Stony Brook University

GETTING STARTED WITH DATA MINING

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

AIR Forum 2017
Washington, D.C.

FAR BEYOND

Using Data Mining for Predicting Outcomes

- 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 rather than finding 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.

Why Should I Do the Data Mining In-house?

- **In-house modeling is cost effective**
- Some software is free and others are only $1,000 to $5,000 to license each year, as opposed hundreds of thousands of dollars for a consulting firm
- Many consulting firms require a lot of man hours for an extended period of set up in order to pull the data together
- Can obtain online or in-house training to teach staff to use the data mining software
- Once the model is set up, it can easily be rerun, and the modeling can be expanded to a variety of different needs
- If there are budget cuts, you will lose the consultants and all of the work, while the in-house model remains in place

What Do I Need to Do to Get Started?

- Data Mining Software

  Need to research software; download trial versions; explore available training

- Access to data

  Contact the IT personnel in charge of your LMS system or other data structures you want to access.

  Try to arrange for data downloads immediately, so it is ready to be used when the software is obtained

- Storage and Data Delivery Systems

  Obtain estimates for the size of the files.

  Initially, external hard drives or cloud storage should be sufficient

Predictive Measures

- **Demographics**
  - Gender, ethnicity, geographic residence when admitted.
- **Pre-college academic characteristics**
  - SAT scores, high school GPA, average SAT scores of the high school (to control for high school GPA).
- **College academic characteristics**
  - Credits accepted when admitted, AP credits, number of STEM and non-STEM courses enrolled in, area of major, enrollment in courses with high rates of D's, F's, or W's.
- **Transactions, service utilization, activities**
  - Learning management system (LMS) logins, advising visits, tutoring center utilization, intramural and fitness class participation.
- **Financial aid**
  - Expected family contribution, AGI, types and amounts of disbursed aid, Pell, Tuition Assistance Program (TAP).

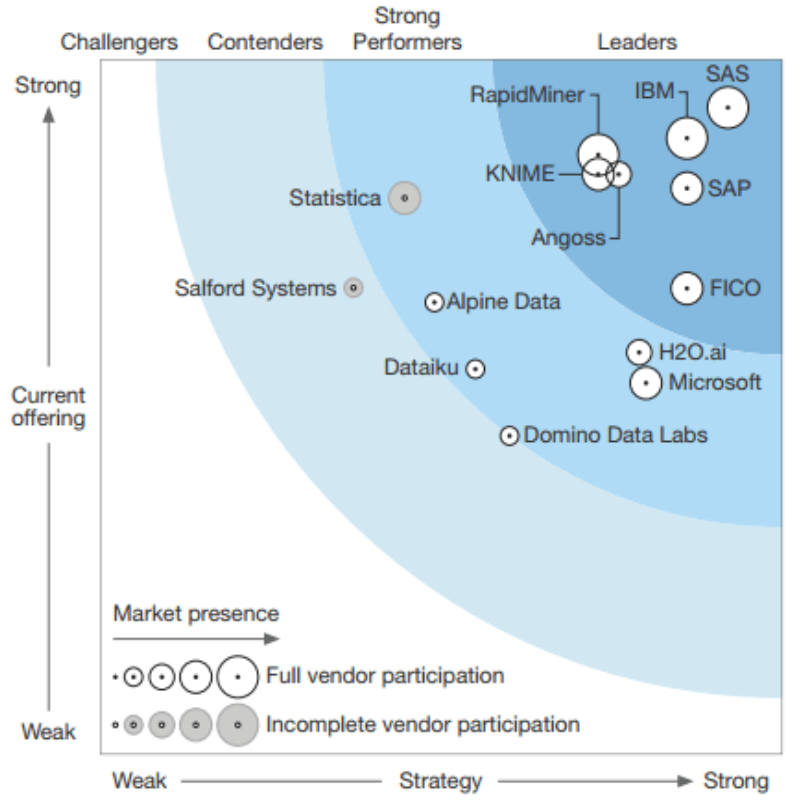
Choosing Data Mining Software

- SAS Enterprise Miner (free demos and trials on request)
- SPSS Modeler (30 day trial)
- Salford Systems, acquired by Minitab; 10-day free trial, which can be extended to 30 days on request
- KNIME (free download)
- R (free download)
- Orange (free download)
- Weka, from the University of Waikato in New Zealand (free download)
- Rapid Miner (10,000 records with free download; unlimited records for higher ed)

