A Two-Step Data Mining Approach for Graduation Outcomes

2013 CAIR Conference

Afshin Karimi (akarimi@fullerton.edu)
Ed Sullivan (esullivan@fullerton.edu)
James Hershey (jrhershey@fullerton.edu)
Sunny Moon (hmoon@fullerton.edu)

November 21, 2013

Data Mining

Science of extracting patterns and knowledge from large data sets to predict future trends and behavior.

- Supervised Learning
- Unsupervised Learning

Two Step Process

- Classification decision tree model to predict six-year graduation of FTF (supervised learning)

- Cluster analysis (K-Means clustering) on the identified at-risk students to reveal patterns and suggest cluster-level intervention (unsupervised learning)

Classification Model Using Decision Tree

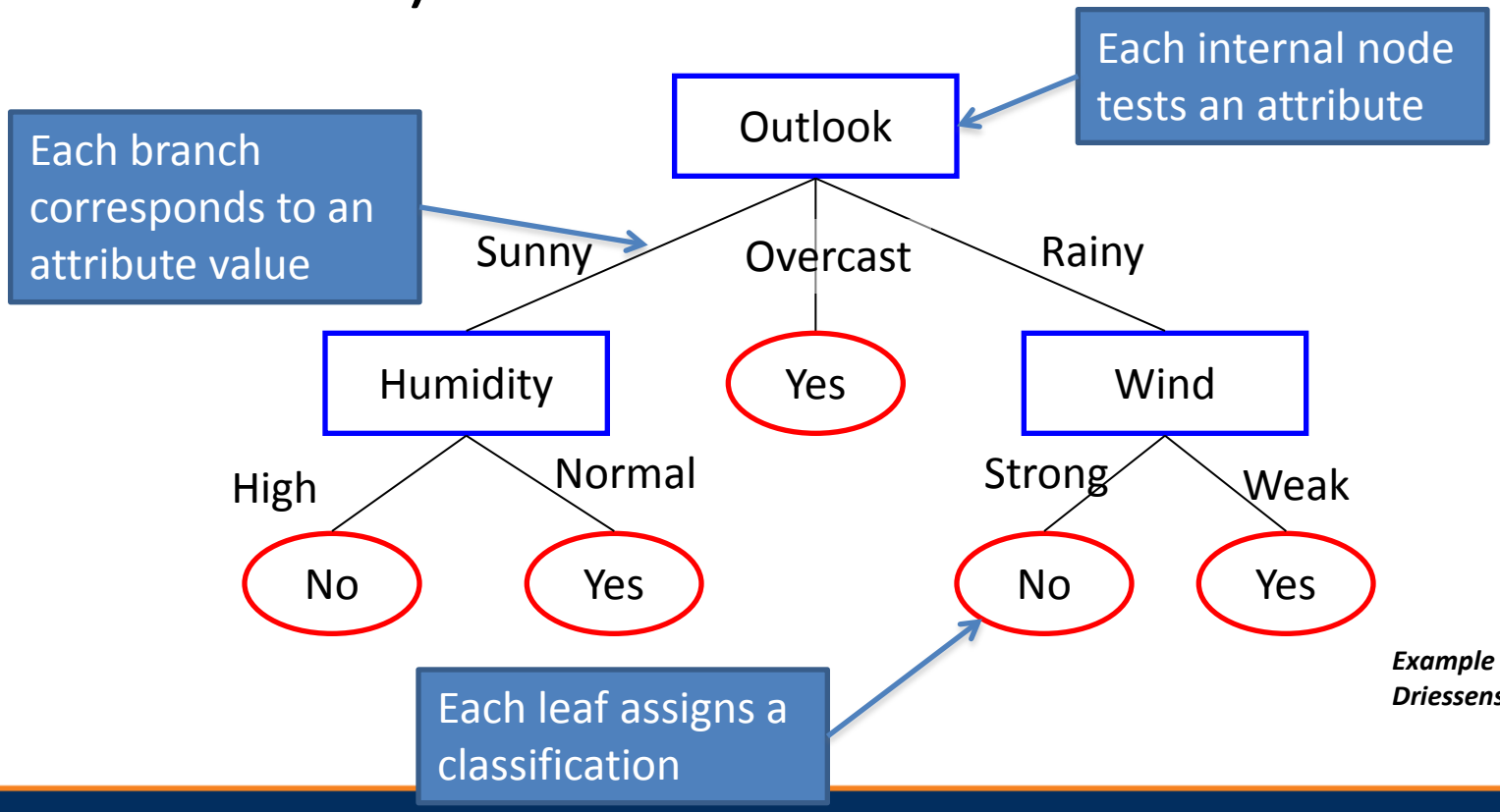
- Decision Tree vs. Neural Networks, Logistic Regression, SVM, etc.
- Decision trees are easy to understand, implement, and visualize

Decision Trees Continued…

- Used in different disciplines including Operations Research
- Inverted trees with root at the top; used to create model that predicts target variable
- Generated by recursive partitioning
- An example of node selection criteria is Information Gain (C5.0) that selects node variable with least entropy with respect to target variable

Example decision tree

- Play tennis or not? (depending on weather conditions)



Example taken from Kurt Driessens slides

Overfitting

- Generated decision tree relies too much on irrelevant feature of training data. The generated model performs poorly on future/unseen data.

