

MINING DATA TO CREATE A SYSTEM TO IDENTIFY AT-RISK FRESHMEN

Nora Galambos, PhD
Senior Data Scientist

NEAIR Annual Conference 2016
Baltimore, MD

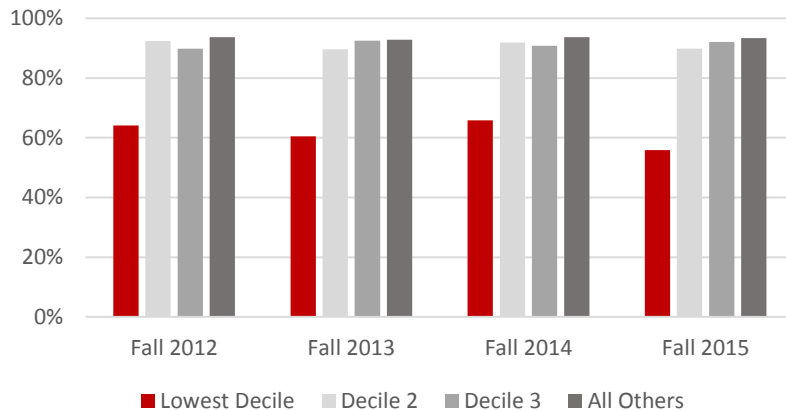
**FAR
BEYOND**

Freshmen GPA Predictions

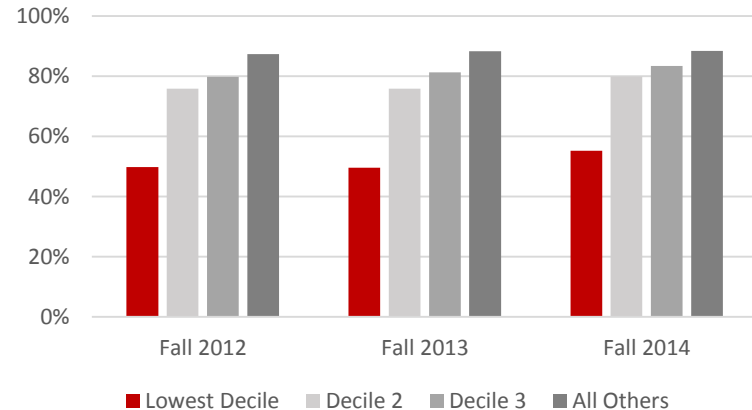
- **Goal: Predict 2016 fall freshmen GPA's at three timepoints during their first semester**
 - **1) End of orientation; 2) End of week 3; 3) End of week 6**
- Summary of data
 - *Demographics*
 - Gender, ethnicity, geographic area of residence when admitted.
 - *Pre-college academic characteristics*
 - SAT scores, high school GPA, average SAT scores of the high school (to control for high school GPA), Common Application data
 - *College academic characteristics*
 - Credits accepted when admitted, AP credits, number of STEM and non-STEM courses enrolled in, enrollment in high DFW courses, area of major
 - *Transactions, service utilization, activities*
 - Learning management system (LMS) logins, advising visits, tutoring center utilization, intramural and fitness class participation, recreation center usage.
 - *Financial aid*
 - Expected family contribution, AGI, types and amounts of disbursed aid, Pell, Tuition Assistance Program (TAP)

One- and Two-year Retention of First-time Full-time Freshmen by First Semester GPA Deciles

One-Year Retention



Two-Year Retention



Model Background

- The initial model for predicting freshmen GPA's was developed in 2015-16 using only one year of data.
- The fall 2014 freshmen cohort data was used to predict the first semester GPA's of the fall 2015 freshmen cohort. Learning management system (LMS) logins were to be incorporated, however the login data were not being archived, so there was no possibility of using multiple years of data.
- One model was developed to predict first term GPA's at week six of the first semester.
- Five data mining models were developed using different methods, including CART, CHAID, and gradient boosting.
- Gradient boosting, CART, and CHAID had the lowest average squared errors, in that order.
- The CART model was the method selected for predicting the fall 2015 freshmen GPA's, because gradient boosting does not yield an easy to use and understand algorithm, coupled with the fact that the gradient boosting model did not result in a substantial error rate improvement.