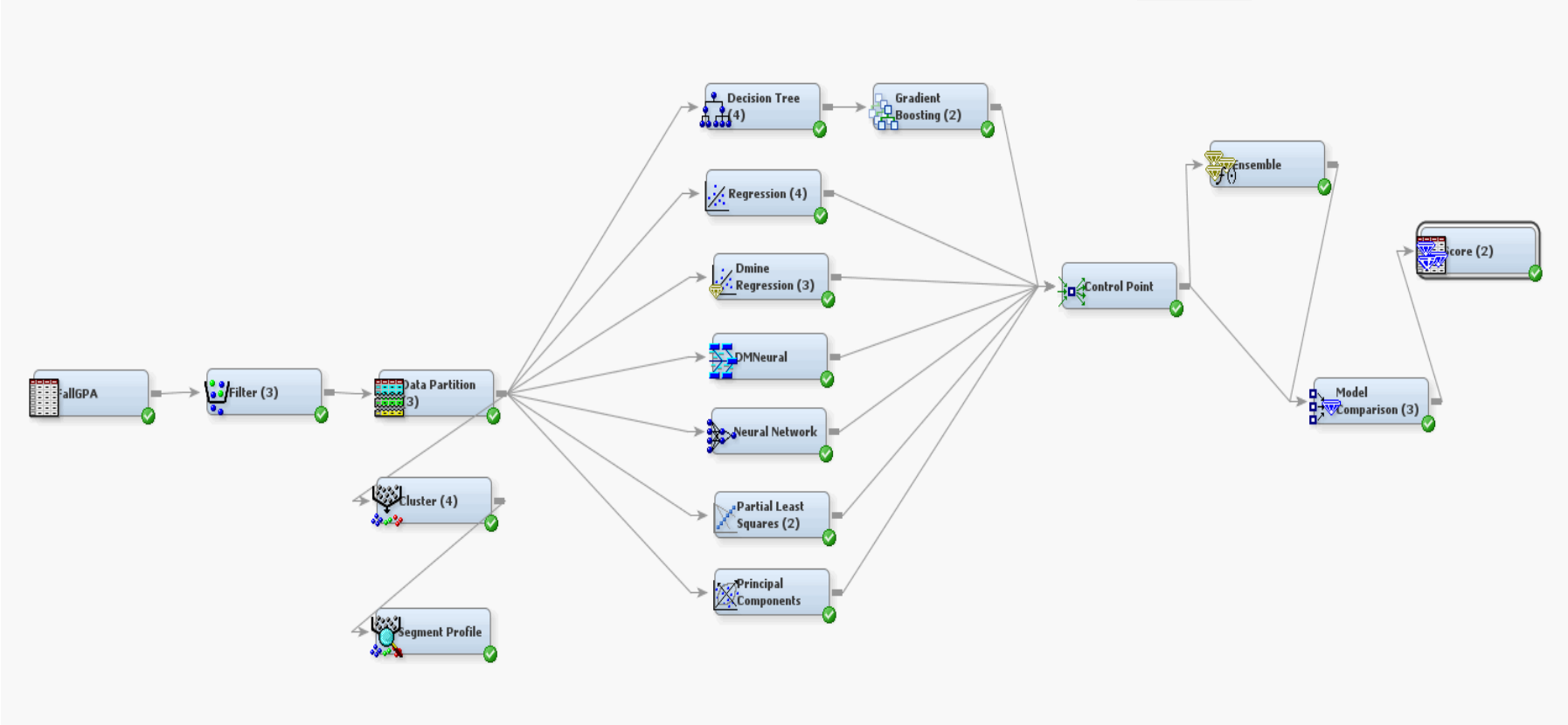
Training Options

- SAS
  - Sample data, support communities, free tutorials and e-learning, courses in major cities, live web classes
- SPSS
  - Online courses
- KNIME
  - Youtube videos, webinars, courses
- Salford Systems
  - Webinars, courses offered throughout the US, private on-site instruction, consulting
- RapidMiner
  - E-books, whitepapers, on demand webinars
- Oranges
  - Online tutorial, Youtube videos
- Weka
  - Online University of Waikato courses
- R
  - Online examples and tutorials

