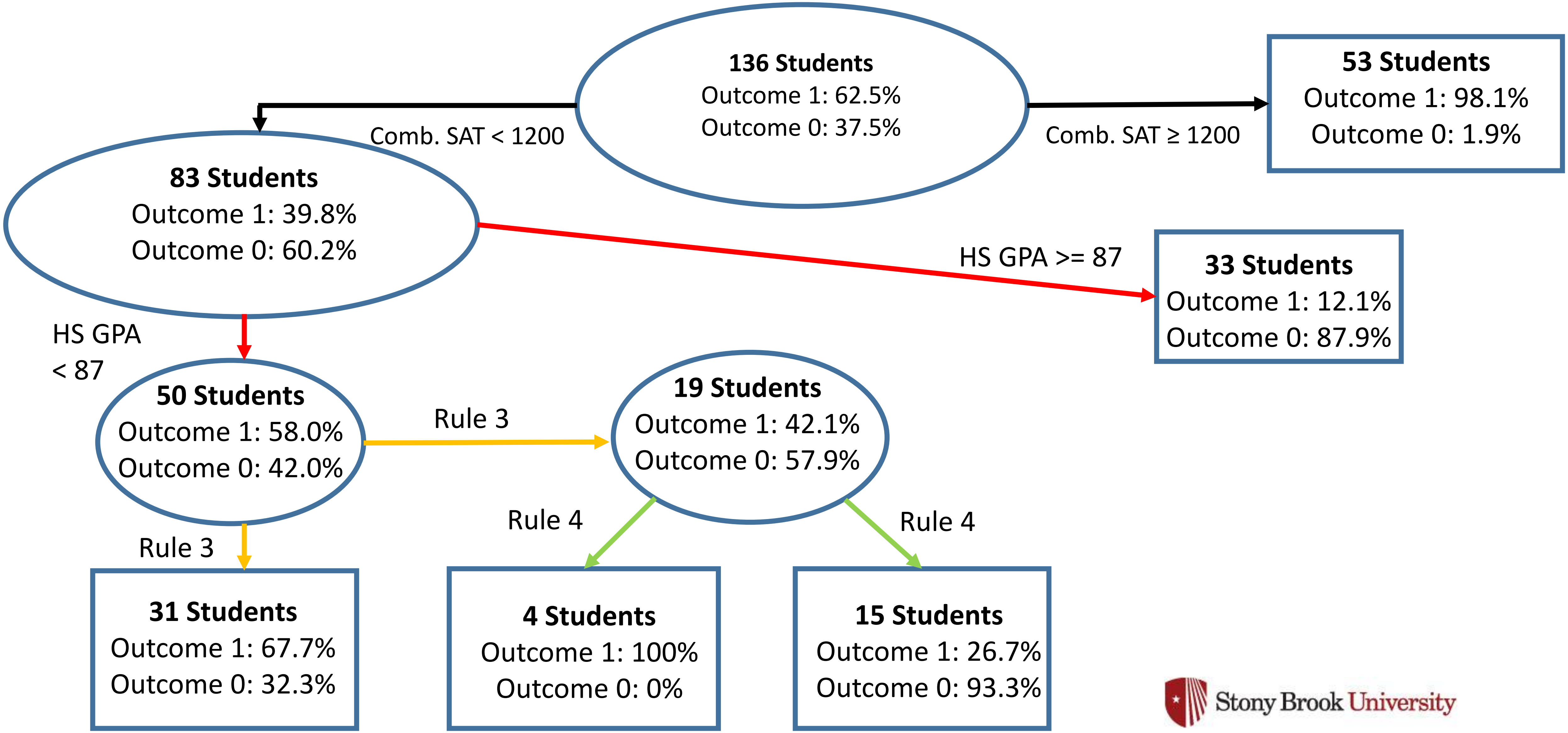


The x_i represent *i* independent predictors and decision rules for the outcome.

