Stony Brook University

DATA MINING DEMYSTIFIED: TECHNIQUES AND INFORMATION FOR MODEL IMPROVEMENT

Nora Galambos, PhD
Senior Data Scientist

AIR Forum 2016
New Orleans, LA

FAR
BEYOND

Why Use Data Mining?

- Enables the extraction of information from large amounts of data.
- Incorporates analytic tools for data-driven decision making.
- Uses modeling techniques to apply results to future data.
  - *The goal is to develop a model, not only to find factors significantly associated with the outcomes.*
- Incorporates statistics, pattern recognition, and mathematics.
- 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 Terminology

Decision tree methods/algorithms:

CHAID and CART

Bagging and Boosting

Regression algorithms:

LARS and LASSO

Diagnostics measures:

ROC Curves

Gini Coefficient and Gini Index

Model testing and improvement:

Cross-validation

Receiver Operating Characteristic (ROC) Curves

- Used to discriminate between binary outcomes.
- Originally used during WWII in signal detection to distinguish between enemy and friendly aircraft on the radar screen. In the 1970’s it was realized that signal detection theory could be useful in evaluating medical test results.
- ROC curves can evaluate how well a diagnostic test can correctly separate results into those with and without a disease. It can be used for other types of binary outcomes, such as models predicting retention and graduation.
- The area under the ROC curve is used to evaluate how well a model correctly discriminates between the outcomes.

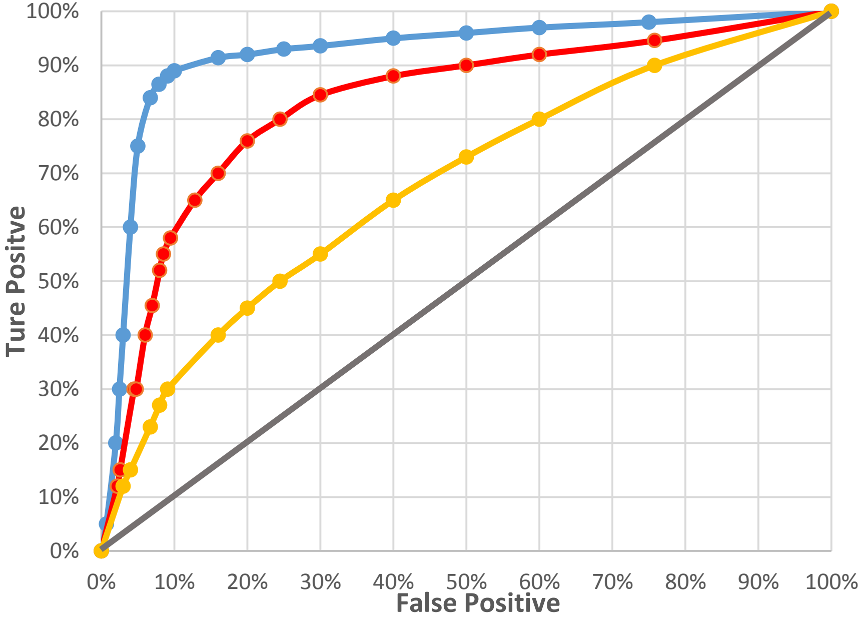
ROC Curves

Sensitivity: True positive rate—probability of a positive result when the disease is present or the outcome occurs. **a/(a+c)**

Specificity: True negative rate—probability of a negative result when the disease is not present or the outcome does not occur. **d/(b+d)**

False positive rate: 1 - specificity

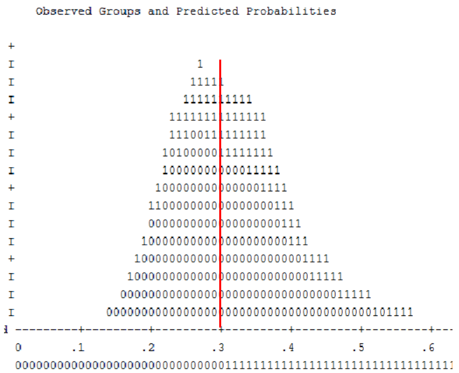
	Observed	
Predicted	**+**	**-**
+	a	b
-	c	d



The greater the area under the ROC curve, the better the discrimination between positive and negative outcomes. It is imperative to examine both the true positive and false positive rates, not just the percentage of correct predictions.