Forrester Wave and Gartner Magic Quadrant for Data Science Platforms



SAS Enterprise Miner Model Diagram



Nodes: Data Utilities

- Drop

  Drop variables and trim size of data sets

- Impute

  Missing data imputation using a variety of methods

- Interactive Binning

  Quintile binning of variables that can be interactively modified and customized

- Principal Components

  Perform principal component analysis for data reduction

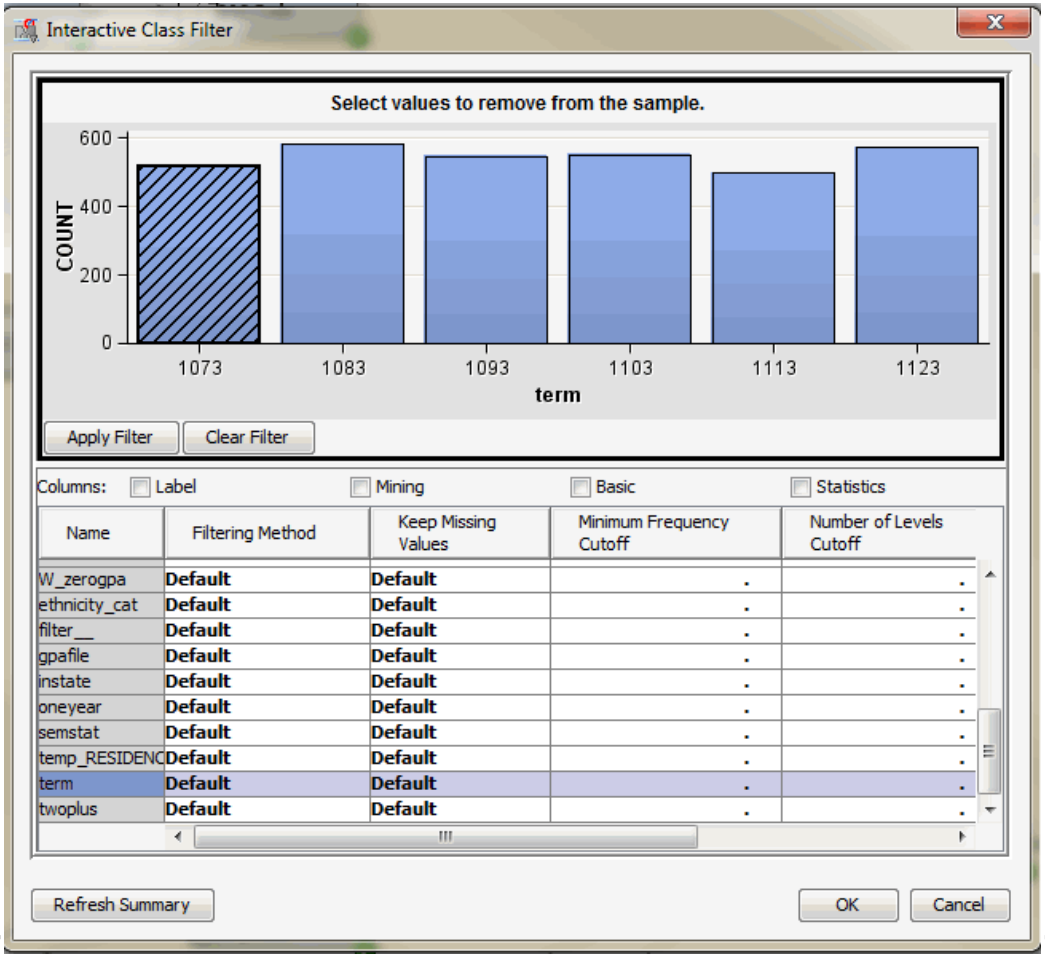
- Transform Variable

  Formula builder for transformations to correct problems like non-normality, non-linearity, stabilize variance, etc.; create interaction variables

Node Data Functions

- Sample
  - Select random samples or stratified random samples from the data
- Data Partitioning
  - Selects random samples to use for training, validation and testing samples
- Filter
  - Interactively filter subsets of data; outlier and missing data handling
- Explore
  - Descriptive statistics; scatter and box plots
- Cluster
  - Grouping of statistically similar observations
- Variable Clustering
  - Grouping similar variables

Filter Node


Interactive Class Filter
Select values to remove from the sample.
600
400
200
0
COUNT
1073
1083
1093
1103
1113
1123
term
Apply Filter
Clear Filter
Columns:
Label
Mining
Basic
Statistics
Name
Filtering Method
Keep Missing Values
Minimum Frequency Cutoff
Number of Levels Cutoff
W_zerogpa Default Default . .
ethnicity_cat Default Default . .
filter__ Default Default . .
gpafile Default Default . .
instate Default Default . .
oneyear Default Default . .
semstat Default Default . .
temp_RESIDENC Default Default . .
term Default Default . .
twoplus Default Default . .
Refresh Summary
OK
Cancel

Assess Nodes

- Model comparison

  Generates comparisons of different modeling methods to determine the best fit and lowest error rate

- Score

  Exports code that can be used to score new data

- Bagging

  Bootstrap aggregation

- Boosting

  Boosted bootstrap aggregation

- Credit scoring

  Assigns scoring points to customer attributes

- Incremental response

  Measures the impact of a treatment, e.g., impact of a response during a promotion

- Text analysis

  Some software has an additional charge for text analysis modules

Methods: Modeling

- Decision Trees
  - Chi Square Automatic Interaction Detection (CHAID)
  - Classification and Regression Trees (CART)
  - Random Forests
- Linear regression
  - Linear regression is an available data mining modeling tool, however it is important to be mindful of missing data and multicollinearity.
  - Unlike decision tree methods linear regression, will listwise delete the missing values. Fortunately, the imputation node can handle that problem
  - Multicollinearity can be handled by a variable clustering node.
- Neural net
- LARS
  - Least angle regression
- Survival analysis
- Time series

Methods: Validation

- 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. A poor fitting model will not do a good job in making predictions using new data.
- Partitioning is used to avoid over- or under-fitting. Divide the data into training, validation, and testing, or use K-fold cross validation.
- 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 ***test*** partition is used for evaluating how the model will work on new data.

