Stony Brook University

An Overview of Data Mining: Predictive Modeling for IR in the 21st Century

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

AIRPO Annual Conference
Lake George 2015

Data Mining

- 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 Wrangling

- According to a 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. The New York Times, August 17, 2014

Data Mining: Initial Steps

- Some of the initial steps are the 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

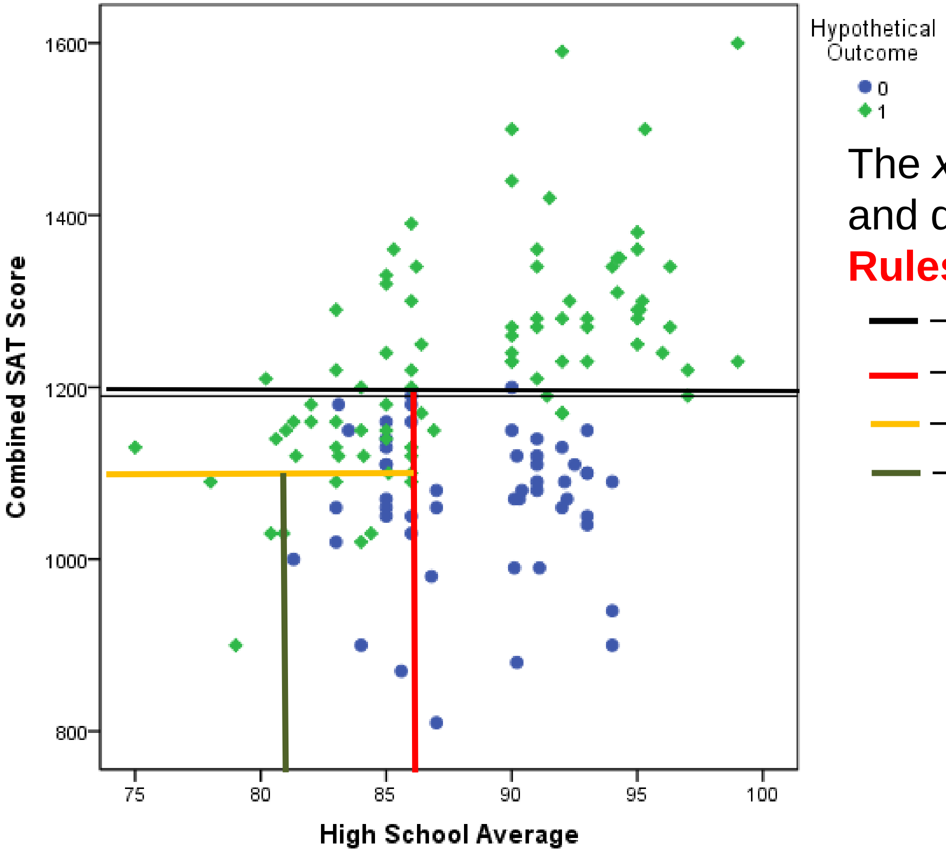
- When the tree grows to use all of the variables, which may be hundreds of levels for large complex datasets, the result may not be useful for making predictions with new data.
- Over-fitting will result in poor predictions when the decision rules are used on new data. The error rate will increase in the validation data.
- The CHAID 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

CHAID

- 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.

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 the predictions to improve weak predictors.

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.

What is a Neural Network?

[1] "A neural network … has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process.
2. Interneuron connection strengths known as synaptic weights are used to store the knowledge. "

Neural networks are especially useful for prediction problems where:

- No mathematical formula is known that relates inputs to outputs.
- Prediction is more important than explanation.
- There is a lot of training data.

[1] Haykin, S. (1994), *Neural Networks: A Comprehensive Foundation,* NY: Macmillan
ftp://ftp.sas.com/pub/neural/FAQ.html#A2
SAS Enterprise Miner Manual

What is a Neural Network?