- To reduce overfitting, use pruning (technique in which leaf nodes that do not add to the discriminative power of the decision tree are removed)

Training/Building the Tree

- Using 24 predictor variables:
  - 12 socio-economic, demographics, HS performance variables
  - 12 first term college variables
  - All converted to nominal variables

- 1 target variable: 6 Yr Degree (with Yes/No values)

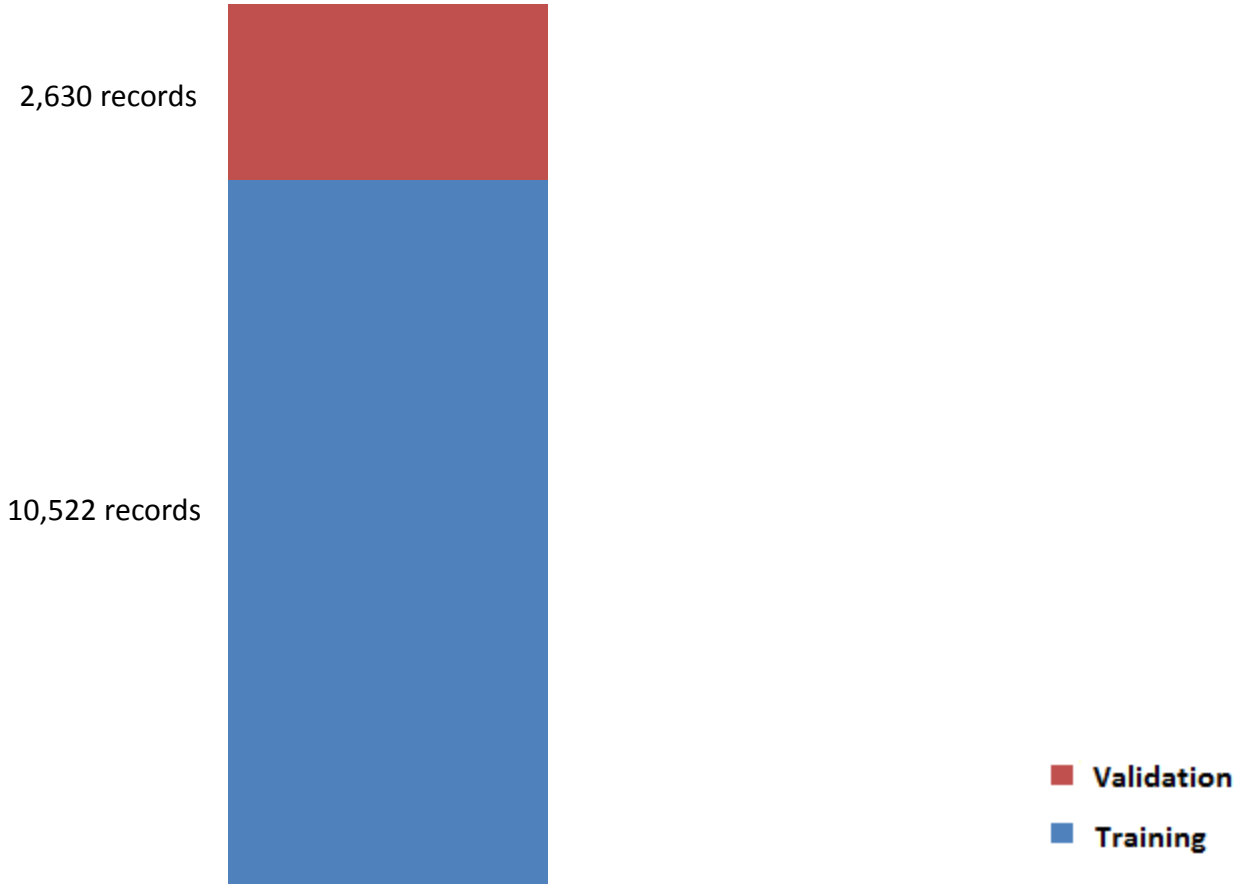
- Using the fall 03, 04, 05, 06 FTF cohorts for training

Predictor Variables
Gender
Under-Represented Status
Residence (county)
Parents Education
HS GPA
of College Prep Math Courses Passed in HS
of College Prep Science Courses Passed in HS
of College Prep Social Science Courses Passed in HS
of College Prep Art Courses passed in HS
SAT Math
SAT Verb
Prior Institution Type
Admission Basis Code
Pell Grant Recepient
Freshman Program Participation
College (Entry)
Entry Level Math Proficiency
English Proficiency
Degree-Applicable Units Earned in First Semester
F,D or WU Grade in 1st Semester
First Term GPA
Math Course (1st term)
English Course (1st term)