ROC Curve: Cutpoints

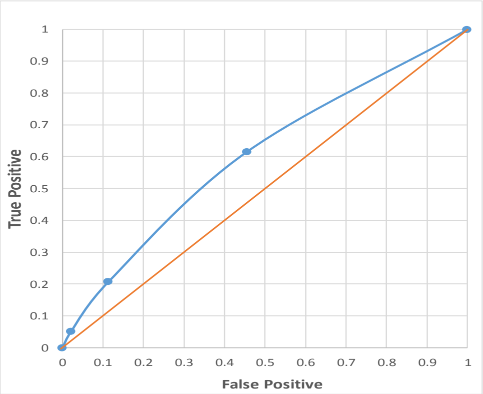
- The sensitivity and specificity vary with the cutpoints of the diagnostic test or the predicted probability of a logistic model. A ROC curve can be created by plotting the true and false positive rates at different model cutpoints.
- As the sensitivity increases the specificity decreases.
- The area under the ROC curve indicates how well the test discriminates between positive and negative outcome groups.
  - AUC = 0.5 indicates no ability to discriminate between groups.
  - Compare ROC curves to determine the model that does the best job at discriminating between groups.



```
    Observed Groups and Predicted Probabilities

+
I                          1
I                         11111
I                        1111111111
+                      11111111111111
I                      11100111111111
I                     1010000011111111
I                     10000000000011111
+                    1000000000000001111
I                   110000000000000000111
I                   0000000000000000000111
I                  100000000000000000000111
+                 1000000000000000000000001111
I                1000000000000000000000000011111
I               000000000000000000000000000000011111
I             00000000000000000000000000000000000000101111
I---------+---------+---------+---------+---------+---------+---
 0        .1        .2        .3        .4        .5        .6
 00000000000000000000000000000011111111111111111111111111111111
```

Classification Plot and Corresponding ROC Curve

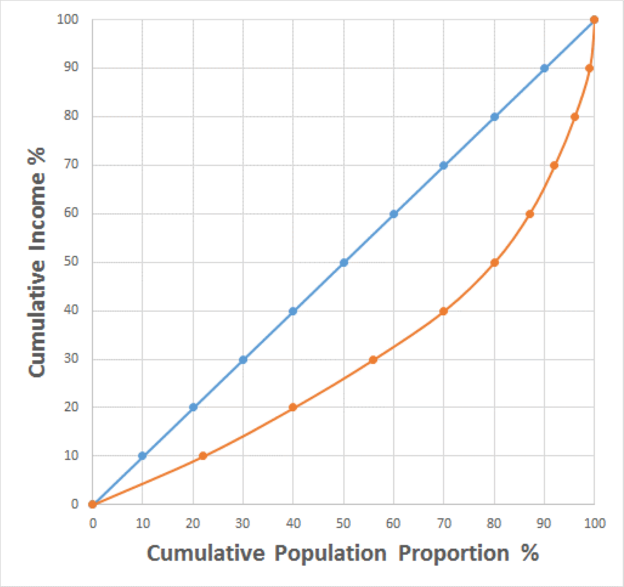


Gini Coefficient

The Lorenz Curve was developed by the economist Max Lorenz in 1905 to represent the distribution of wealth and the Italian statistician Corrado Gini developed the Gini Coefficient in 1912 as a measure of income inequality.

The x-axis is the cumulative population proportion. The y-axis is the cumulative income. Each point on the Lorenz curve gives the proportion of income (y value) possessed by the corresponding proportion of the population. For example on the curve to the right, 70% of the population has only 40% of the wealth.

The closer the Lorenz curve is to the diagonal, the more equal the income distribution.



Gini Coefficient

$$Gini = \frac{A}{A + B}$$

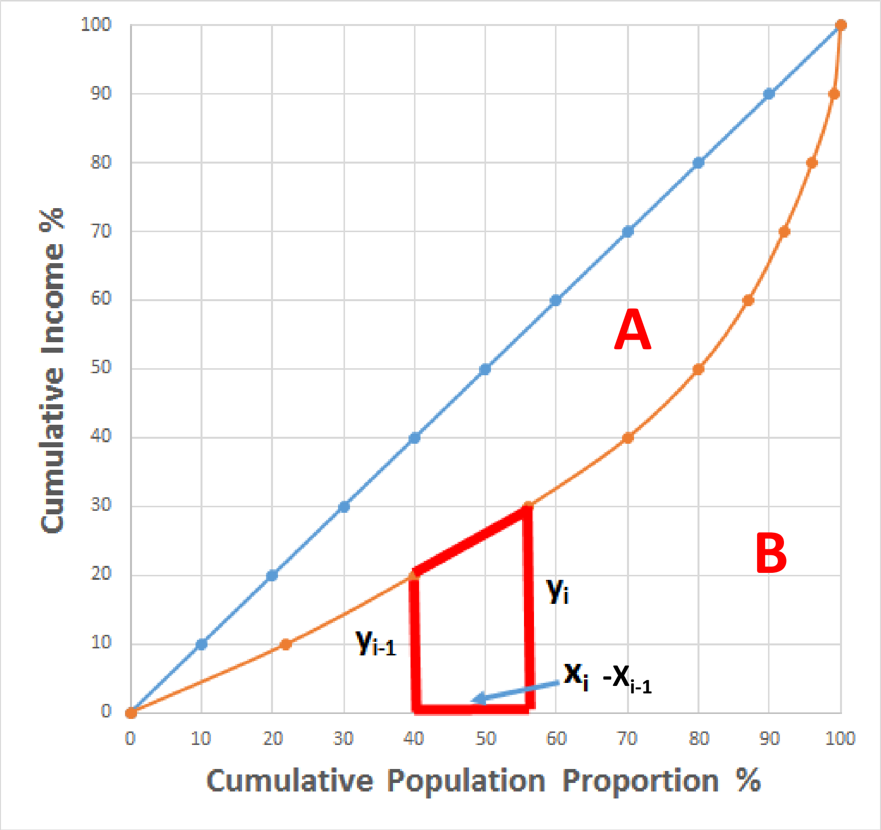
$$A + B = \frac{1}{2} base\ X\ height = \frac{1}{2}$$