Development of Current Model

- Fall 2014 and fall 2015 first-time full-time freshmen cohort data were used to predict the fall 2016 GPA's. N= 5,664 (after 34 students who withdrew prior to the end of the term were removed).
- The extensive modeling work on the fall 2014 cohort data was utilized to motivate the development of the three new models
- 5,000 plus observations are not enough for partitioning into training and validation sets to avoid over-fitting the model, so K-fold cross validation was used instead.
- In K-fold cross validation, the data are subdivided into K equal groups. K-1 groups are for training and the remaining group is for validation. This is done K times. Each time a different group is used for the validation set. Often, five folds are used.
- Models were compared using averaged squared errors.

5-Fold Cross Validation Plan

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



CART Method: Classification and Regression Trees

- The method does an exhaustive search for the best binary split.
- It splits categorical predictors into a smaller number of groups or finds the optimal 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 used 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
- CART creates a set of nested binary decision rules to predict an outcome.

CHAID Method: 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 categorical 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.

Cross Validation Results: Average Squared Error (ASE) for Freshmen GPA Models at Three Timepoints

K Folds	Day 1 Model (CHAID)		Week 3 (CART)		Week 6 (CART)	
	Validation ASE	Training ASE	Validation ASE	Training ASE	Validation ASE	Training ASE
1	0.46	0.41	0.44	0.46	0.43	0.46
2	0.48	0.41	0.45	0.44	0.43	0.44
3	0.50	0.40	0.45	0.46	0.44	0.43
4	0.51	0.40	0.45	0.43	0.46	0.43
5	0.56	0.39	0.51	0.43	0.51	0.42
Average ASE	0.50	0.40	0.46	0.44	0.45	0.44

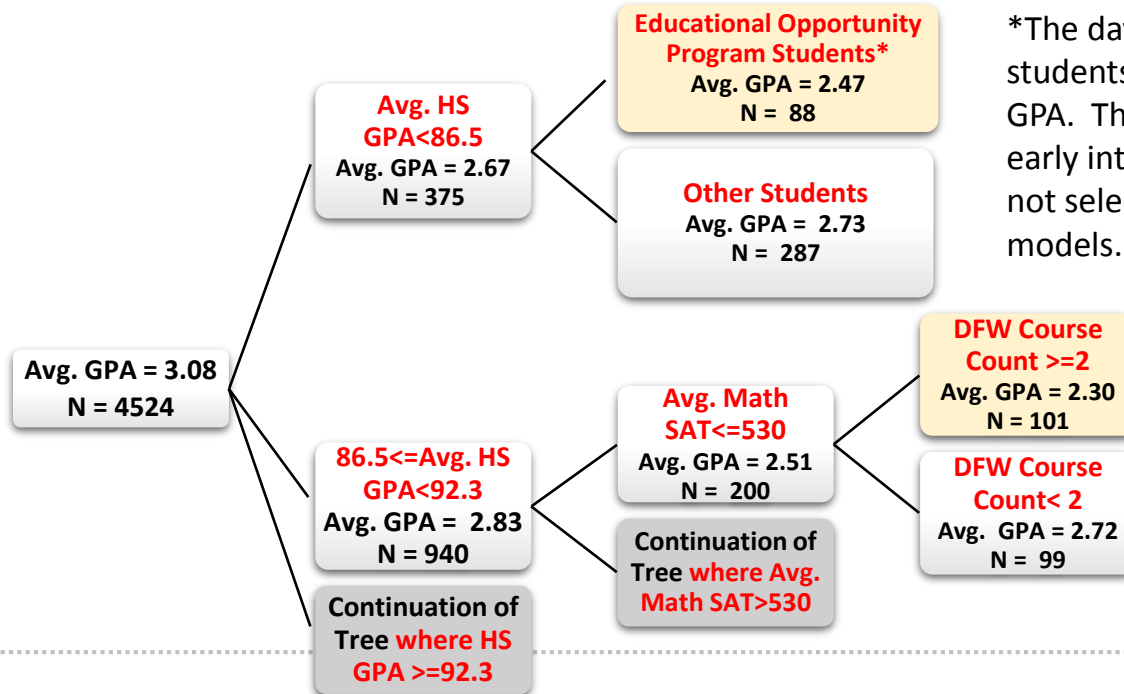
$$\text{ASE} = (\text{Sum of Squared Errors})/N$$