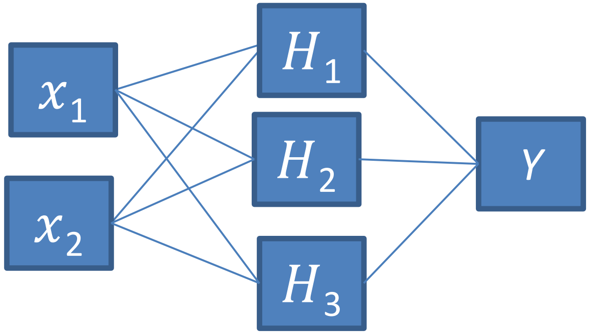
- Developed by researchers to mimic the neurophysiology of the human brain.
- By combining many simple computing elements (neurons or units) into a highly interconnected system, these researchers hoped to produce complex phenomena such as intelligence.
- In recent years, neural network researchers have incorporated methods from statistics and numerical analysis into their networks.
- The feedforward neural networks are a class of flexible nonlinear regression, discriminant, and data reduction models, which detect complex nonlinear relationships in data.

SAS Enterprise Miner Manual

Neural Network Prediction Formula

$$\hat{y} = \widehat{w}_{00} + \widehat{w}_{01} \cdot H_1 + \widehat{w}_{02} \cdot H_2 + \widehat{w}_{03} \cdot H_3$$

$$H_1 = \tanh(\widehat{w}_{10} + \widehat{w}_{11}\, x_1 + \widehat{w}_{12}\, x_2)$$
$$H_2 = \tanh(\widehat{w}_{20} + \widehat{w}_{21}\, x_1 + \widehat{w}_{22}\, x_2)$$
$$H_3 = \tanh(\widehat{w}_{30} + \widehat{w}_{31}\, x_1 + \widehat{w}_{32}\, x_2)$$



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 during the weekend
  - Library use

Using SAS Enterprise Miner for Predictive Modeling

- The model presented is a preliminary version for the prediction of the first semester GPA of first-time full-time fall 2014 freshmen.
- Measures included in the model
  - The results incorporate transactional data to provide different insights into student outcomes. Those data include:
    - Blackboard logins
    - Advising visits
    - Tutoring center usage
  - The model also includes traditional demographics and pre-college characteristics, e.g., SAT scores, gender, ethnicity, and transfer and AP credits upon admissions
  - Financial aid measures: AGI, EFC, and disbursed amounts of different aid types
  - DFW rates
  - Average SAT scores of the students' high schools.
- Preliminary results demonstrate that high school GPA is the strongest predictor and that BlackBoard logins as a proxy for student academic engagement appears to play a more important role than SAT scores. Unfortunately, BlackBoard logins are regularly purged, so using the results of previous cohorts to improve the model is not possible.

Variable Importance List: SAS Enterprise Miner Output from Preliminary Decision Tree Analysis