$$A = \frac{1}{2} - \sum_{k=0}^{n} \frac{1}{2}(x_i - x_{i-1})(y_i + y_{i-1})$$

$$G = \frac{\frac{1}{2} - \frac{1}{2}\sum_{k=0}^{n}(x_i - x_{i-1})(y_i + y_{i-1})}{\frac{1}{2}}$$

$$=1 - \sum_{k=0}^{n}(x_i - x_{i-1})(y_i + y_{i-1})$$

$$= 1 - 2B$$



The Gini Coefficient is 1 minus twice the area under the Lorenz Curve.

In predictive modeling the Gini Coefficient is used to evaluate the strength of a model, in which case a greater Gini coefficient is better.

Gini Index:
An Impurity Measure

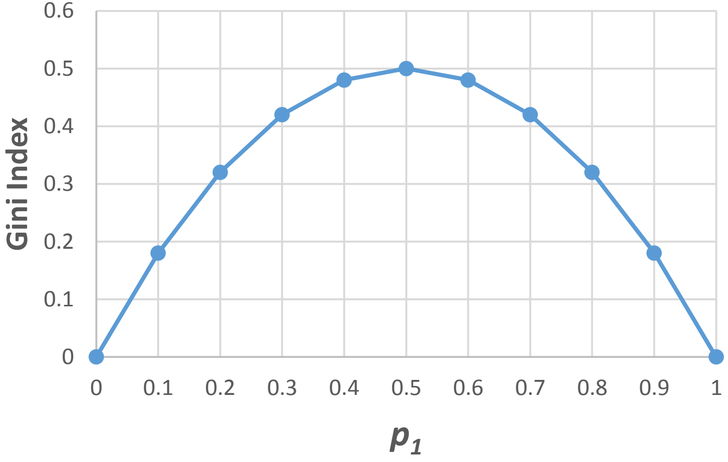
The Gini Index is used for selecting the best way to split data between nodes in decision trees by measuring the degree of impurity of the nodes. When there are only two categories the maximum value for the Gini Index = 0.5, indicating the greatest degree of impurity.

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

c = the number of classes, t represents the node and p is the probability of group i membership.

A word of caution: Sometimes the Gini Coefficient is referred to as the Gini Index.

Gini Index Graph: Two Categories



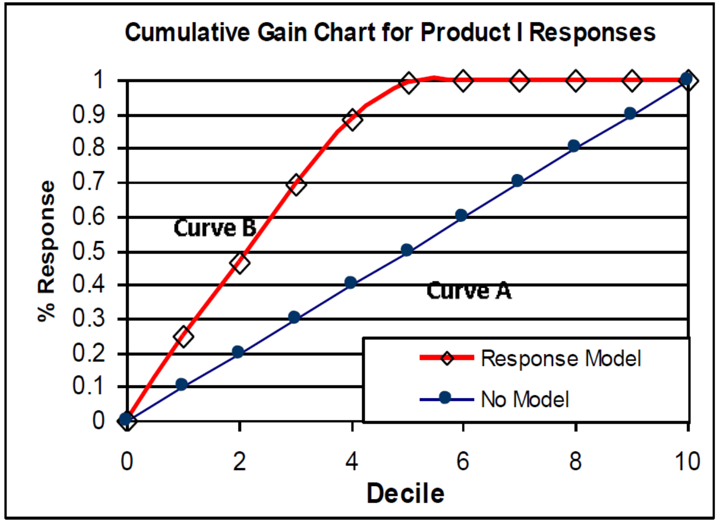
For two categories—x axis: group 1 probability, y axis: Gini Index

Gain Chart

Gain and lift are often used to evaluate models used for direct marketing, but they can be used to compare decision tree models for other uses, as well.