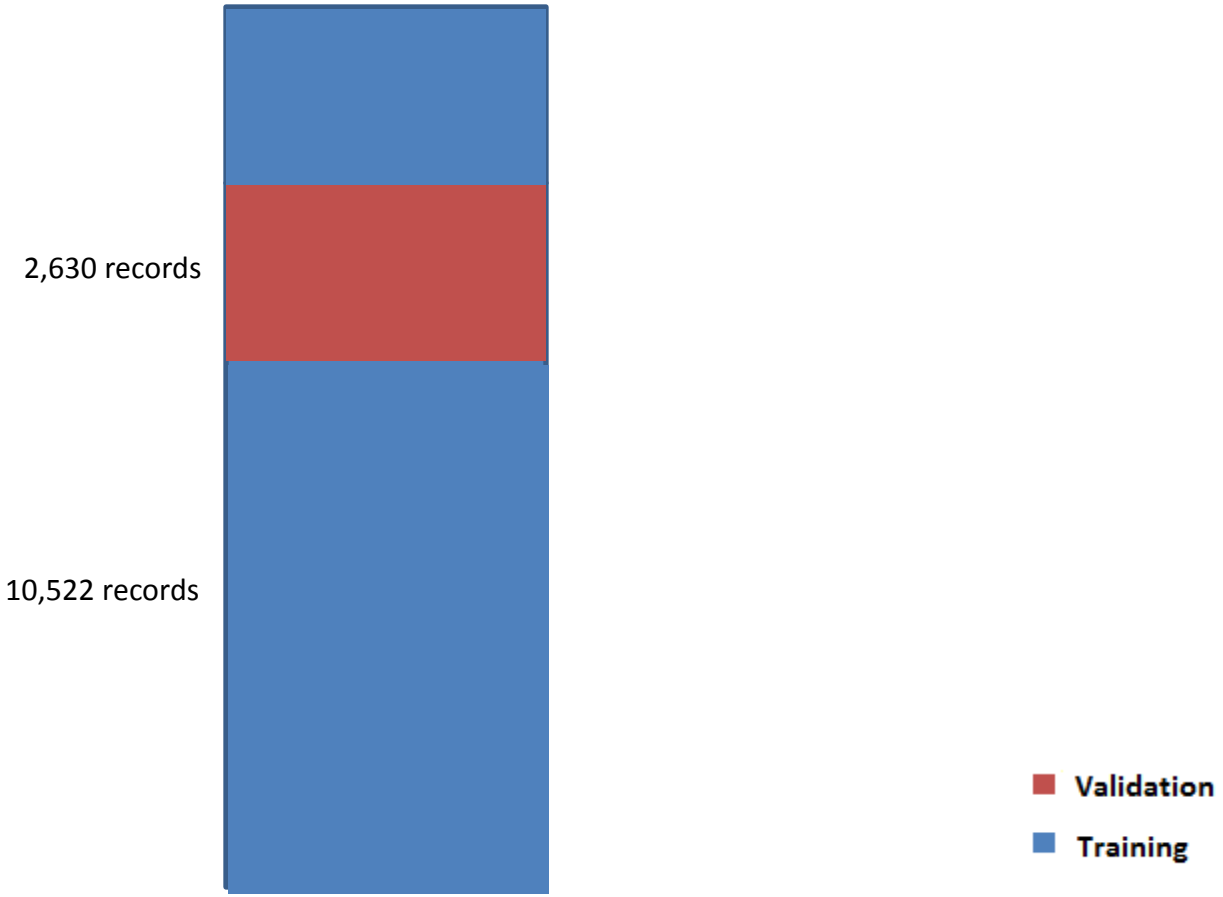
Model Validation & Testing

- Total of 14,152 records from fall 03, 04, 05, 06 cohorts (missing HS GPAs, SATs excluded) for model training
- Random 1,000 records removed and set aside for future testing
- Remaining 13,152 records used for training/validation using a 5-fold cross validation