Fall Semester Day 1: First Term GPA CHAID Model

Students Suggested for Interventions

Avg. GPA < 2.5
& ≥ 2.0 .

Avg. GPA < 2.0



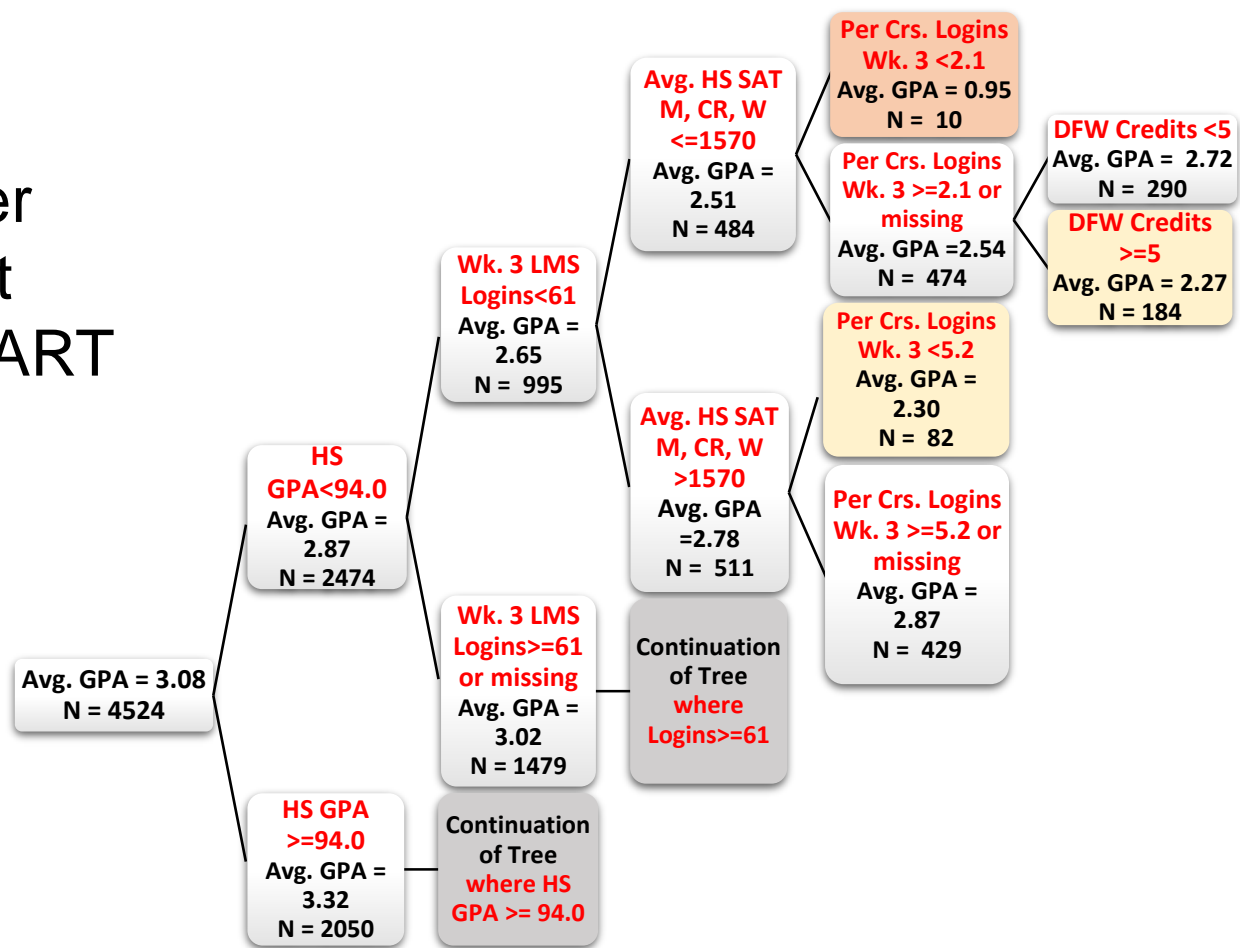
*The day 1 model selected EOP students based on their low HS GPA. They were assigned to early interventions and were not selected in subsequent models.