The diagonal line represents random response with no model. To evaluate the effectiveness of a model, the data are sorted in descending order of the probability of a positive response, often binned by deciles. The cumulative rate of a positive response rate at each decile is evaluated.

In the example 70% of the positive responses came from the first three deciles or 30% of the subjects.



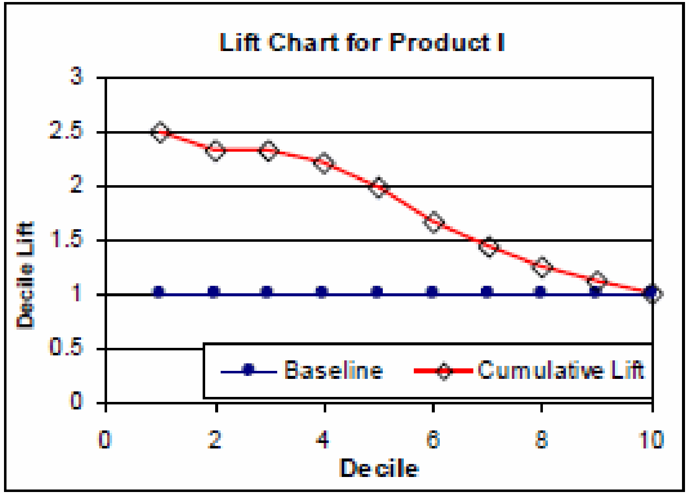
Jaffery T., Liu SX. *Measuring Campaign Performance using Cumulative Lift and Gain Chart.* SAS Global Forum 2009.

FAR BEYOND

Lift Charts

The cumulative lift chart evaluates how many times better the model performs compared to having no model at all. Considering the example on the previous slide at the third decile, 30% of the subjects accounted for 70% of the total responses. 70/30 = 2.33, so the model performs 2.33 times better than having no model.

Lift and gain charts are used to compare models and to determine whether using a particular model is worthwhile. If this example was for a direct mailing, we would expect that 70% of the total responses could be obtained by sending the mailing to just 30% of the sample.



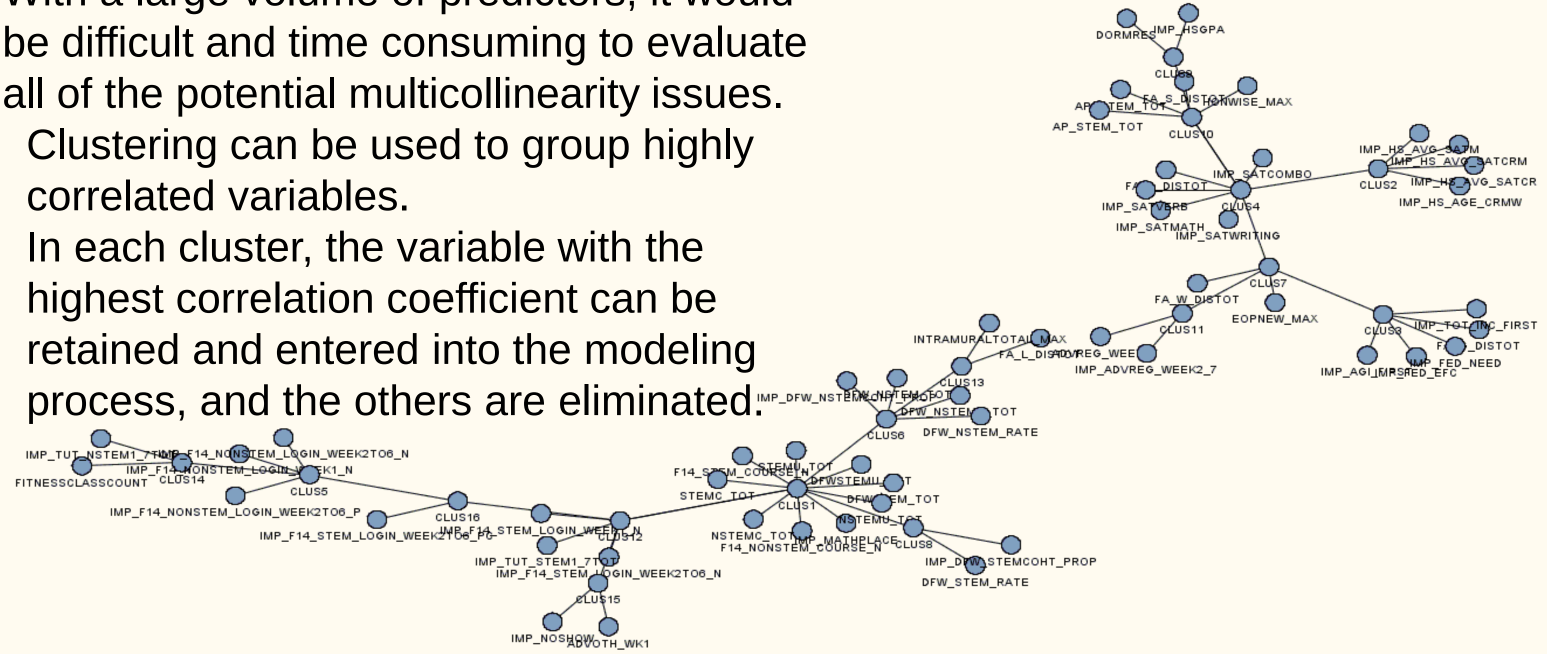
Jaffery T., Liu SX. *Measuring Campaign Performance using Cumulative Lift and Gain Chart.* SAS Global Forum 2009.