Rules for Hypothetical Outcome = 1

- x_{SAT} → Combined SAT <= 1190
- $x_{HS\ GPA}$ → HS GPA < 87.0
- x_3 → Decision rule for factor 3
- x_4 → Decision rule for factor 4

CART: Tree with Hypothetical Data



136 Students
Outcome 1: 62.5%
Outcome 0: 37.5%

Comb. SAT < 1200

Comb. SAT ≥ 1200

53 Students
Outcome 1: 98.1%
Outcome 0: 1.9%

83 Students
Outcome 1: 39.8%
Outcome 0: 60.2%

HS GPA >= 87

33 Students
Outcome 1: 12.1%
Outcome 0: 87.9%

HS GPA
< 87

50 Students
Outcome 1: 58.0%
Outcome 0: 42.0%

Rule 3

19 Students
Outcome 1: 42.1%
Outcome 0: 57.9%

Rule 3

31 Students
Outcome 1: 67.7%
Outcome 0: 32.3%

Rule 4

4 Students
Outcome 1: 100%
Outcome 0: 0%

Rule 4

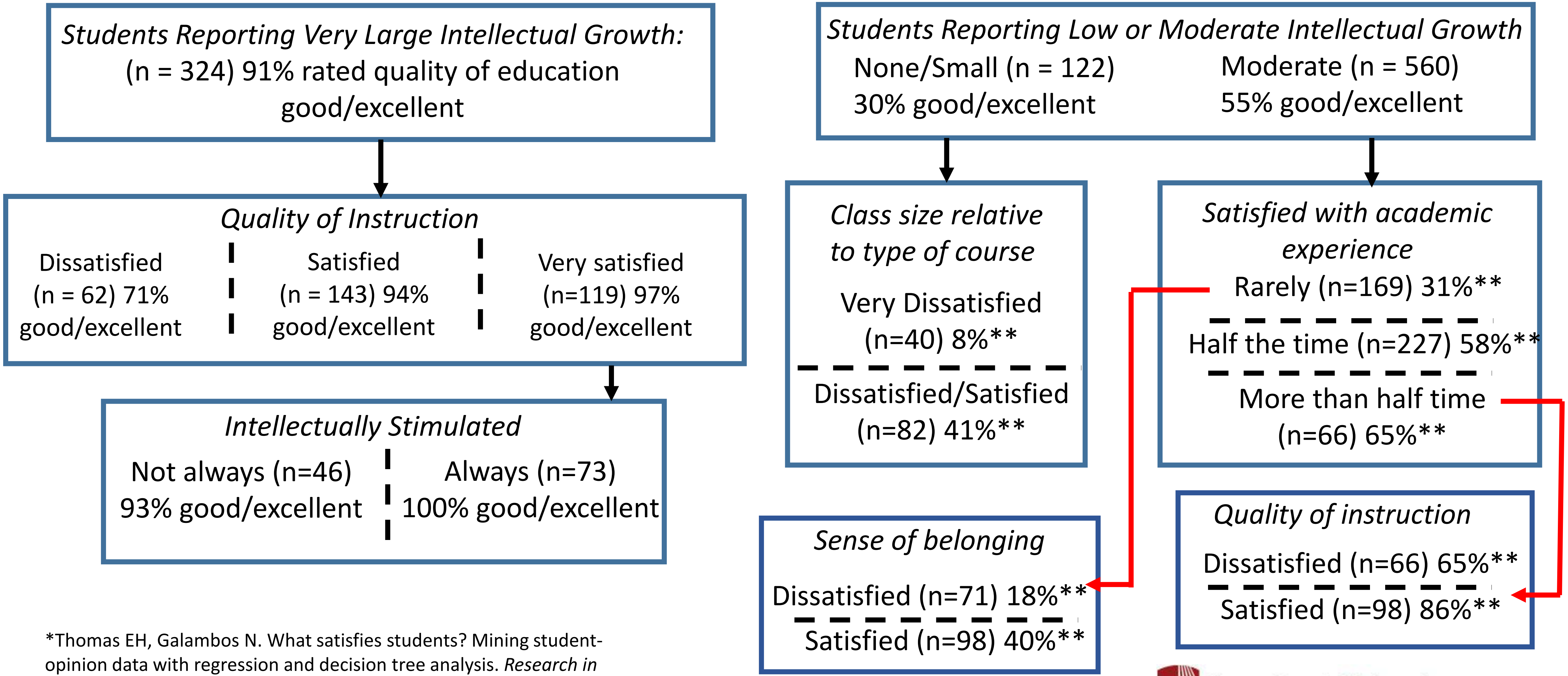
15 Students
Outcome 1: 26.7%
Outcome 0: 93.3%