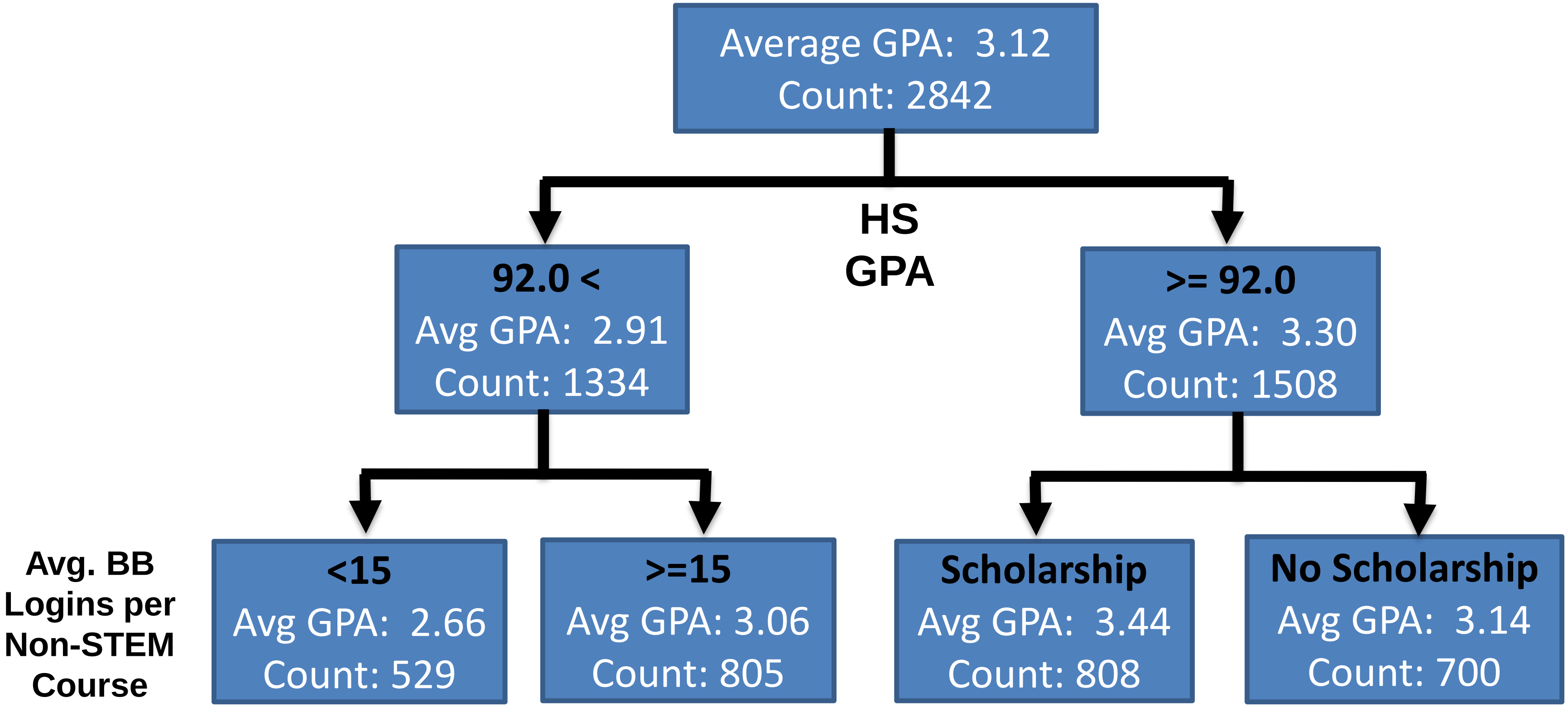
Predictive Measures	Relative Importance
High School GPA	1.000
Avg. BB* logins per non-STEM courses that use BB	0.682
Received merit scholarship (y/n)	0.544
BB non-stem course logins in week one	0.517
Total units transferred in at time of admission	0.490
BB non-STEM course logins during weeks 2 to 7	0.456
Avg. SAT critical reading score by high school (from College Board data)	0.425
Math Placement Score	0.419
BB STEM course logins during weeks 2 to 7	0.387
BB non-STEM course logins through week 7	0.350
BB STEM logins weeks 1 to 7	0.349
Avg. SAT math score by high school (from College Board data)	0.317
BB avg. logins per STEM course	0.303
Total AP STEM units	0.290
DFW STEM rate in first semester courses	0.273
Number of STEM credits with DFW rates >= 10%	0.253
Number of non-STEM credits with DFW rates >= 10%	0.244
Received Perkins loan (y/n)	0.223
Gender	0.212
Avg. SAT math and verbal score by high school (from College Board data)	0.211
Received TAP (y/n)	0.193
Amount of federal financial aid need	0.192
Amount of disbursed scholarship aid	0.183
Combined SAT math and verbal score	0.149
Advising visits during week 2 to 7	0.130
Level of math courses, e.g., MAP, calculus or higher level	0.124
SAT writing score	0.099
SAT verbal score	0.088