Linear Regression in a Data Mining Environment

- Linear regression is an available data mining modeling tool, however it is important to be mindful of missing data and multicollinearity.
- Decision tree methods are able to handle missing values by combining them with another category or placing them in a category with other values. They can also be replaced by using surrogate rules. Linear regression, on the other hand, will listwise delete the missing values.
- When using data having dozens or even hundreds of potential predictors it could happen that not much data remains.
- It is important to note the number of observations remaining in your model and consider using an imputation method provided by the software package if listwise deletion is a serious problem.

Clustering

- With a large volume of predictors, it would be difficult and time consuming to evaluate all of the potential multicollinearity issues.
  - Clustering can be used to group highly correlated variables.
  - In each cluster, the variable with the highest correlation coefficient can be retained and entered into the modeling process, and the others are eliminated.



LASSO: Least Absolute Shrinkage and Selection Operator

The LASSO algorithm shrinks the independent variable coefficients and sets some of them to zero to simplify the model. Ordinary Least Squares regression equation:

$$y = x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

The sum of the squared residuals is minimized in OLS regression. However, LASSO requires that this minimization is subject to the sum of the absolute values of the coefficients being less than or equal to some value t, which is referred to as a tuning parameter.

$$\sum_j |\beta_j| \leq t$$

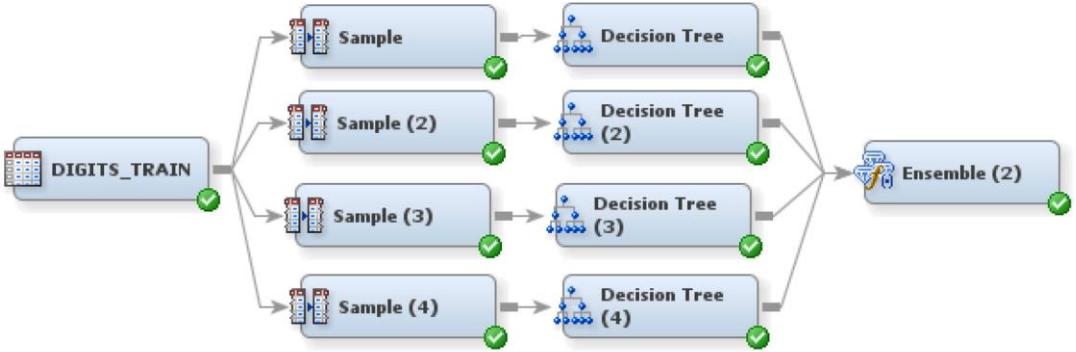
The constraint forces some coefficients to be set to zero and results in a simpler model. As the tuning parameter increases to a very large value, it approaches the OLS regression model. It is important to test a range of values for t.

LARS: Least Angle Regression

The LARS procedure is an efficient method computationally and begins the like ordinary least squares regression by adding the predictor x_i with the highest correlation to ***y***. Next the coefficient β_i is increased (or decreased if β_i is negative) until some other predictor x_k is just as correlated with the residuals as is x_i. The coefficient of that predictor is increased until that predictor is no longer the one most correlated with the residual ***r***. "*At this point LARS parts company with Forward Selection. Instead of continuing along x_{j1}, LARS proceeds in a direction equiangular between the two predictors until a third variable x_{j3} earns its way into the 'most correlated' set. LARS then proceeds equiangularily between x_{j1},x_{j2} and x_{j3}, i.e. along the 'least angle direction', until a fourth variable enters, etc.*"[1]

[1] Efron B, et al., Least Angle Regression. 2002. retrieved from http://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf

Bagging: Bootstrap Aggregating



- Bootstrapping is used to form datasets
  - We have ***M*** datasets by sampling with replacement
- A separate tree is created using each of the ***M*** datasets.
- The improved final model is a combination or average of the ***M*** decision trees.