CHAID Example

- CHAID is another type of tree-based analysis and stands for chi-squared automatic interaction detection.
- Unlike CART with binary splits evaluated by misclassification measures, the CHAID algorithm uses the chi-square test to determine significant splits, as well as the independent variables with the strongest association with the outcome.
- It may find multiple splits in continuous variables, and allows splitting of categorical data into more than two categories.
- As with CART, CHAID allows different predictors for different sides of the binary split.

CHAID Example: Student Opinion Survey*

Percentages Reflect Ratings of Good or Excellent on Quality of Education



Students Reporting Very Large Intellectual Growth: (n = 324) 91% rated quality of education good/excellent

Quality of Instruction

Dissatisfied (n = 62) 71% good/excellent

Satisfied (n = 143) 94% good/excellent

Very satisfied (n=119) 97% good/excellent

Intellectually Stimulated

Not always (n=46) 93% good/excellent

Always (n=73) 100% good/excellent

Students Reporting Low or Moderate Intellectual Growth

None/Small (n = 122) 30% good/excellent

Moderate (n = 560) 55% good/excellent

Class size relative to type of course

Very Dissatisfied (n=40) 8%**

Dissatisfied/Satisfied (n=82) 41%**

Satisfied with academic experience

Rarely (n=169) 31%**

Half the time (n=227) 58%**

More than half time (n=66) 65%**

Sense of belonging

Dissatisfied (n=71) 18%**

Satisfied (n=98) 40%**

Quality of instruction

Dissatisfied (n=66) 65%**

Satisfied (n=98) 86%**

*Thomas EH, Galambos N. What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*. 2004, 45:251-269.

**Rating of good or excellent on quality of education

Stony Brook University

Bagging: Bootstrap Aggregation

- Method of decreasing the variance of the predictive model.
- Bootstrap samples are created by sampling the data with replacement.
  - Assuming the original sample has N observations, each m_i bootstrap sample has *n* observations sampled with replacement.
- The statistic of interest is computed for each sample.
  - For example, we may calculate the mean for each sample. The result will be a distribution of means allowing for a determination of the value of the mean.
- In bagging, multiple CART models are created using bootstrap samples and the results are combined to reduce the variance of the prediction.
  - For regression the results are averaged. For classification, voting algorithms are used whereby the final classification is the one most frequently predicted by the sample results.

CART: Boosting

- Two computer scientists, Yoav Freund and Robert Schapire, from AT&T Labs developed boosting in 1997.
- One common boosting algorithm is AdaBoost or Adaptive Boosting, which adds weights to observations to improve the error rate of predictors that do not perform much better than guessing.
- It will only work for analyses having a binary response variable.
- Boosting is an iterative procedure with the weights updated at each iteration to improve weak predictors.
- In the first step equal weights are assigned to each observation in the training set:

$$w_i = \frac{1}{n} \quad for\ i = 1, \ldots, n$$

CART: Boosting Continued

- Next the misclassification rate for the first iteration is calculated: ϵ_1 which is the proportion of misclassifications in the first iteration.
- Calculate the classifier weight: $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
- The weighted misclassification error:

$D_{t+1}(i) = \frac{D_t(i)}{Z_t}\ e^{-\alpha_t y_i}$, $y_i \in \{-1,1\}$ where -1 is a misclassification, 1 is a correct classification, t is the iteration, and Z_t is a normalization factor that allows the weights to sum to 1.