*BB = BlackBoard

Preliminary Decision Tree Model

Predicting First Semester GPA for First-Time Full-Time Freshmen

Average First Semester Freshmen GPA



Average GPA: 3.12
Count: 2842

HS GPA

92.0 <
Avg GPA: 2.91
Count: 1334

>= 92.0
Avg GPA: 3.30
Count: 1508

Avg. BB Logins per Non-STEM Course

<15
Avg GPA: 2.66
Count: 529

>=15
Avg GPA: 3.06
Count: 805

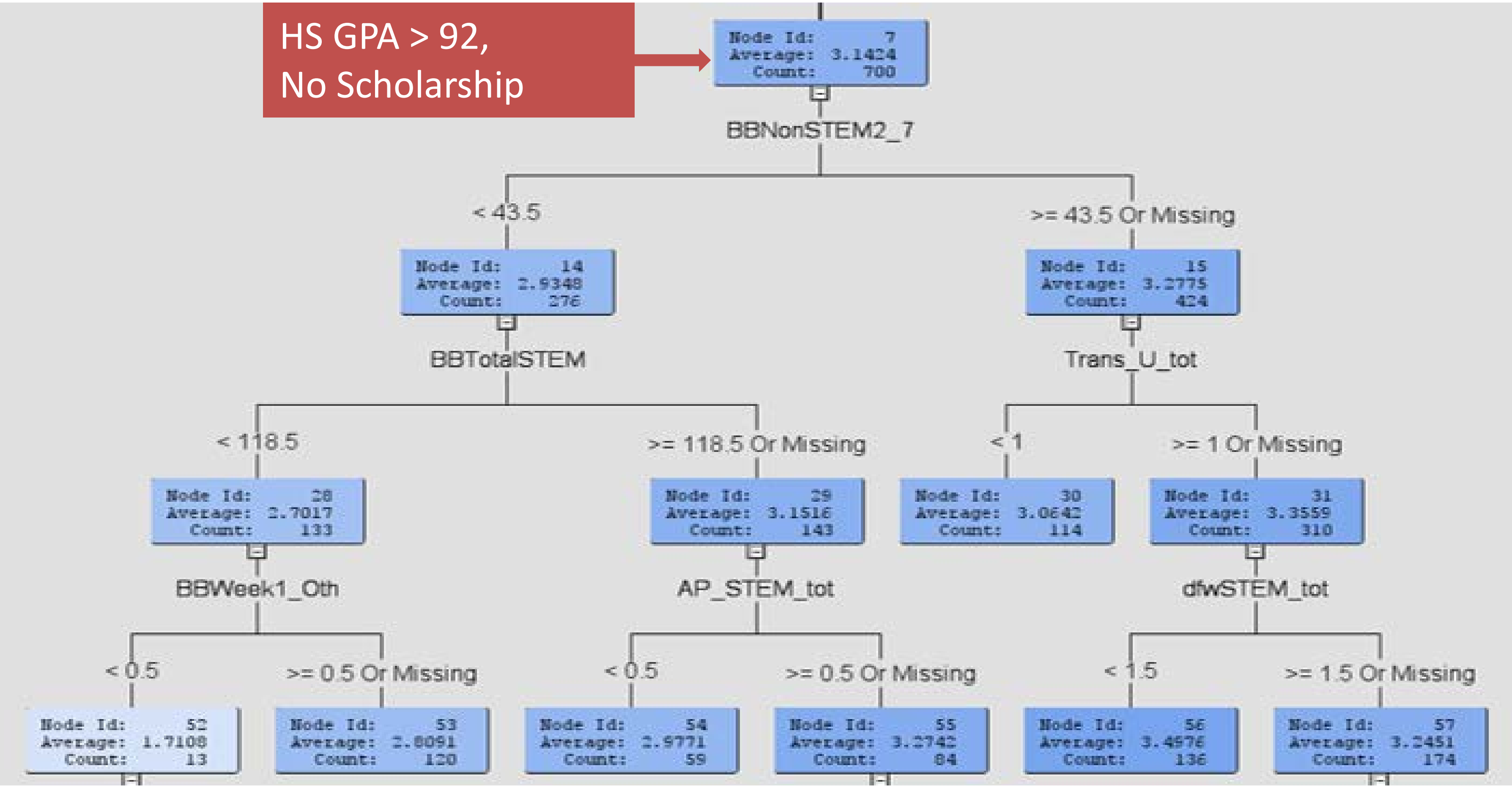
Scholarship
Avg GPA: 3.44
Count: 808

No Scholarship
Avg GPA: 3.14
Count: 700

Stony Brook University

Preliminary Decision Tree Model: First Semester GPA for First-Time Full-Time Freshmen

First Level: High School GPA; Second Level: Received Scholarship (Y/N)