AdaBoost: Adaptive Boosting

- The decision tree is fit to the entire training set.
  - Misclassified observations are weighted to encourage correct classification.
  - This is done repeatedly with new weights assigned at each iteration to improve the accuracy of the predictions.
- The result is a series of decision trees, each one adjusted with new weights based on the accuracy of the estimates or classifications of the previous tree.
- Although boosting generally results in an improved model, because the results at each stage are weighted and combined into a final model, there is no resulting tree diagram.
  - Scoring code is generated by the software package allowing the model to be used to score new data for predicting outcomes.

BFOS CART Method
Breiman, Friedman, Olshen, Stone
Classification and Regression Trees

- The method does an exhaustive search for the best binary split.
- It splits categorical predictors into two groups, or finds the optimal binary split in numerical measures.
  - Each successive split is again split in two until no further splits are possible.
  - The result is a tree of maximum possible size, which is then pruned back.
  - For interval targets the variance is use to assess the splits; For nominal targets the Gini impurity measure is used.
  - Pruning starts with the split that has the smallest contribution to the model
  - The missing data is assigned to the largest node of a split
- Creates a set of nested binary decision rules to predict an outcome.

CHAID: Chi-squared Automatic Interaction Detection

Unlike CART with binary splits evaluated by misclassification measures, the CHAID algorithm uses the chi-square test (or the F test for interval targets) to determine significant splits and find independent variables with the strongest association with the outcome. A Bonferroni correction to the p-value is applied prior to the split.

It may find multiple splits in continuous variables, and allows splitting of data into more than two categories.

As with CART, CHAID allows different predictors for different sides of a split.

The CHAID algorithm will halt when statistically significant splits are no longer found in the data.

Data Partitioning

- Partitioning is used to avoid over- or under-fitting. The data are divided 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.
  - Often 60% is used for training the model and 40% is used for validation--or 40% for training, 30% for validation, and 30% for testing
  - ***Problem:*** What if the sample size is small? E.g., predicting freshmen retention where the retention rate is usually around 90% and there are 3,000 freshmen. That means that the training sample may only have around 180 students who are not retained to develop a model that may have 50 or more predictors. Using K-fold cross validation is the answer.

K-fold Cross-validation
for Evaluating Model Performance

Why use k-fold cross-validation?

It works with limited data.

- 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 K test samples.

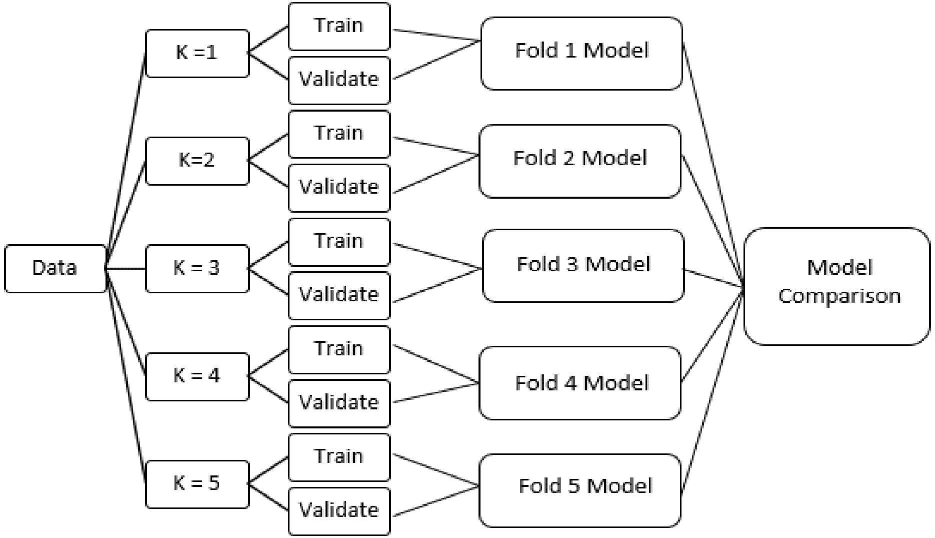
5-Fold Cross Validation Plan

For each K_i the entire dataset was divided into 5 equal parts

K = 1	Train	Train	Train	Train	Validate
K = 2	Train	Train	Train	Validate	Train
K = 3	Train	Train	Validate	Train	Train
K = 4	Train	Validate	Train	Train	Train
K = 5	Validate	Train	Train	Train	Train