Fall Semester Week 3: First Term GPA CART Model

Students Suggested for Interventions

Avg. GPA < 2.5
& ≥ 2.0

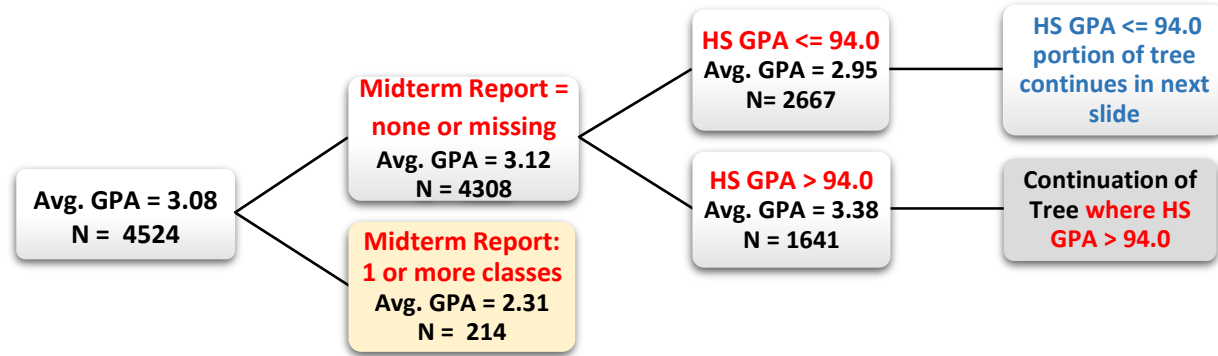
Avg. GPA
< 2.0



Fall Semester Week 6: First Term GPA CART Model, Part 1

Students Suggested for Interventions

- Avg. GPA < 2.5
& ≥ 2.0
- Avg. GPA
< 2.0

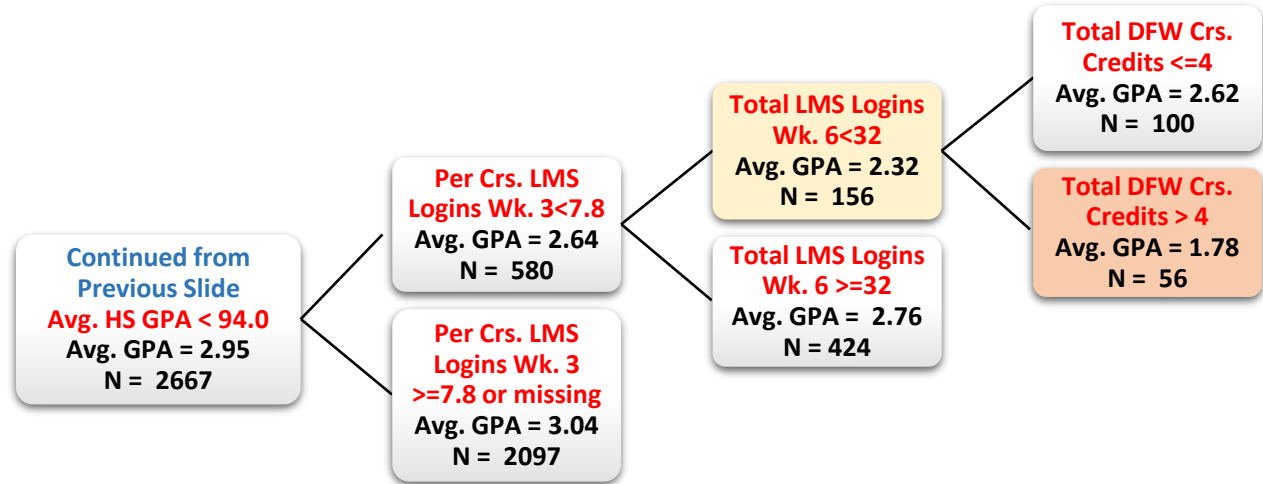


Midterm Report:
Midterm course feedback from participating professors.
Predictive data only available for Fall 2015

Fall Semester Week 6: First Term GPA CART Model, Part 2

Students Suggested for Interventions