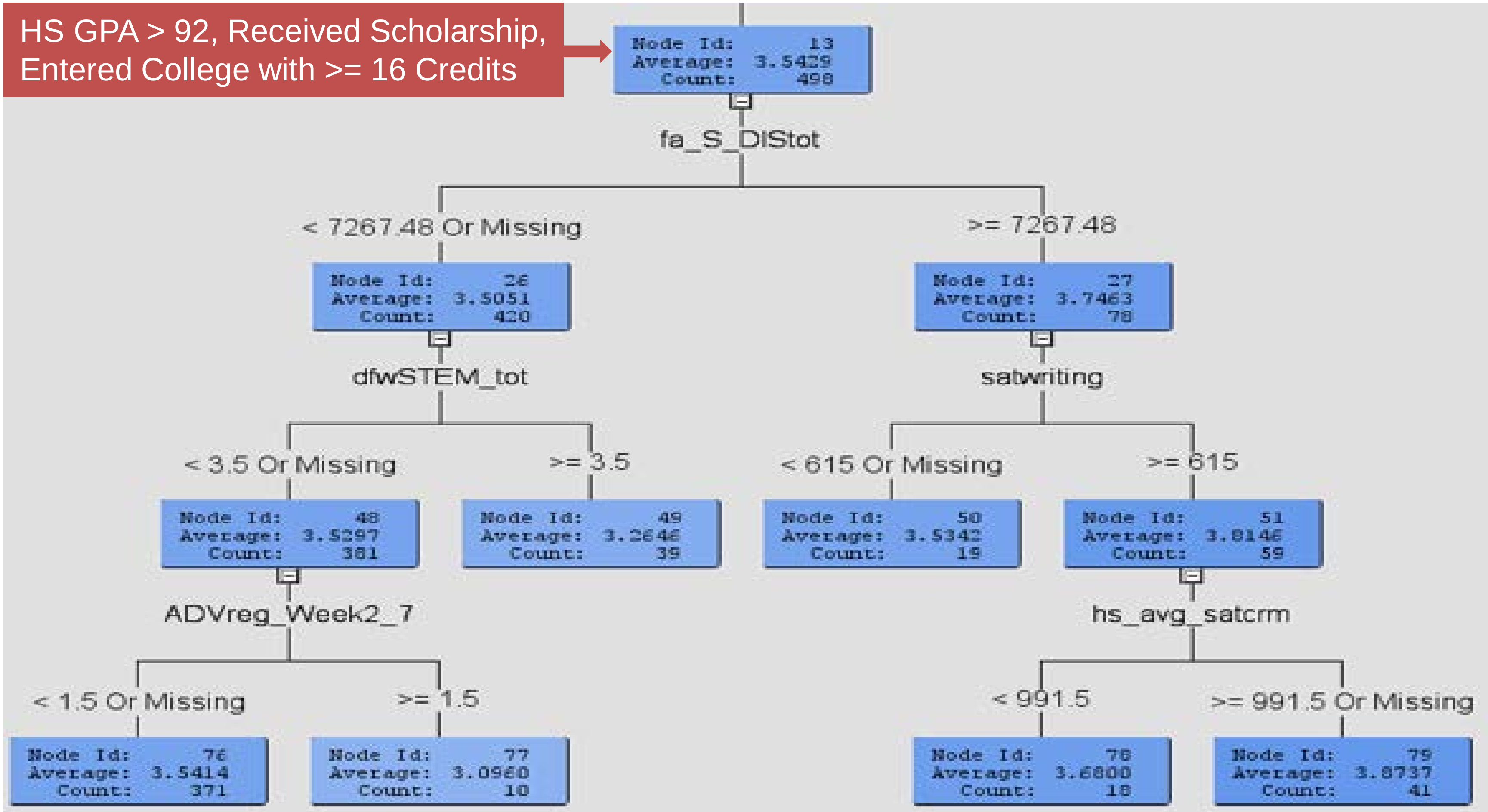


HS GPA > 92,
No Scholarship

BB = BlackBoard; DFW refers to courses/credits taken in high DFW rate courses, not the students' grades.

Preliminary Decision Tree Model: First Semester GPA for First-Time Full-Time Freshmen

First Level: High School GPA; Second Level: Received Scholarship (Y/N); Third Level: Entered College with Credits