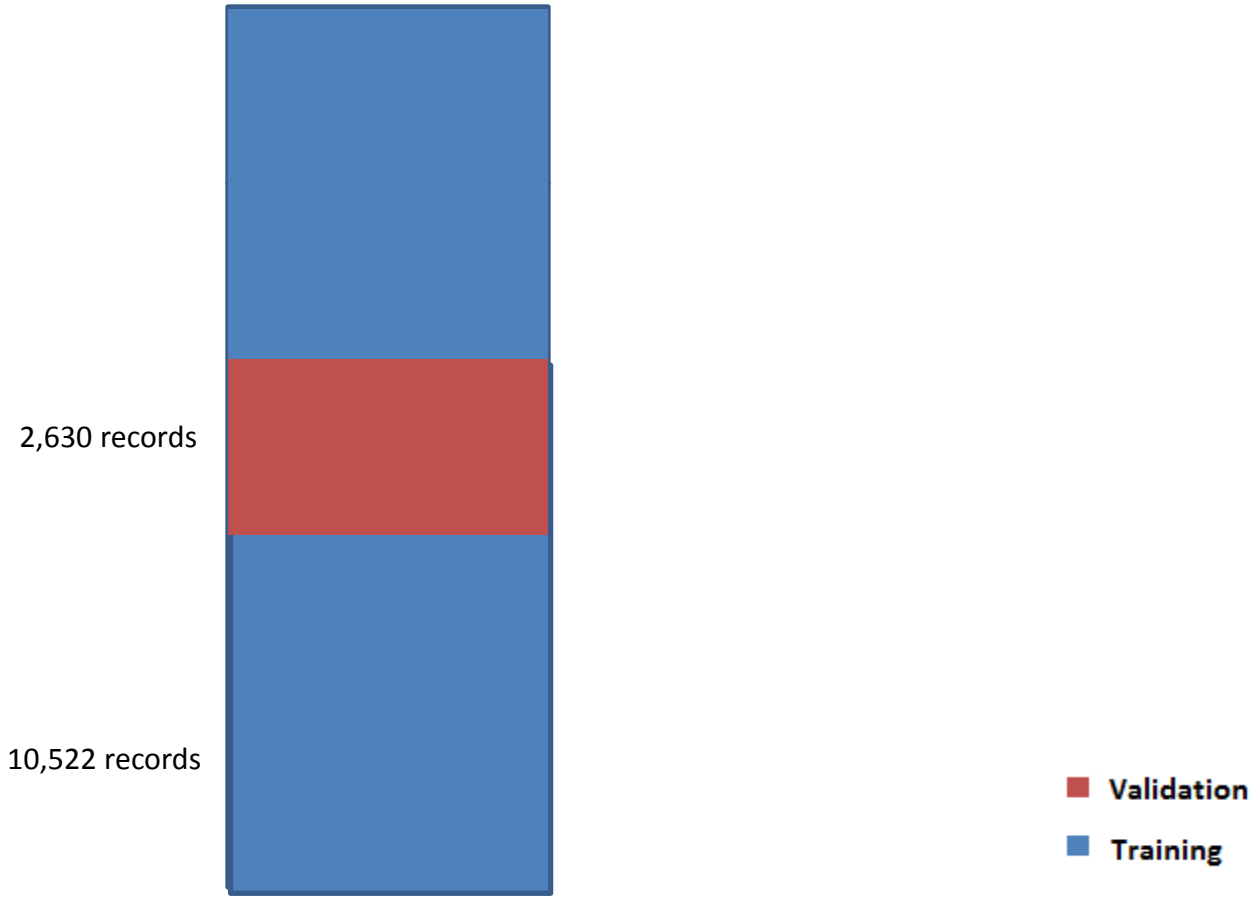
5-Fold Cross Validation



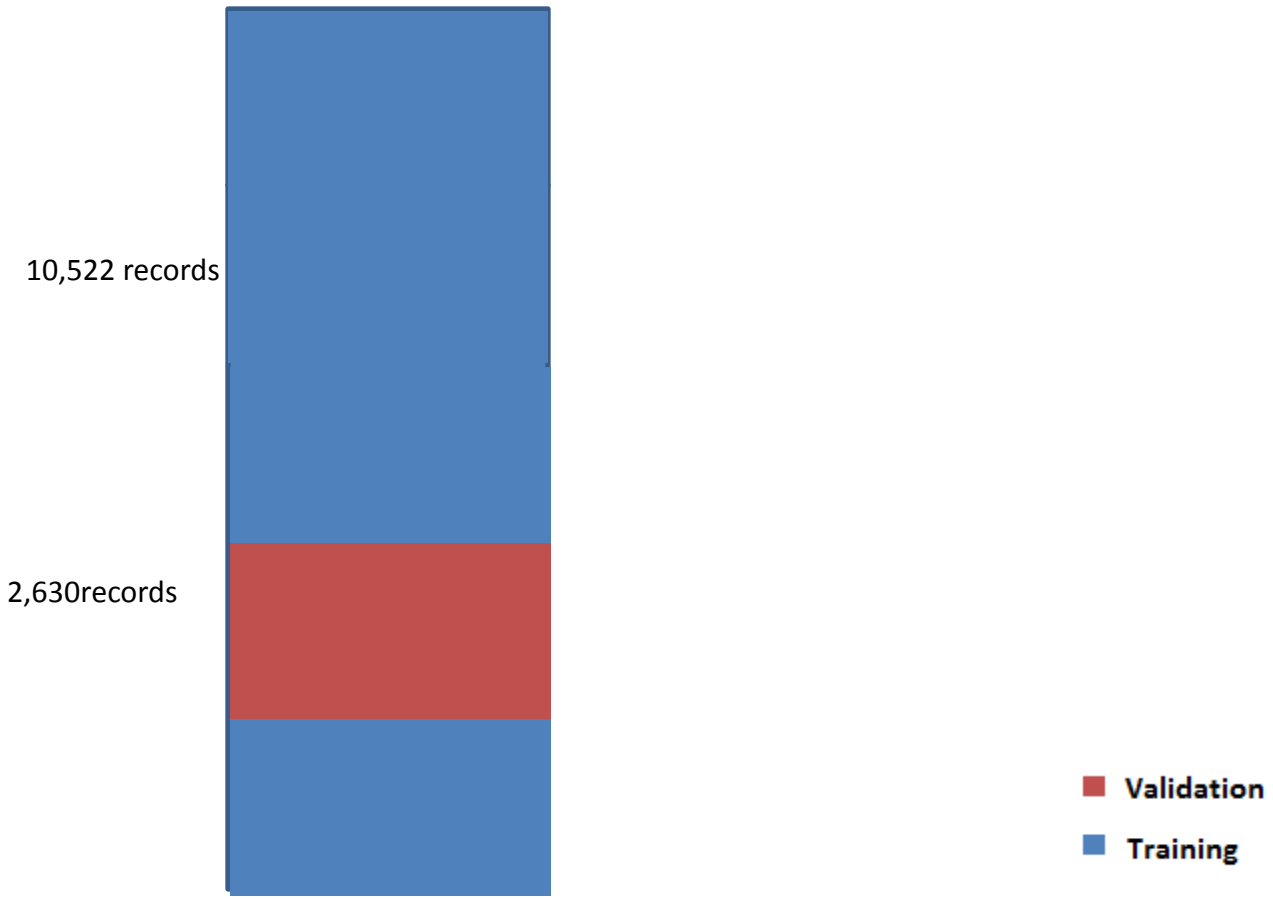
5-Fold Cross Validation



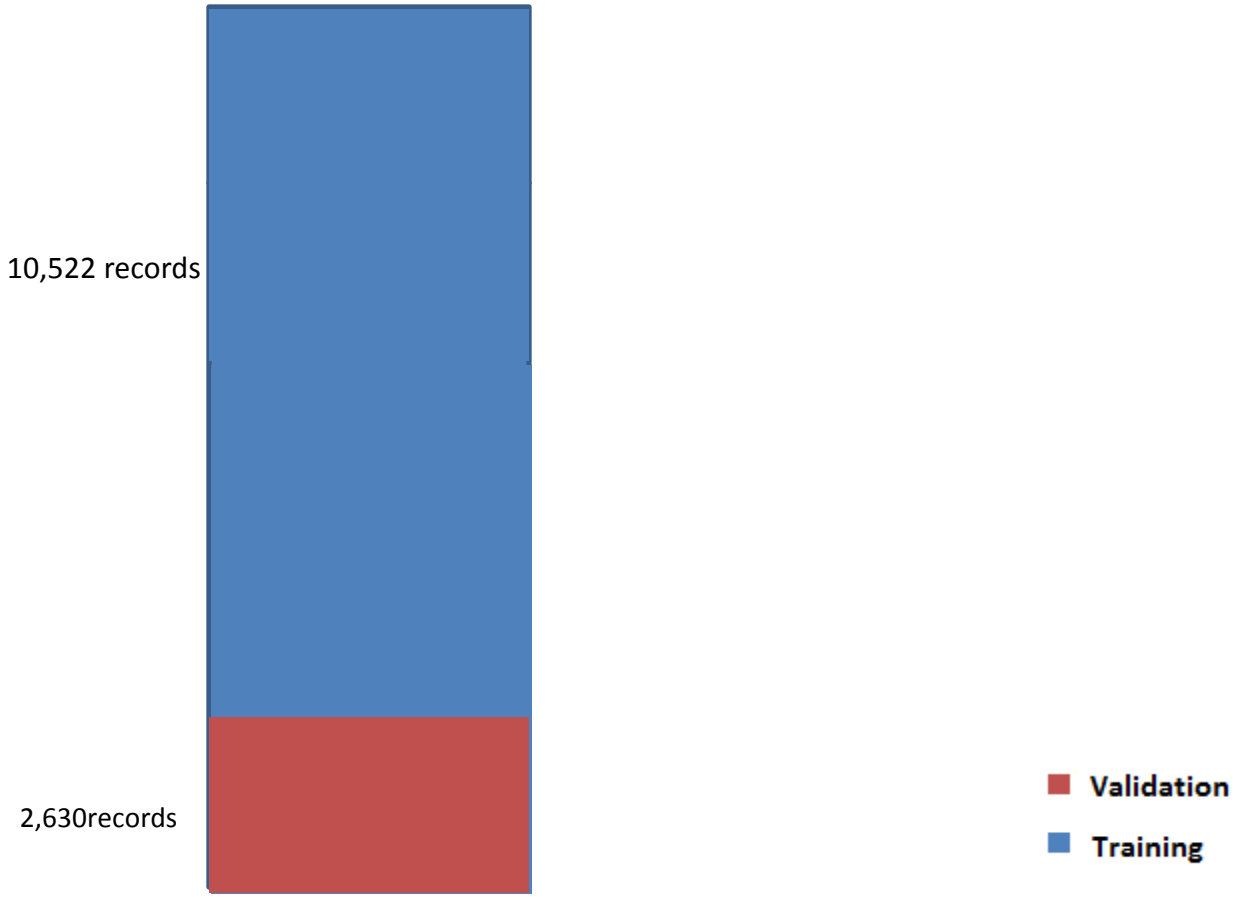