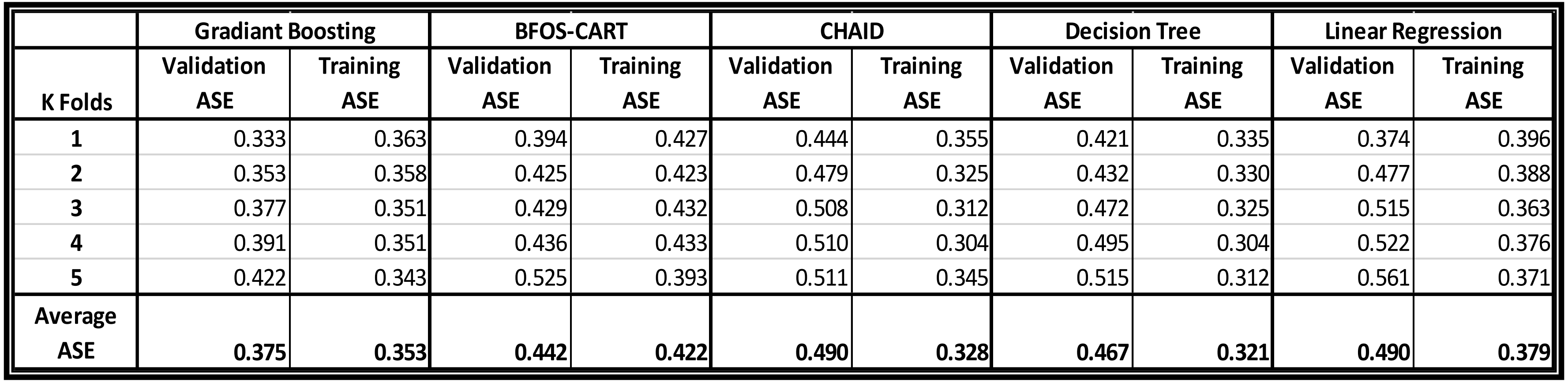
5-Fold Cross Validation Plan

For each K_i the entire dataset was divided into 5 equal parts



Stony Brook University

Cross Validation Results: Average Squared Error (ASE) Results for Five Data Mining Methods to Predict Freshmen GPA