HS GPA > 92, Received Scholarship, Entered College with >= 16 Credits

Node Id: 13
Average: 3.5429
Count: 498

fa_S_DIStot

< 7267.48 Or Missing

Node Id: 26
Average: 3.5051
Count: 420

dfwSTEM_tot

< 3.5 Or Missing

Node Id: 48
Average: 3.5297
Count: 381

ADVreg_Week2_7

< 1.5 Or Missing

Node Id: 76
Average: 3.5414
Count: 371

>= 1.5

Node Id: 77
Average: 3.0960
Count: 10

>= 3.5

Node Id: 49
Average: 3.2646
Count: 39

>= 7267.48

Node Id: 27
Average: 3.7463
Count: 78

satwriting

< 615 Or Missing

Node Id: 50
Average: 3.5342
Count: 19

>= 615

Node Id: 51
Average: 3.8146
Count: 59

hs_avg_satcrm

< 991.5

Node Id: 78
Average: 3.6800
Count: 18

>= 991.5 Or Missing

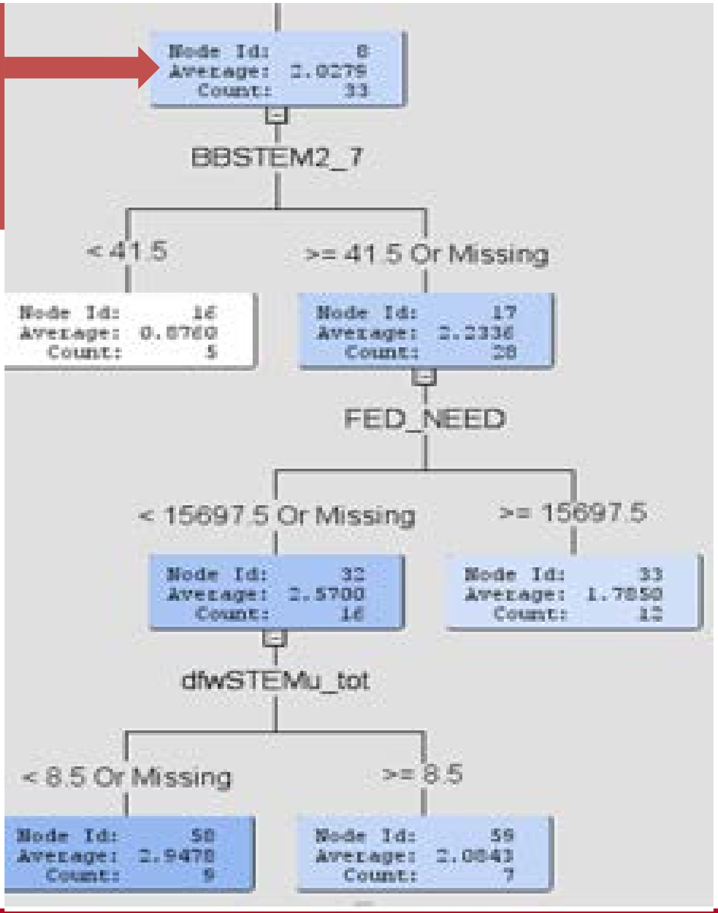
Node Id: 79
Average: 3.8737
Count: 41

ADV refers to advising visits; hs_avg_satcrm is the average SAT CR and Math Score by high school as reported by The College Board.

Preliminary Decision Tree Model: First Semester GPA for First-Time Full-Time Freshmen

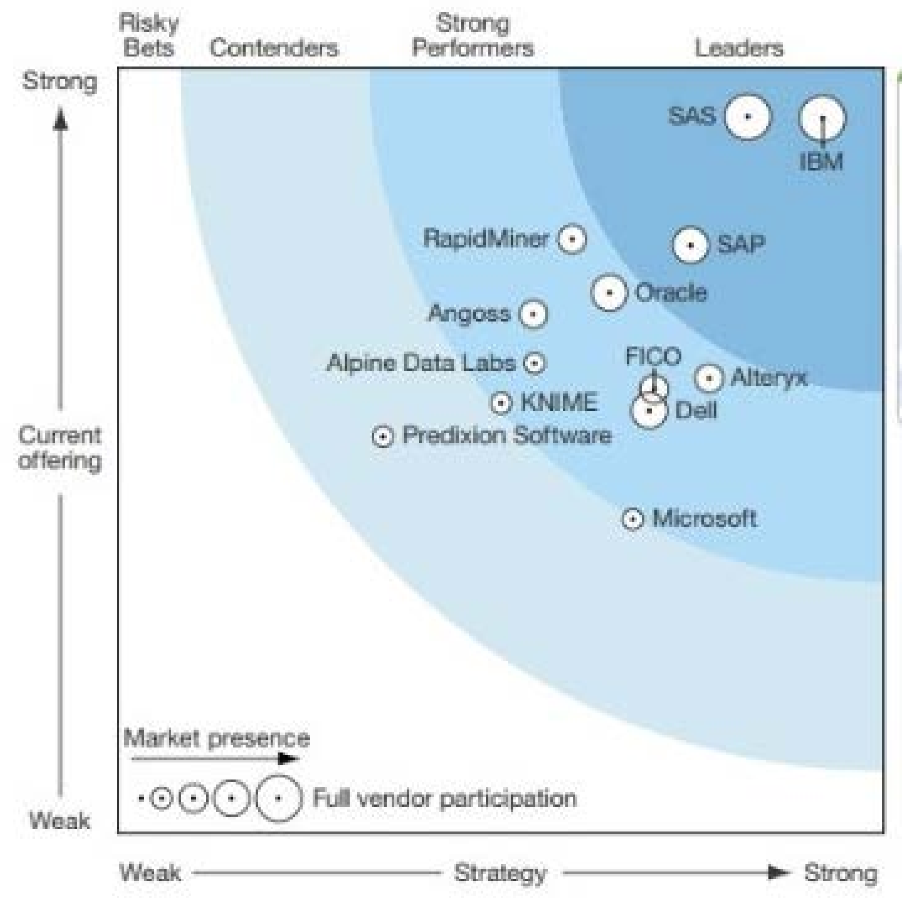
First Level: High School GPA; Second Level: Avg. BlackBoard Logins per Non-STEM Course; Third Level: BlackBoard Total Non-STEM Logins

HS GPA <= 92, Avg. BlackBoard Logins per non-STEM Course <15.1, Total BlackBoard non-STEM Logins < 8.5