Data Partitioning

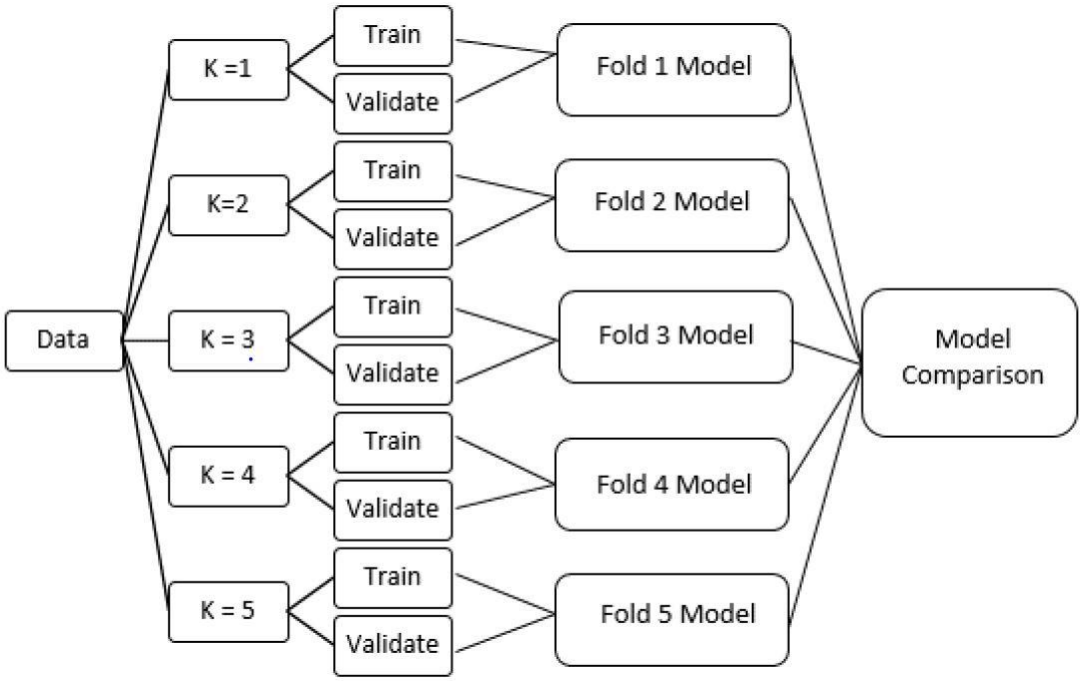
Property	Value
General	
Node ID	Part8
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	60.0
Validation	40.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	3/31/17 5:13 PM
Run ID	3a3eab47-781d-4cb2-9afb-8fe8dadfb9e0