- Avg. GPA < 2.5
& ≥ 2.0
- Avg. GPA
< 2.0

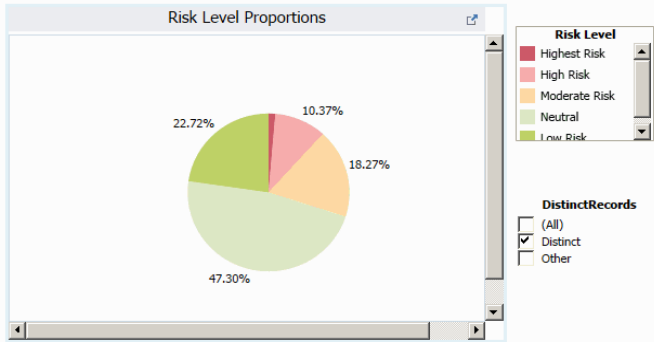


How Can the Results be Used?

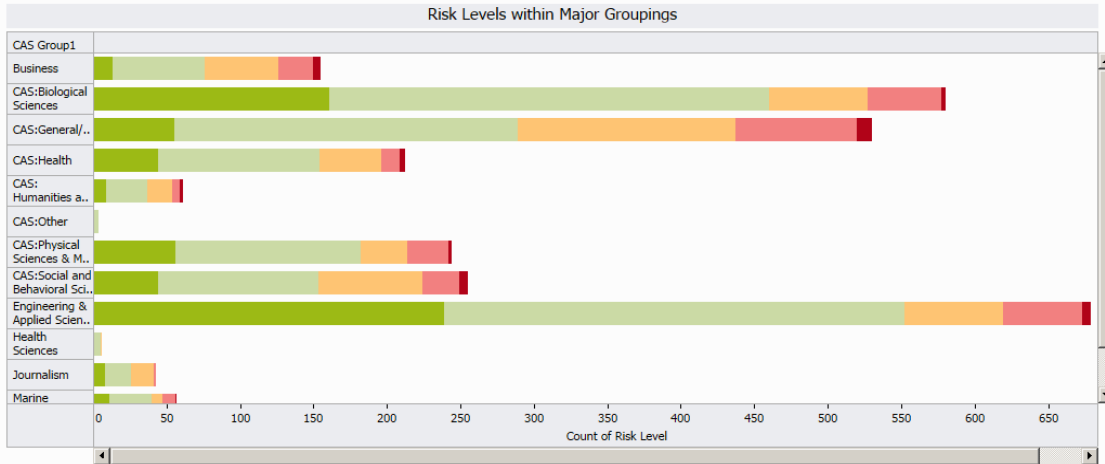
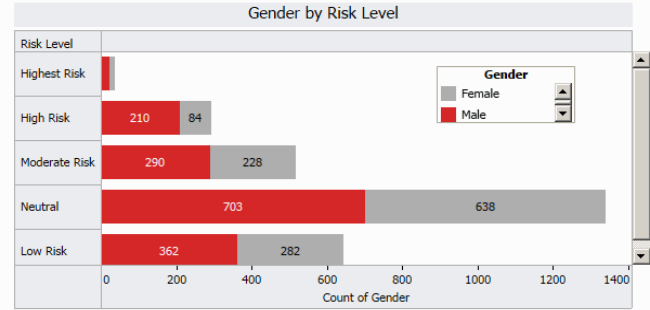
- The model as presented can be used to assign students to designed interventions.
 - The results were distributed to appropriate stakeholders.
 - Students were assigned to interventions such as tutoring, pairing them with peer mentors, and sending communications from campus advising.
- The early model 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 Model Dashboard