Predictive Analytics Rankings: The Forrester Wave

Figure 3 The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2

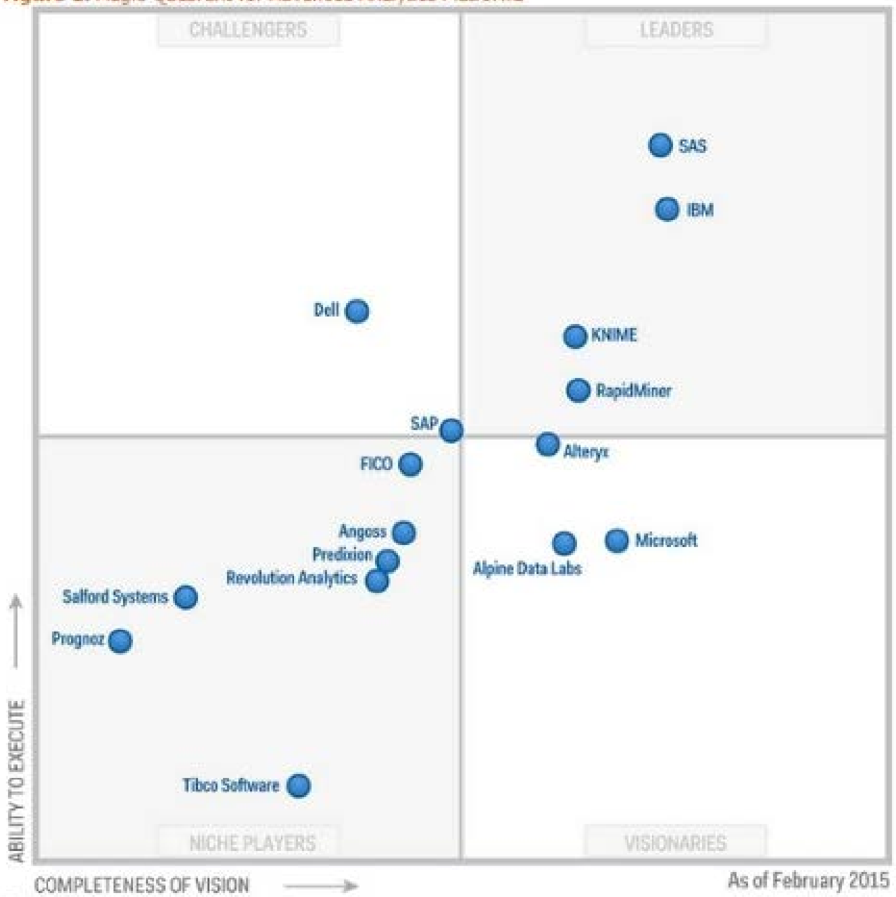


http://www.forbes.com/sites/louiscolumbus/2015/05/25/roundup-of-analytics-big-data-business-intelligence-forecasts-and-market-estimates-2015/

Forrester Research (Nasdaq: FORR) is an influential research and advisory firm. They work with business and technology leaders to develop “customer-obsessed” strategies that drive growth.

Predictive Analytics Rankings: Gartner Magic Quadrant

Figure 1. Magic Quadrant for Advanced Analytics Platforms



Source: Gartner (February 2015)

http://www.forbes.com/sites/louiscolumbus/2015/05/25/roundup-of-analytics-big-data-business-intelligence-forecasts-and-market-estimates-2015/

Positioning Technology Players Within a Specific Market: Gartner, Inc. (NYSE: IT) is a leading information technology research and advisory company. Gartner delivers the technology-related insight for client decision-making.

Software Vendors

SAS Enterprise Miner

http://www.sas.com/en_us/software/analytics/enterprise-miner.html

SPSS Modeler

http://www-01.ibm.com/software/analytics/spss/products/modeler/index.html

Rapid Miner

https://rapidminer.com/

Salford Systems

https://www.salford-systems.com/