5-Fold Cross Validation



5-Fold Cross Validation



5-Fold Cross Validation



Model's Accuracy

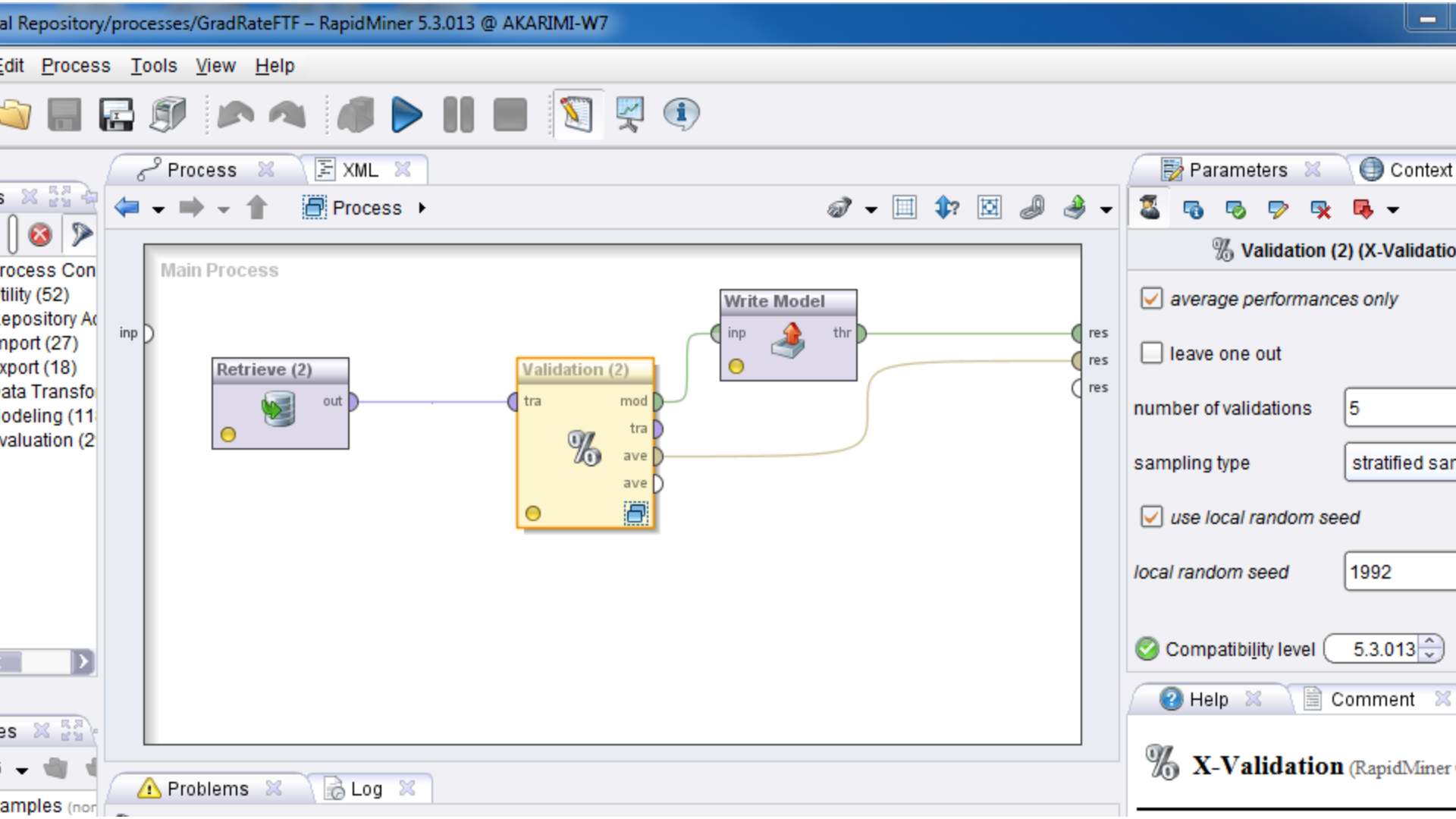
- Classification accuracy is the average accuracy of the 5 runs:
  - Classification Accuracy: 66.4%
  - Sensitivity (true positive rate): 72.4%
  - Specificity (true negative rate): 60.3%