General

M-fold Cross-validation for Evaluating Model Performance

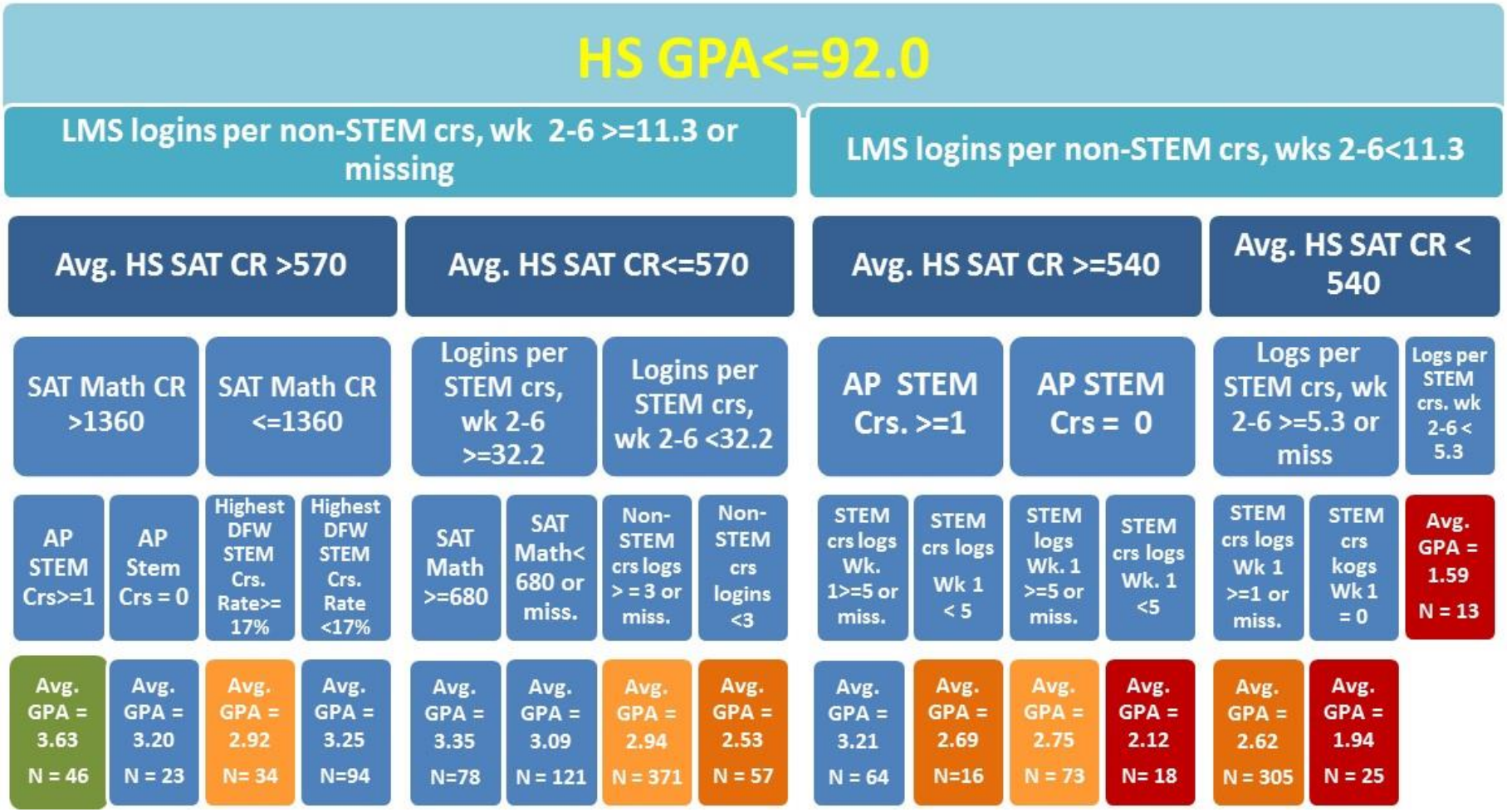
- The sample is divided into M equal groups, or folds. Many sources recommend 10 folds if there is enough data.
- Next the model is run M times, however each time, one fold is left out.
- For five folds, four are for training and one is for validation
- The procedure is performed M times (in this case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

Cross Validation Diagram



Stony Brook University

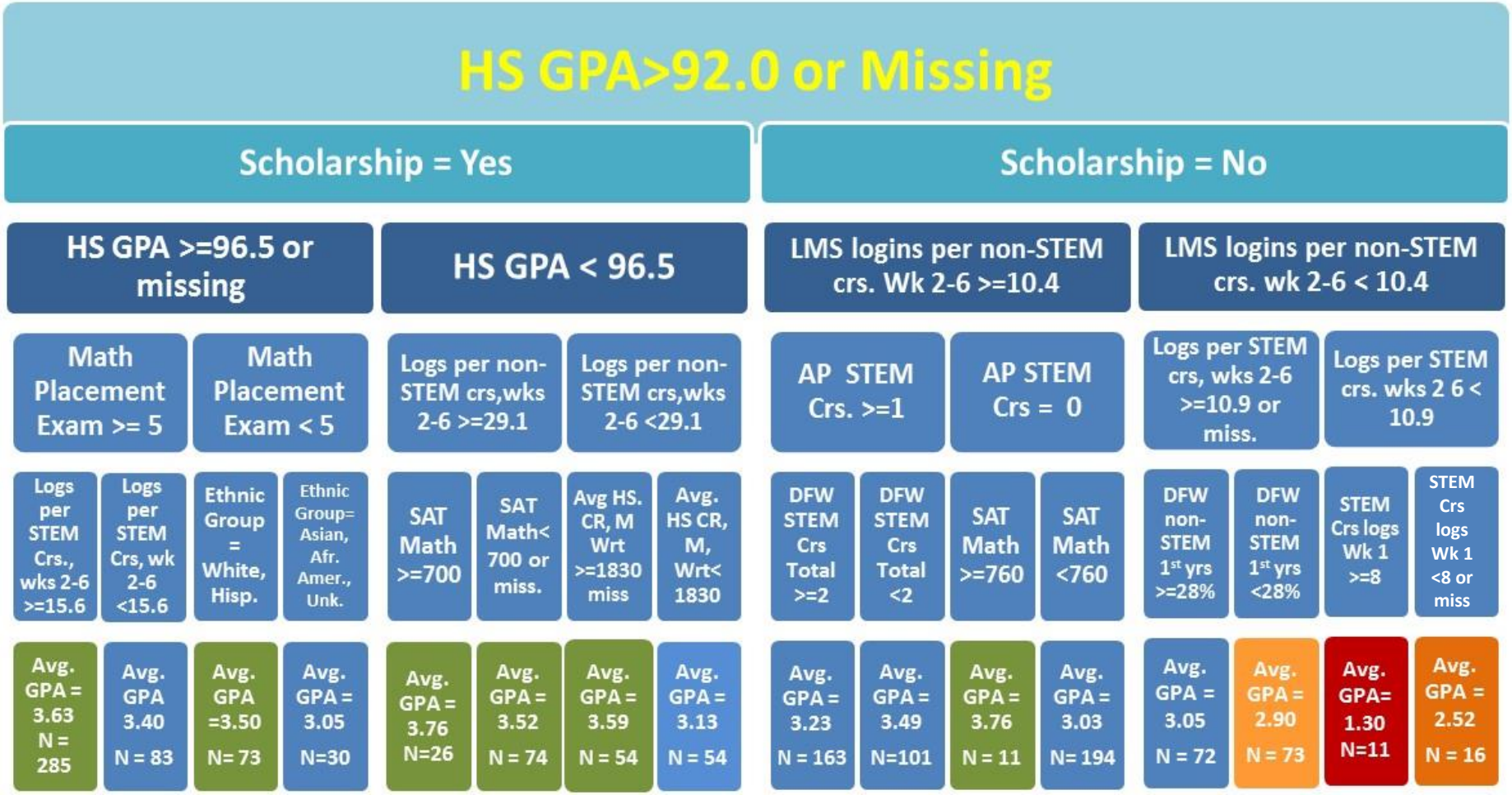
PREDICTING F15 FRESHMEN GPA: Part 1—All HS GPA Nodes<= 92.0



