Data Mining

One Institution's Experience with Getting Started

Emily Berg
North Dakota State University
AIRUM 2012

Outline of Presentation

- Introduction to data mining
- Some data mining studies in IR
- Data mining outside of IR
- Using data mining to analyze a new proposed admission method

What is data mining?

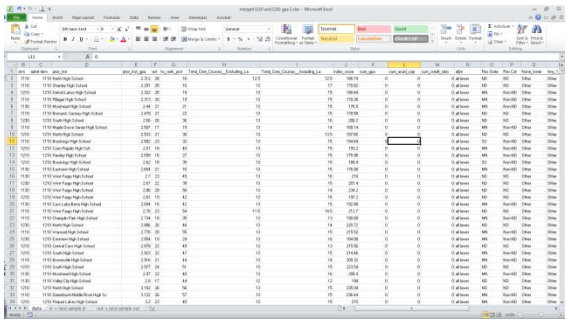
- Data mining is a method of uncovering hidden trends and patterns that lend themselves to predictive modeling using a combination of explicit knowledge base, sophisticated analytical skills, and academic domain knowledge. – *Jing Luan*
- Exploratory & Predictive
- Automatic or semi-automatic analysis of large datasets
- Unsupervised (no knowledge of outcome) or supervised (know outcome beforehand)

Is data mining a legitimate form of data analysis?

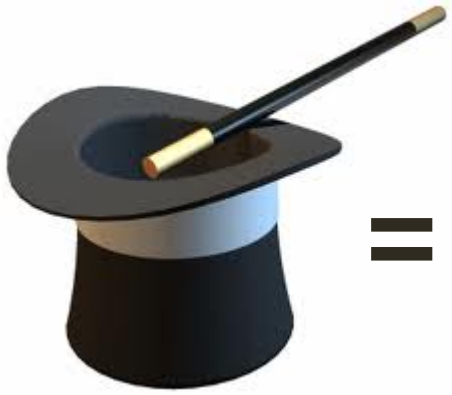
Yes

However… data fishing is not!

What data mining is NOT…



+



=

Everything you ever needed to know about our students… and then some.

By NDSU OIRA

1,437,761 pages

Advantages to using data mining

- Exploratory
- Predictive
- Takes advantage of massive stores of data
- No *a priori* hypotheses required