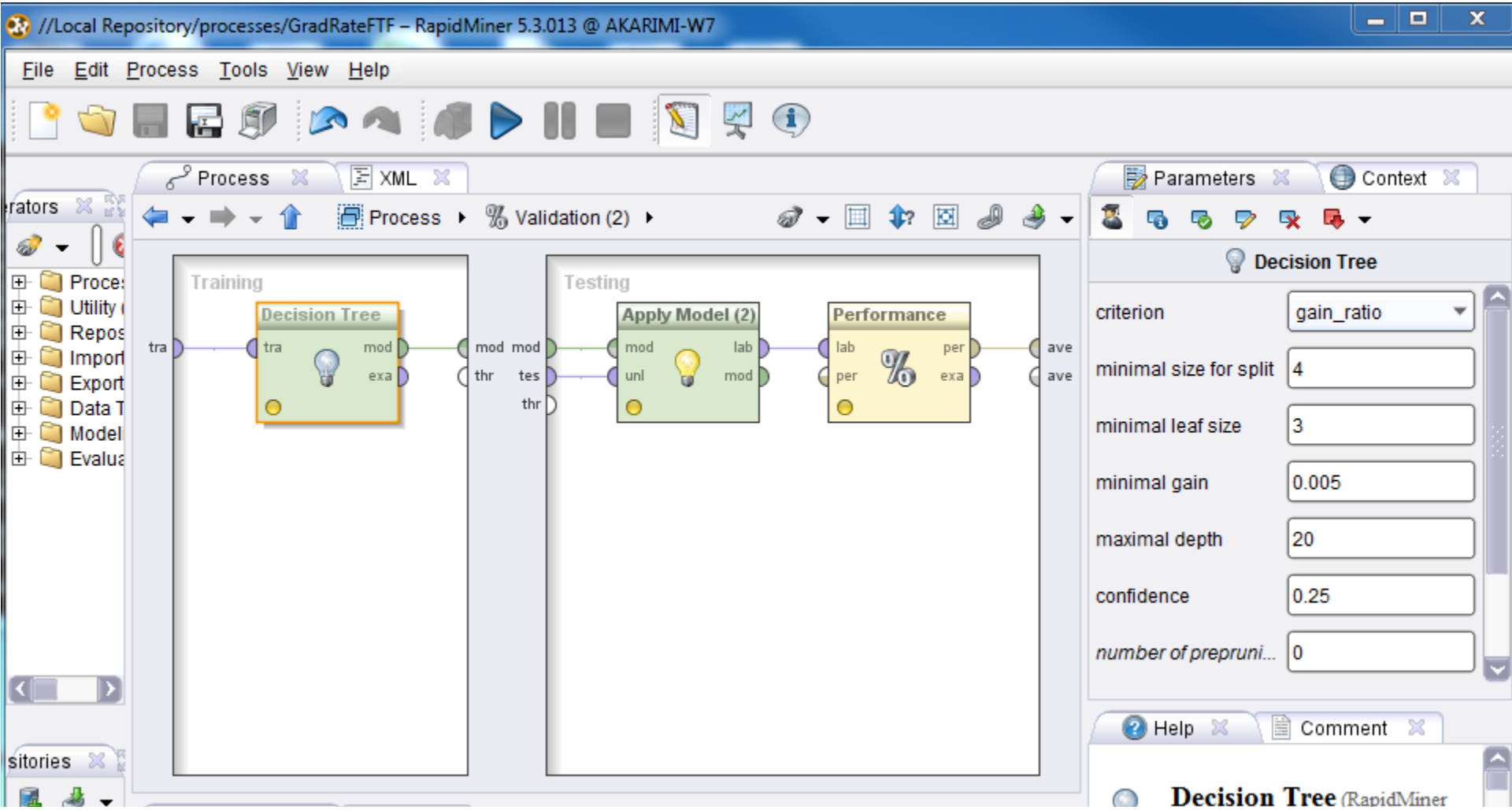
Table View Plot View

classification_error: 33.61% +/- 0.56% (mikro: 33.61%)

	true 0	true 1	class precision
pred. 0	3949	1819	68.46%
pred. 1	2601	4783	64.78%
class recall	60.29%	72.45%	

RapidMiner 5.0




//Local Repository/processes/GradRateFTF – RapidMiner 5.3.013 @ AKARIMI-W7
File Edit Process Tools View Help
Process
XML
Process
Validation (2)
Training
Decision Tree
tra
mod
exa
Testing
Apply Model (2)
mod
lab
unl
Performance
lab
per
exa
ave
thr
tes
Parameters
Context
Decision Tree
criterion
gain_ratio
minimal size for split
4
minimal leaf size
3
minimal gain
0.005
maximal depth
20
confidence
0.25
number of prepruni...
0
Help
Comment
Decision Tree (RapidMiner

Relevance (weights) of the variables on the Information Gain Ratio

Variable	Weight (normalized)
F,D or WU Grade in 1st Semester	0.075
Degree-Applicable Units Earned in First Semester	0.042
First Term GPA	0.036
Math Course (1st term)	0.033
Admission Basis Code	0.015
HS GPA	0.01
Gender	0.009
Freshman Program Participation	0.008
Entry Level Math Proficiency	0.007
English Course (1st term)	0.007
Under-represented Status	0.007
# of College Prep Math Courses Passed in HS	0.004
English Proficiency	0.004
College (entry)	0.004
Parents Education	0.003
SAT Verbal	0.003
Pell Grant Recepient	0.002
SAT Math	0.002
Prior Institution Type	0.002
Residence (county)	0.001
# of College Prep Social Science Courses Passed in HS	0.001
# of College Prep Science Courses Passed in HS	0.001
# of College Prep Art Courses Passed in HS	0.001

Generated Tree…

Testing

- Tested the model using the 1,000 records that were *NOT* used in building the model.

- Also, later (when summer 13 degrees were posted) tested the model using the Fall 07 cohort

Testing with Fall 07 FTF Cohort (Sept 13)

- Model predicts 1,717 (out of 4,026) students not to graduate in 6 years

- Model's classification accuracy: 68%

  (1183+1567)/4026

  sensitivity: 1567/2101 = 75%

  specificity: 1183/1925 = 61%

Count

		Graduated (actual)		
		0	1	Total
Graduated (predicted)	0	1183	534	1717
	1	742	1567	2309
Total		1925	2101	4026

- Top half of predicted non-graduates predicted with 82% accuracy

Count

		Graduated (actual)		
		0	1	Total
Graduated (predicted)	0	702	157	859
Total		702	157	859

Clustering

- Place these 859 students who were predicted *not* to graduate in clusters such that:
  - Students in each cluster are as similar as possible (based on their HS and 1st term college academic performances) and
  - Clusters are as different from each other as possible (again based on students' HS and 1st-term college academic performances)

K-Means Clustering-Using Mixed Euclidean Distance (both numeric and nominal variables)