FAR BEYOND

Stony Brook University

PREDICTING F15 FRESHMEN GPA: Part 2—All HS GPA Nodes > 92.0



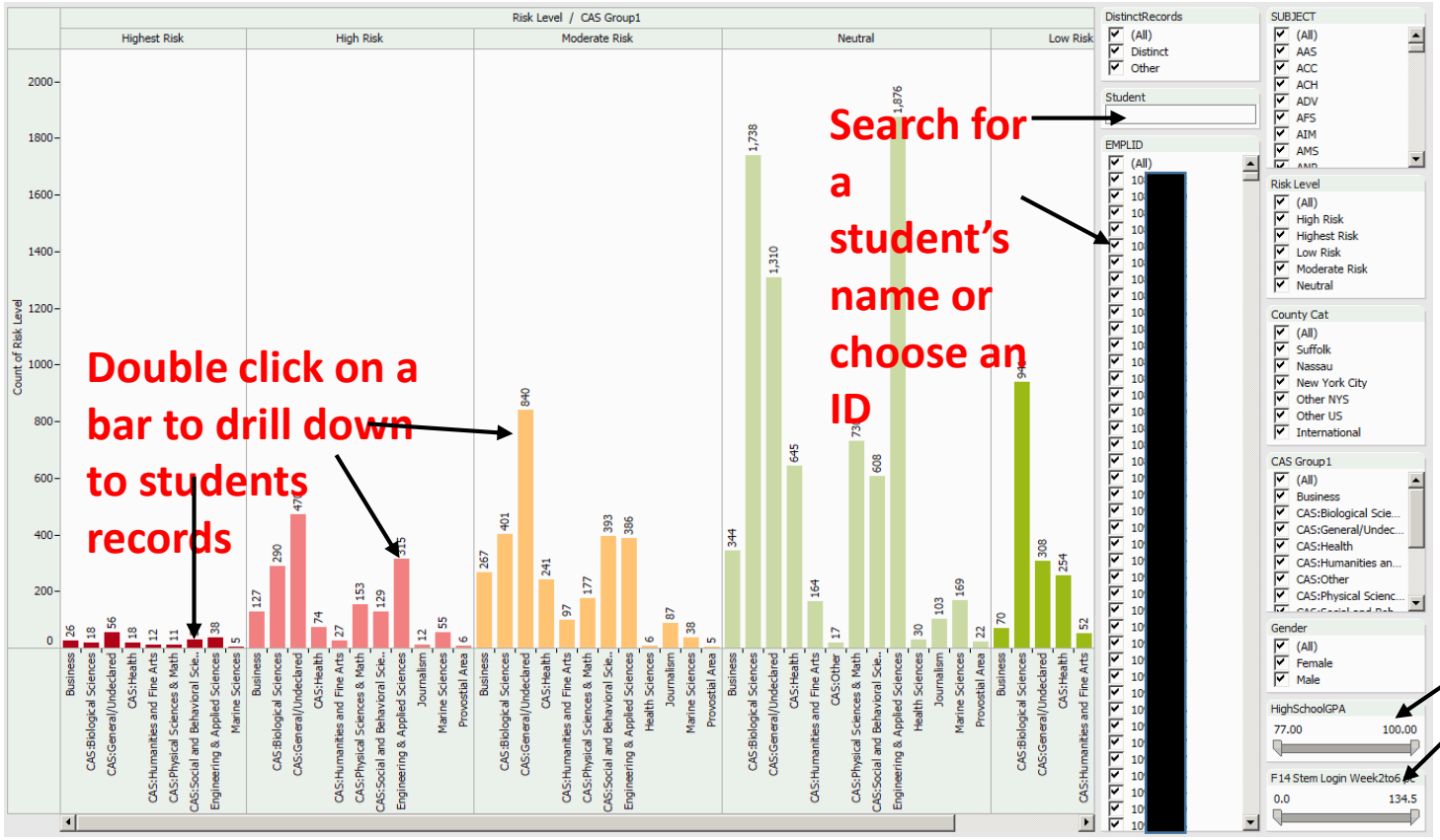
FAR BEYOND

How Can the Results Be Used?

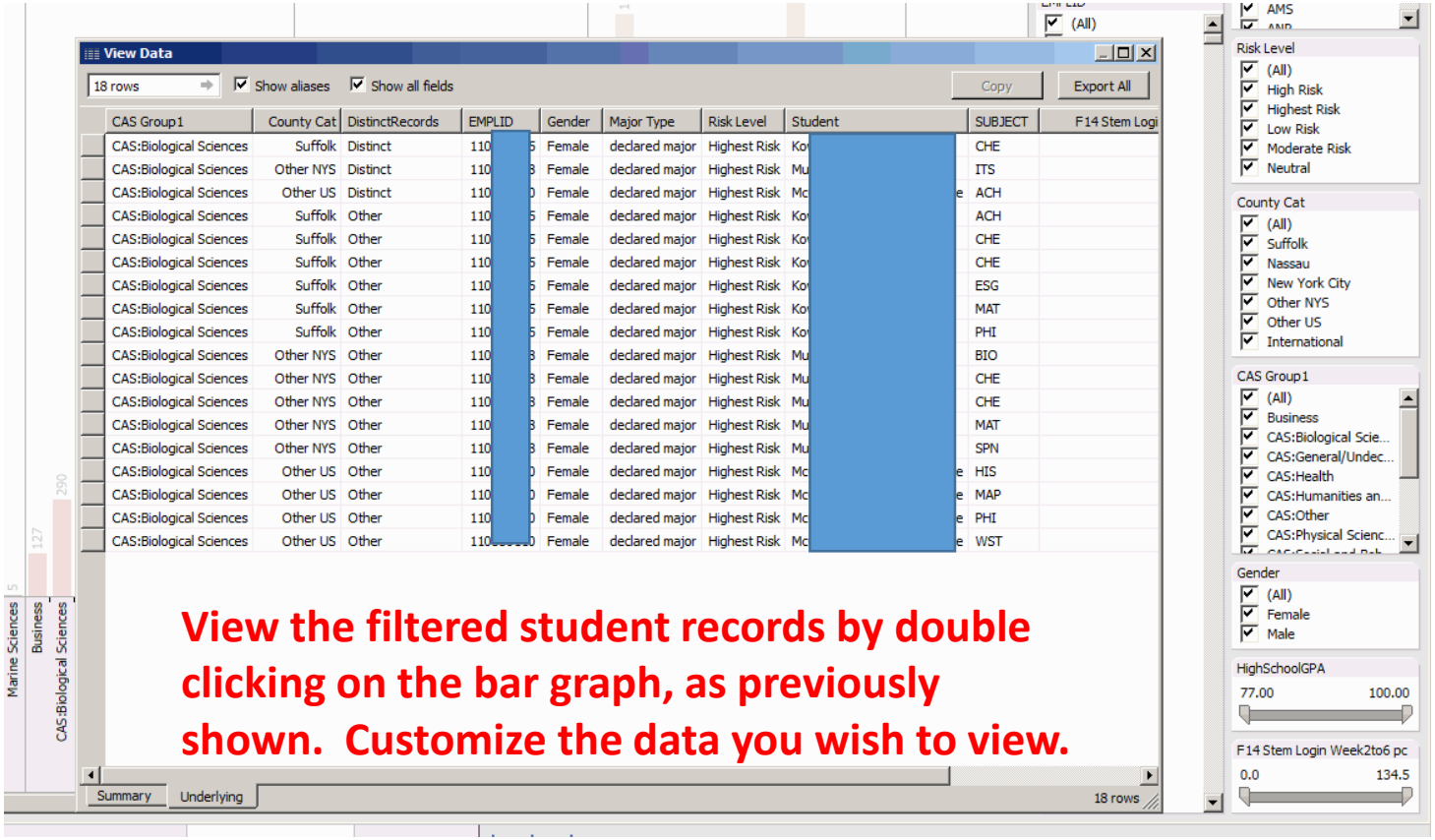
- The models as presented can be used to assign students to designed interventions.
  - The results were distributed to appropriate stakeholders.
  - Students were assigned to interventions such as tutoring, pairing them up with peer mentors, and sending communications from campus advising.
- The early model results can be shared with departments to inform their advising and intervention efforts.
- *The goal is to find the students who need assistance to fulfill their potential, and reduce the number who end up leaving due to poor performance.*