- The next iteration gives misclassified observations heavier weights for successive iterations.
- At each iteration the weights are summed for all different cutpoints and the cutpoint with the lowest sum is selected.
- The final cutpoints are selected using the combination of all of the results.

CART: Boosting--Comments

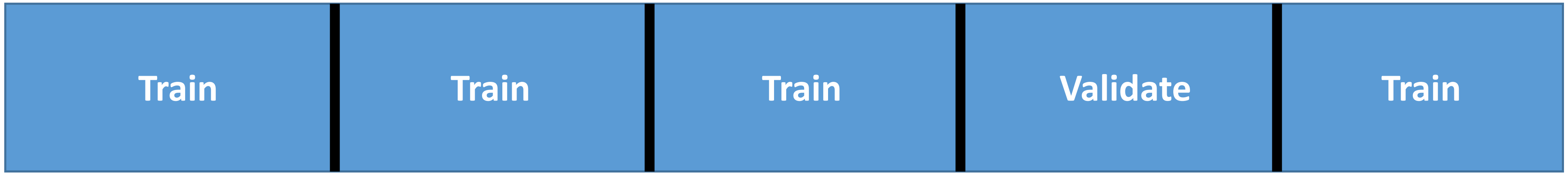
- A disadvantage is that the result is a weighted sum of trees, which can be difficult to interpret.
- Since some higher education data, such as SAT scores, may be difficult to split into a binary decision to predict retention or graduation, boosting may improve the model.
    - There is often not a clear cut SAT score value, below which there is an extremely low misclassification of students predicted to leave a university.
    - High and low SAT score students may leave their institutions for very different reasons.
    - Boosting may be able to lower the misclassification rate in such situations.

M-fold Cross-validation for Evaluating Model Performance

- Why use m-fold cross-validation?
  - Error on the training set does not predict future performance.
  - In order to generalize the prediction on new data it is important to have an accurate estimate of the error.
  - Often the error rate in the training data is overly optimistic. The error rate in the training data may greatly underestimate the test data error.
  - Helps guard against over-fitting
  - Works with limited data.

M-fold Cross-validation for Evaluating Model Performance

- The sample is divided into M equal groups, or folds. Many sources recommend 10 folds if there is enough data.
- Next the model is run M times, however each time, one fold is left out.
- For five folds, four are for training and one is for validation
- The procedure is performed M times (in this case five times), each time leaving out a different validation sample



Train | Train | Train | Validate | Train

M-fold Cross-validation for Evaluating Model Performance

- The initial steps are the similar to traditional data analysis.
- The entire dataset is used to choose the predictors.
- Cross-validation is used to evaluate the model, not to develop the model.
- The error is estimated by averaging the error of the M test samples.

“Listening” to Social Media

- The Census Bureau collected social media posts to use the information to adjust their communications message.
- They tracked the frequency of news mentions of the census over time, as well as the frequency of positive, negative, and neutral posts.[1]
- This method can be used to follow posts on university social media sites.
- Although the results are not from scientific samples, it may be useful to follow what is on the minds of students. Note: There is no link to determine the name of the students who are posting, unless they are voluntarily supplying their names.
- Some type of web scraping software may be useful to collect the data.
- Text mining software can be employed to analyze the results.

[1] Childs JH, Wroblewski M. Replacing public opinion polling with social media listening for purposes of communications research: can we do it? (presentation) Joint Statistical Meetings, Boston MA. August 2-7, 2014.

Use of Transaction Data

- Goal: Assemble various sources of transaction data to add to the more traditional metrics to measure the interaction of students with their college environment.
- Some sources to explore:
  - Interactions with the Blackboard course management system—login info only; no actual course information
  - Academic advising visits
  - Food service card swipes
    - Interest in knowing what students remain on campus over the weekend
  - Library use