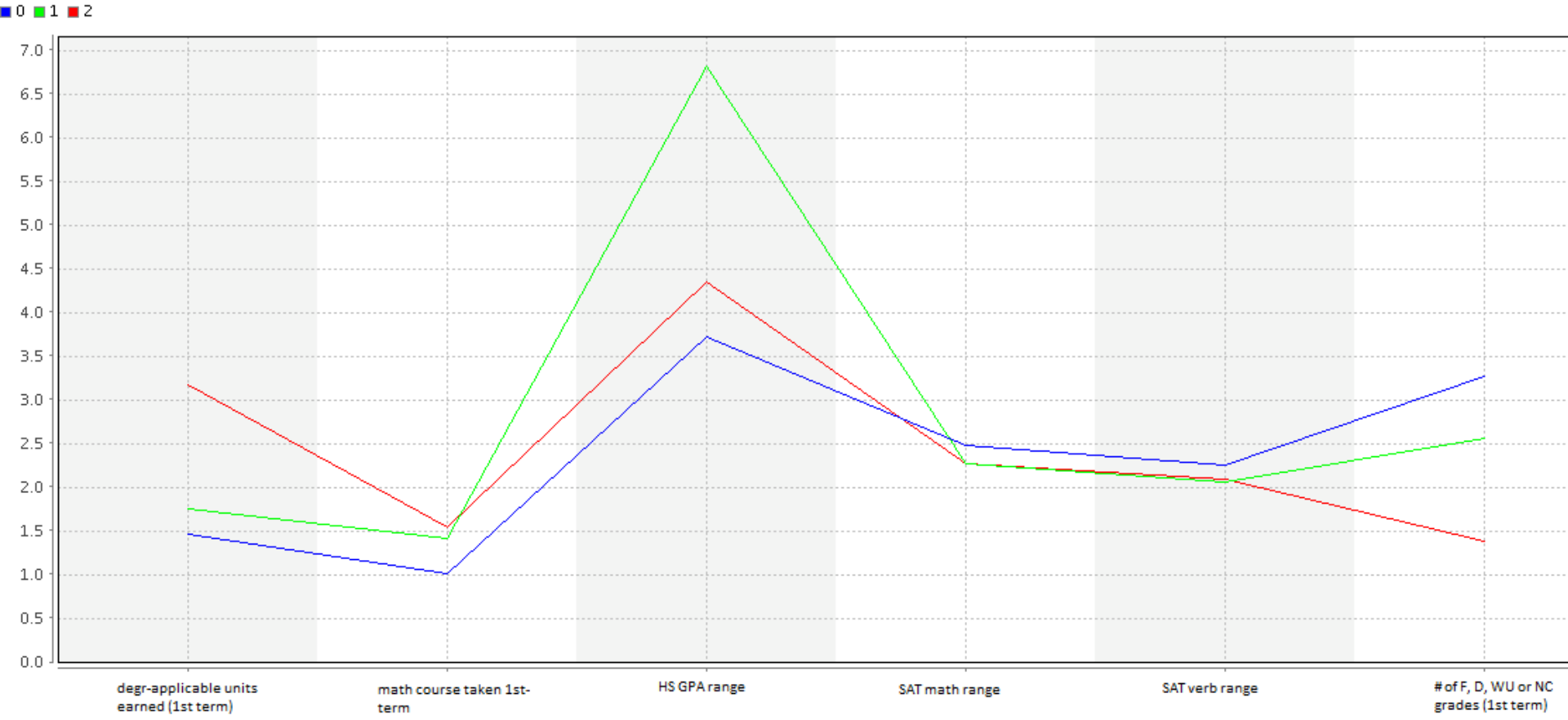
- Focus is on the HS to college transition
- Variables used (only academic performance pre-college and 1st term):

- HS GPA
- SAT Verb
- SAT Math
- Number of degree-applicable units earned in 1st term
- Number of F, D, WU or NC grades in 1st term
- 1st term type of math course passed/failed

Cluster Model

```
Cluster 0: 324 items
Cluster 1: 208 items
Cluster 2: 327 items
Total number of items: 859
```

Clusters' Centroid Plot



Clusters Analysis

Cluster	N	High School GPA		SAT Math		SAT Verb		Degree-applicable Units Earned		# of F, D, WU or NC grades	
		Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
0	324	2.84	0.22	493	88.2	469	83	1.57	1.95	3.27	1.01
1	208	3.45	0.23	472	87.6	451	77	2.41	2.41	2.57	1.11
2	327	2.96	0.23	471	81.6	453	75	6.35	3.06	1.39	0.59

Clusters Analysis Continued…

Cluster	1st Term Math Course Outcome				
	Failed Remedial Math	Failed GE Math	Passed Remedial Math	Passed GE Math	None
0	20%	57%	16%	6%	2%
1	15%	45%	29%	6%	5%
2	18%	30%	29%	20%	3%

- **Cluster 0 (The Un-motivated)**

- HS GPA 2.8
- SAT Math 493, SAT Verb 469
- 1st term college:
  - Earned 1.6 degree-applicable units
  - # of F, D, WU or NC grades: 3.3
  - 57% took & failed GE math, 20% took and failed remedial math
  - 1st term GPA: 0.58
- Mostly men (59% men, 41% women)
- College of major group mode: hierarchical, followed by semi-hierarchical
- **Benefits from (Probation) Advisement**

- **Cluster 2 (The Slow Starters)**

- HS GPA 2.9
- SAT Math 471, SAT Verb 453
- 1st term college:
  - Earned 6.3 degree-applicable units
  - # of F, D, WU or NC grades: 1.4
  - 30% took & failed GE math, 30% took and passed remedial math
  - 1st term GPA: 1.63
- Mostly women (47% men, 53% women)
- College of major group mode: semi-hierarchical, followed by non-hierarchical

- **Benefits from Academic Support**

Cluster 1 (The Disconnected)

- HS GPA: 3.4 (above avg. HS GPA of fall 07 incoming freshmen)
- SAT Math 472, SAT Verb 451
- 1st term college:
  - Earned 2.4 degree-applicable units
  - # of F, D, WU or NC grades: 2.6
  - 45% took & failed GE math, 29% took and passed remedial math
  - 1st term GPA: 0.83
- Largely 1st generation college students (40.4%)
- Majority underrepresented students (55.3%)
- Majority from outside local area high schools (57%)
- Mostly Women (36% men, 64% women)

Benefits from Practices that Promote Campus Engagement, Early Warning System

Summary

- Predictive model for early identification of at-risk students using early indicators (not past 1st term in college)…
- Provides insight into clusters of at-risk students; suggests cluster-level intervention
- Don't need expertise in machine learning, AI, statistics (data mining tools handle algorithms)
- Need to know the data intimately (data compilation & preparation most critical, most time-consuming)

Questions/Comments?

Contact email: akarimi@fullerton.edu