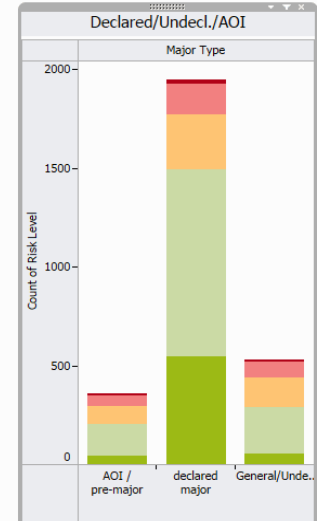
Risk Levels



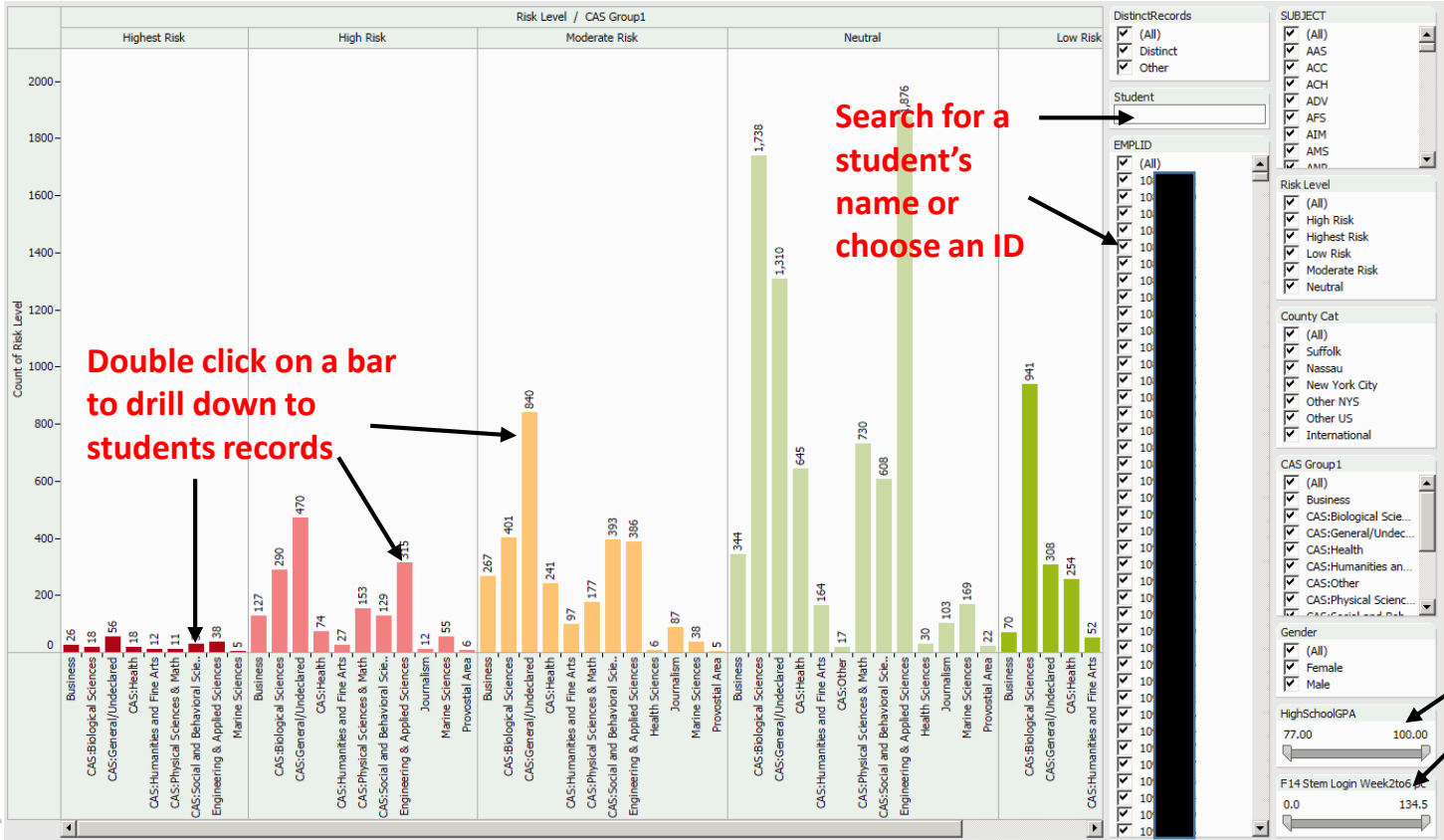
- Risk Level**
- (All)
 - High Risk
 - Highest Risk
 - Low Risk
 - Moderate Risk
 - Neutral
- Gender**
- (All)
 - Female
 - Male