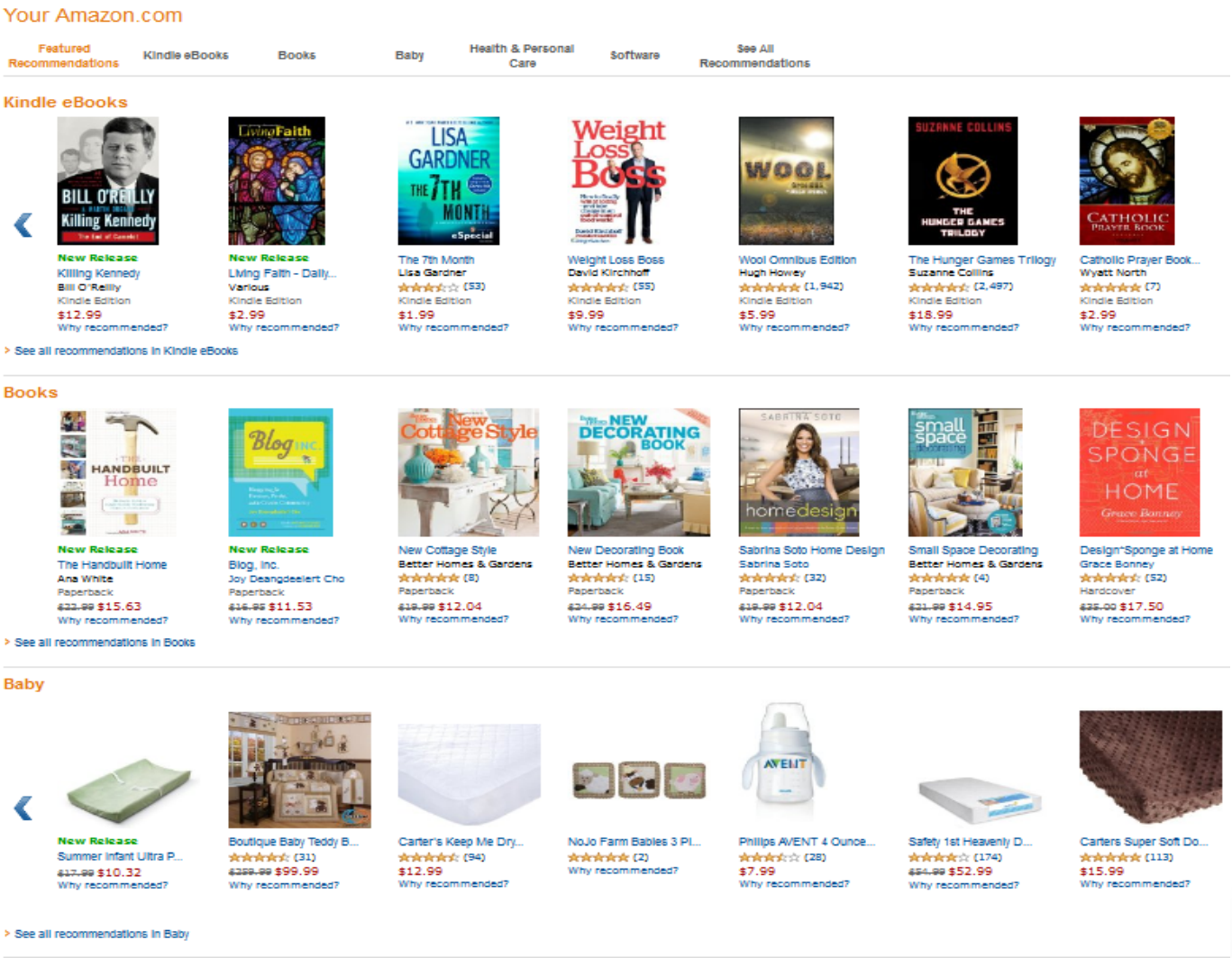
Data mining in IR

- Still relatively new in last 5-10 years
- Estimating retention and degree completion time (Herzog, 2006)
- Evaluating achievements of distance education students (Sen and Ucar, 2012)
- Recommending course enrollment, matching student ability to course difficulty (Vialardi et al., 2011)
- Institutional Learning Engagement Typology (Luan et al., 2009)

Data mining in other fields

- Mitt Romney campaign: Identified potential wealthy donors
- Credit card companies: Fraud detection
- Netflix movie recommendations
- Health insurance companies: Plans tailored based on risk factors

Amazon.com: Recommended for you…



Software

- IBM SPSS Modeler 14.2
  - Selected following the OIRA Director's attendance at IBM Business Analytics Conferences and Workshops in 2009 and 2010.
  - Used often in IR, so transition was easy
  - Support seemed good prior to purchase; So far, so good post-purchase.
  - Point & click user interface was compatible with the (lack of) SQL skills of the Assistant Director

- Other options:
  - SAS Enterprise Miner
  - Oracle Data Mining
  - STATISTICA

Training

- IBM Business Analytics Conferences and Workshops in 2009 & 2010
- AIR 2012 Workshop
  - Both theoretical and hands-on
  - Highly recommend if offered again
- IBM Support documents and demo streams
- Learning as we go for some things

Long-term data mining goals at NDSU

- Define “successful student”
  - What do they look like when they are admitted?
  - What courses do they take and in what sequence?
  - What other factors contribute to “success”?
    - How do you define “success”?

- Predict graduation (y/n) and time to graduation

Long-term goals: Data collection

- Worked with software engineer from IT department to extract as much data as possible from online data repositories
- Collected every demographic, admissions, and academic variable we could find that seemed relevant.
- Still hope to add:
  - Advisor meetings
  - Blackboard is an untapped resource (turning in homework, class attendance, online participation, etc.)

Sample Project:
Admissions Criteria Revision at North Dakota University System

- Admissions methods revision proposed by new Chancellor

GOALS

- Promote success by enrolling the student at the appropriate level
- Lower number of remedial courses being taught at research universities
- Address the potential population boom of college students in North Dakota

Admissions Index

- Model - Iowa's Regent Admission Index
- Uses HS GPA, ACT, and number of core high school courses to create index score (IA also uses class rank)
  - Highest group of index scores = automatic admission into research universities (NDSU and UND) and all other institutions
  - Next group = admission into non-research universities
  - Third group = admissions into junior colleges

Admissions Index

- Does it accurately reflect NDSU student retention and achievement?
- Will we be keeping out too many students that would do well?
- Will we be admitting too many students who will struggle?
- Can we create a better index on our own?

Admissions Index

Can we predict first year academic performance based on admissions data?