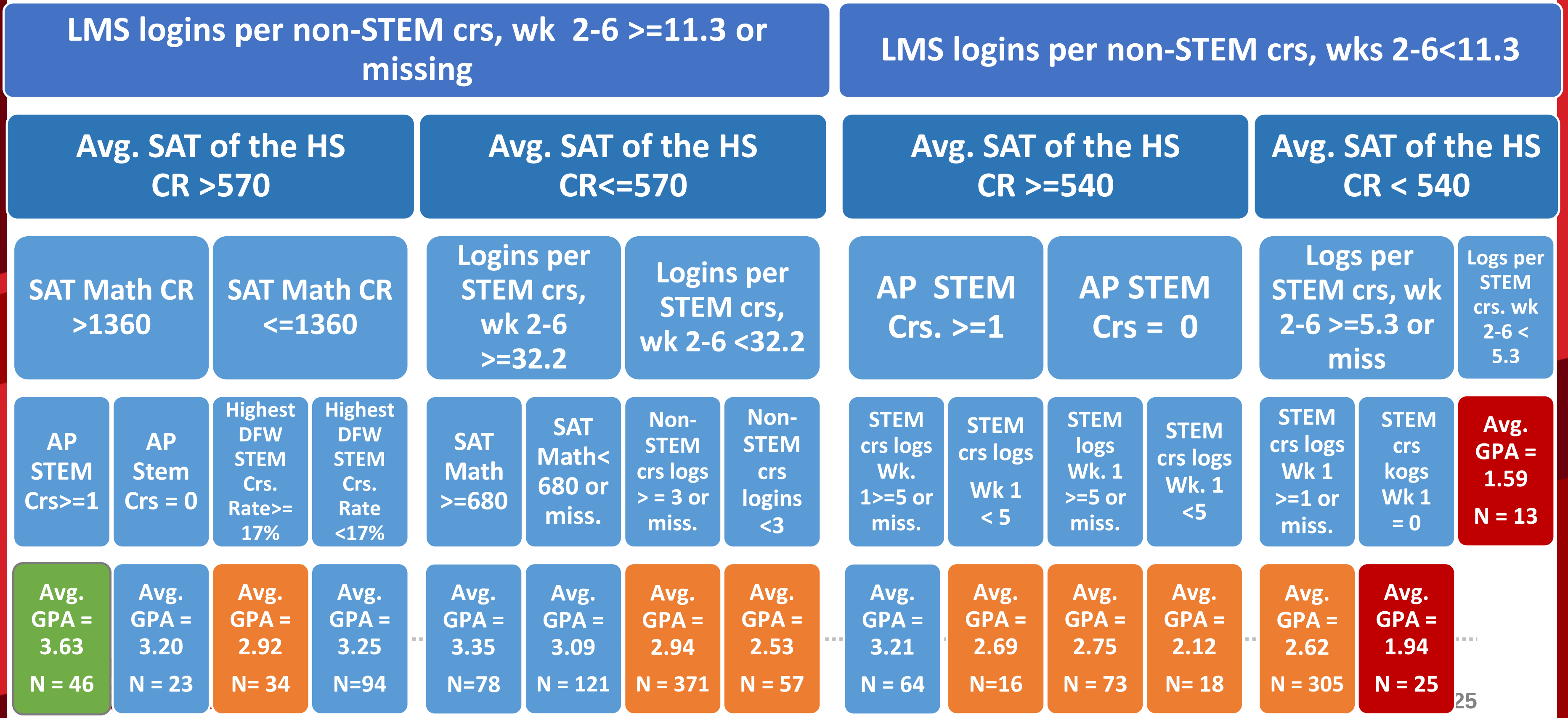


	Gradiant Boosting		BFOS-CART		CHAID		Decision Tree		Linear Regression	
K Folds	Validation ASE	Training ASE	Validation ASE	Training ASE	Validation ASE	Training ASE	Validation ASE	Training ASE	Validation ASE	Training ASE
1	0.333	0.363	0.394	0.427	0.444	0.355	0.421	0.335	0.374	0.396
2	0.353	0.358	0.425	0.423	0.479	0.325	0.432	0.330	0.477	0.388
3	0.377	0.351	0.429	0.432	0.508	0.312	0.472	0.325	0.515	0.363
4	0.391	0.351	0.436	0.433	0.510	0.304	0.495	0.304	0.522	0.376
5	0.422	0.343	0.525	0.393	0.511	0.345	0.515	0.312	0.561	0.371
Average ASE	**0.375**	**0.353**	**0.442**	**0.422**	**0.490**	**0.328**	**0.467**	**0.321**	**0.490**	**0.379**

ASE = (Sum of Squared Errors)/N

Gradient boosting had the smallest average ASE followed by CART. Gradient boosting and BFOS-CART, on average, had the smallest differences between the validation and training errors. The CART method was chosen for the modeling process—Had relatively low ASE. Gradient boosting, without an actual tree diagram, would make the results more difficult to explain.

Portion of CART Tree for HS GPA<=92.0



LMS logins per non-STEM crs, wk 2-6 >=11.3 or missing

LMS logins per non-STEM crs, wks 2-6<11.3

Avg. SAT of the HS CR >570

Avg. SAT of the HS CR<=570

Avg. SAT of the HS CR >=540

Avg. SAT of the HS CR < 540

SAT Math CR >1360

SAT Math CR <=1360

Logins per STEM crs, wk 2-6 >=32.2

Logins per STEM crs, wk 2-6 <32.2

AP STEM Crs. >=1

AP STEM Crs = 0

Logs per STEM crs, wk 2-6 >=5.3 or miss

Logs per STEM crs. wk 2-6 < 5.3

AP STEM Crs>=1

AP Stem Crs = 0

Highest DFW STEM Crs. Rate>= 17%

Highest DFW STEM Crs. Rate <17%

SAT Math >=680

SAT Math< 680 or miss.

Non-STEM crs logs > = 3 or miss.

Non-STEM crs logins <3

STEM crs logs Wk. 1>=5 or miss.

STEM crs logs Wk 1 < 5

STEM logs Wk. 1 >=5 or miss.

STEM crs logs Wk. 1 <5

STEM crs logs Wk 1 >=1 or miss.

STEM crs kogs Wk 1 = 0

Avg. GPA = 1.59 N = 13

Avg. GPA = 3.63 N = 46

Avg. GPA = 3.20 N = 23

Avg. GPA = 2.92 N= 34

Avg. GPA = 3.25 N=94

Avg. GPA = 3.35 N=78

Avg. GPA = 3.09 N = 121

Avg. GPA = 2.94 N = 371

Avg. GPA = 2.53 N = 57

Avg. GPA = 3.21 N = 64

Avg. GPA = 2.69 N=16

Avg. GPA = 2.75 N = 73

Avg. GPA = 2.12 N= 18

Avg. GPA = 2.62 N = 305

Avg. GPA = 1.94 N = 25