- Major Type**
- (All)
 - AOI / pre-major
 - declared major
 - General/Undeclared



Sample Dashboard Filters

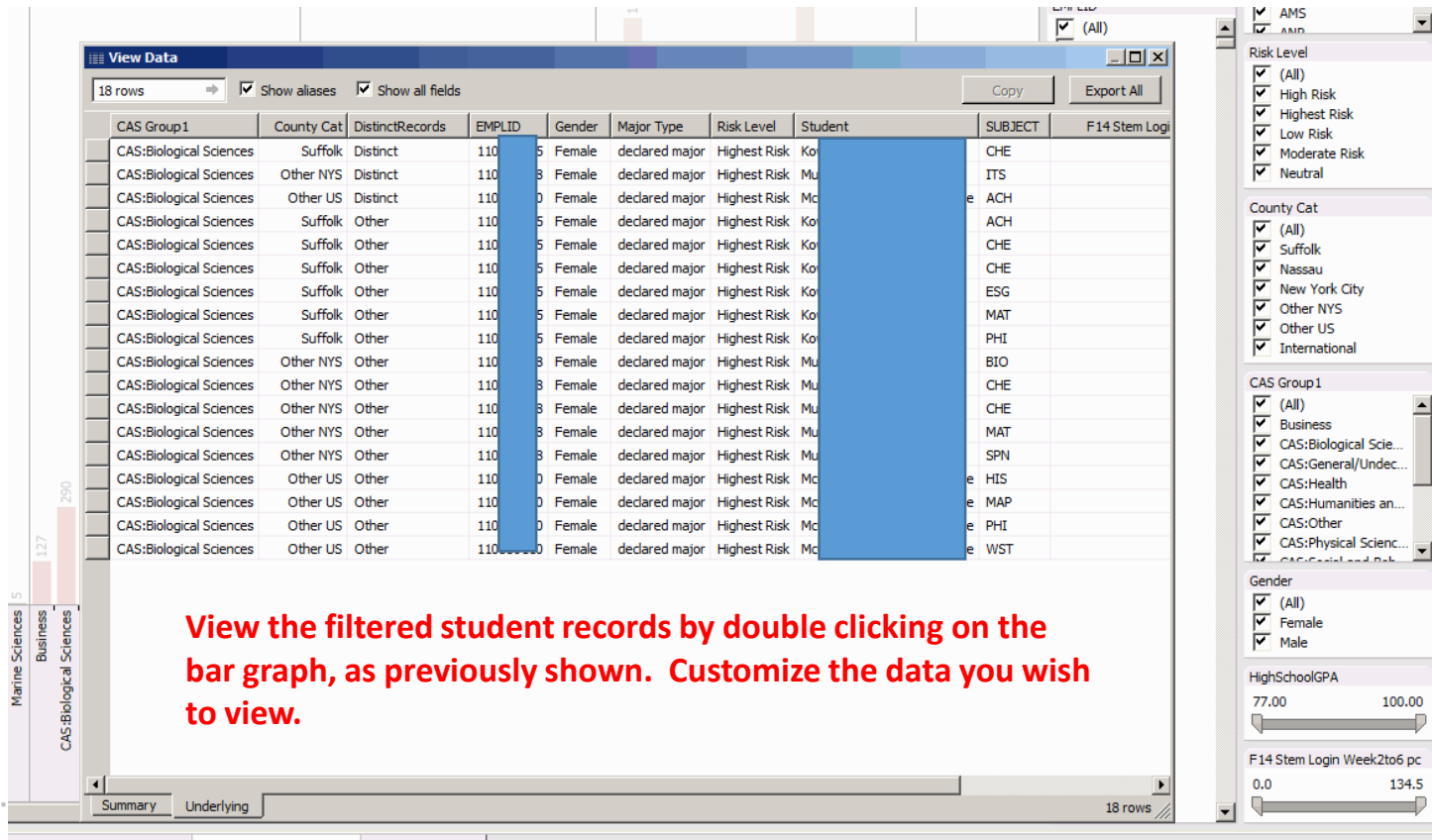


Double click on a bar to drill down to students records

Search for a student's name or choose an ID

HS GPA and LMS login sliders

Drilling Down to Student Records



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

CAS Group1	County Cat	DistinctRecords	EMPLID	Gender	Major Type	Risk Level	Student	SUBJECT	F14 Stem Log
CAS:Biological Sciences	Suffolk	Distinct	11000005	Female	declared major	Highest Risk	Ko	CHE	
CAS:Biological Sciences	Other NYS	Distinct	11000008	Female	declared major	Highest Risk	Mu	ITS	
CAS:Biological Sciences	Other US	Distinct	11000010	Female	declared major	Highest Risk	Mc	e ACH	
CAS:Biological Sciences	Suffolk	Other	11000005	Female	declared major	Highest Risk	Ko	ACH	
CAS:Biological Sciences	Suffolk	Other	11000005	Female	declared major	Highest Risk	Ko	CHE	
CAS:Biological Sciences	Suffolk	Other	11000005	Female	declared major	Highest Risk	Ko	CHE	
CAS:Biological Sciences	Suffolk	Other	11000005	Female	declared major	Highest Risk	Ko	ESG	
CAS:Biological Sciences	Suffolk	Other	11000005	Female	declared major	Highest Risk	Ko	MAT	
CAS:Biological Sciences	Suffolk	Other	11000005	Female	declared major	Highest Risk	Ko	PHI	
CAS:Biological Sciences	Other NYS	Other	11000008	Female	declared major	Highest Risk	Mu	BIO	