Sample Dashboard

Dashboards for Results Delivery



Drilling Down to Student Data



View the filtered student records by double clicking on the bar graph, as previously shown. Customize the data you wish to view.

Data Storage

- Initially when getting started, use an external hard drive or cloud storage, like Google Drive.
- Uploading and downloading to a Google Drive can be slow when down on a regular basis.
- Until a system and repository is in place, departments and/or individuals will need to be contacted regularly to obtain the transaction data. This will become tedious and time consuming.

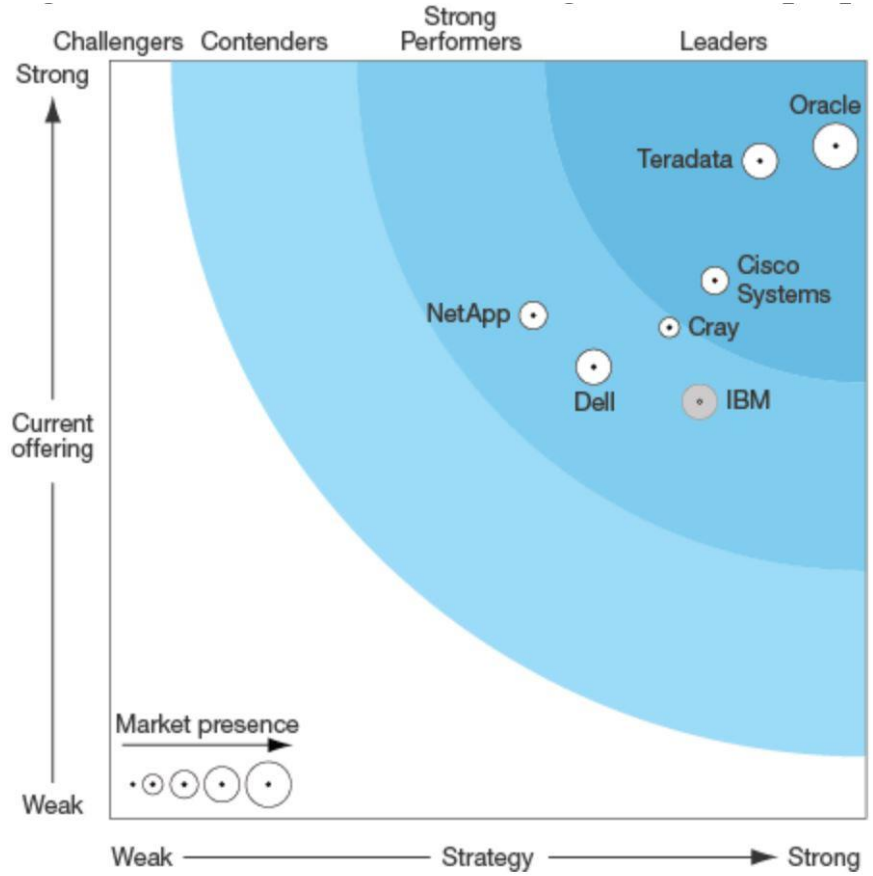
Data Storage: Hadoop

- Hadoop platforms handle the 3 V's of data:
  - Large **volumes** of data
  - A **variety** of data
  - High **velocity** data that has a small window of utility
- Stores data on large clusters of affordable hardware.
- Applications are divided between the machines in the Hadoop cluster. Several computers can share the computational workload.

Hadoop Systems

- Hadoop distributions
  - Cloudera (www.cloudera.com)
  - EMC (www.gopivotal.com)
  - Hortonworks (www.Hortonworks.com)
  - IBM (www.ibm.com/software/data/infosphere/biginsights)
  - Intel (hadoop.intel.com)
  - MapR (www.mapr.com)
- Hadoop toolboxes—tools to use with Hadoop implementations
  - Amazon (aws.amazon.com/ec2)
  - Hadapt (www.hadapt.com)
  - Karmasphere (www.karmasphere.com)
  - WANdisco (www.wandisco.com)
  - Zettaset (www.zettaset.com)

Forester Wave: Big Data Hadoop-Optimized Systems



Questions?
Please contact me!!

Nora.Galambos@stonybrook.edu

http://www.stonybrook.edu/commcms/irpe/reports/index.php