CAS:Biological Sciences	Other NYS	Other	11000008	Female	declared major	Highest Risk	Mu	CHE	
CAS:Biological Sciences	Other NYS	Other	11000008	Female	declared major	Highest Risk	Mu	CHE	
CAS:Biological Sciences	Other NYS	Other	11000008	Female	declared major	Highest Risk	Mu	MAT	
CAS:Biological Sciences	Other NYS	Other	11000008	Female	declared major	Highest Risk	Mu	SPN	
CAS:Biological Sciences	Other US	Other	11000010	Female	declared major	Highest Risk	Mc	e HIS	
CAS:Biological Sciences	Other US	Other	11000010	Female	declared major	Highest Risk	Mc	e MAP	
CAS:Biological Sciences	Other US	Other	11000010	Female	declared major	Highest Risk	Mc	e PHI	
CAS:Biological Sciences	Other US	Other	11000010	Female	declared major	Highest Risk	Mc	e WST	

Summary of Alert System Development and Use

1. Identify data sources on campus and begin the process of collecting, archiving and recoding.
2. Don't skimp on model development. Be sure to hold out data for validating the model.
3. Plan a system to distribute the model results and lists of students suggested for interventions.
4. Work with stakeholders to track interventions.
5. Campus service data being collected is not only useful to determine if students are having improved outcomes, but can be used to study campus service utilization, like tutoring and advising.