Admissions Data:

- High school GPA
- ACT Composite score
- Count of core high school courses

- Tuition Residency (North Dakota vs. Non-North Dakota)

Admissions Index: Data

Measures of Success

- Retention
  - Fall to Spring
  - Fall to Fall

- Three measures of Academic Success after first year
  - Academic Capacity **- Grade points/number of courses**
  - GPA **- Grade points/number of credits**
  - Credits Earned to Credits Attempted Ratio **– Credits earned/credits attempted**

Admissions Index: Study Population

- 3,614 First-Time, Full-Time students entering NDSU the fall after their spring high school graduation

Cohort	North Dakota	Non-North Dakota	Total
Fall 2010	685	1126	1811
Fall 2011	711	1092	1803
Total	1396	2218	3614

- No international, home school, GED, non-traditional, etc.

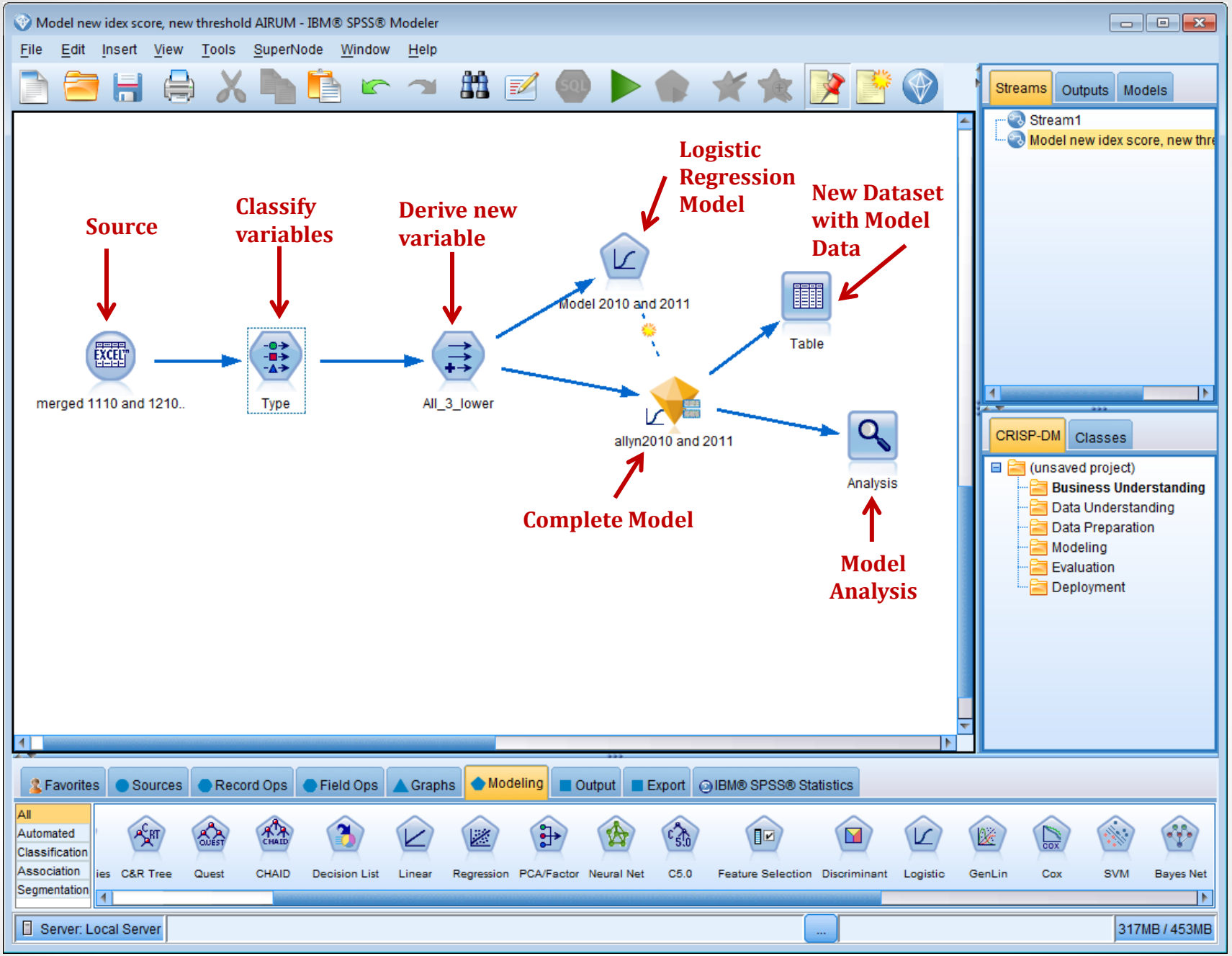
Predicting academic success for FYR students

- Started with AutoClassifier Node
  - Don’t know which test to start with? Automate it!
  - Runs all models types with classification outcomes

- Selected Logistic Regression because it was ranked highest for overall accuracy

Stage 1: 3 Low Measures

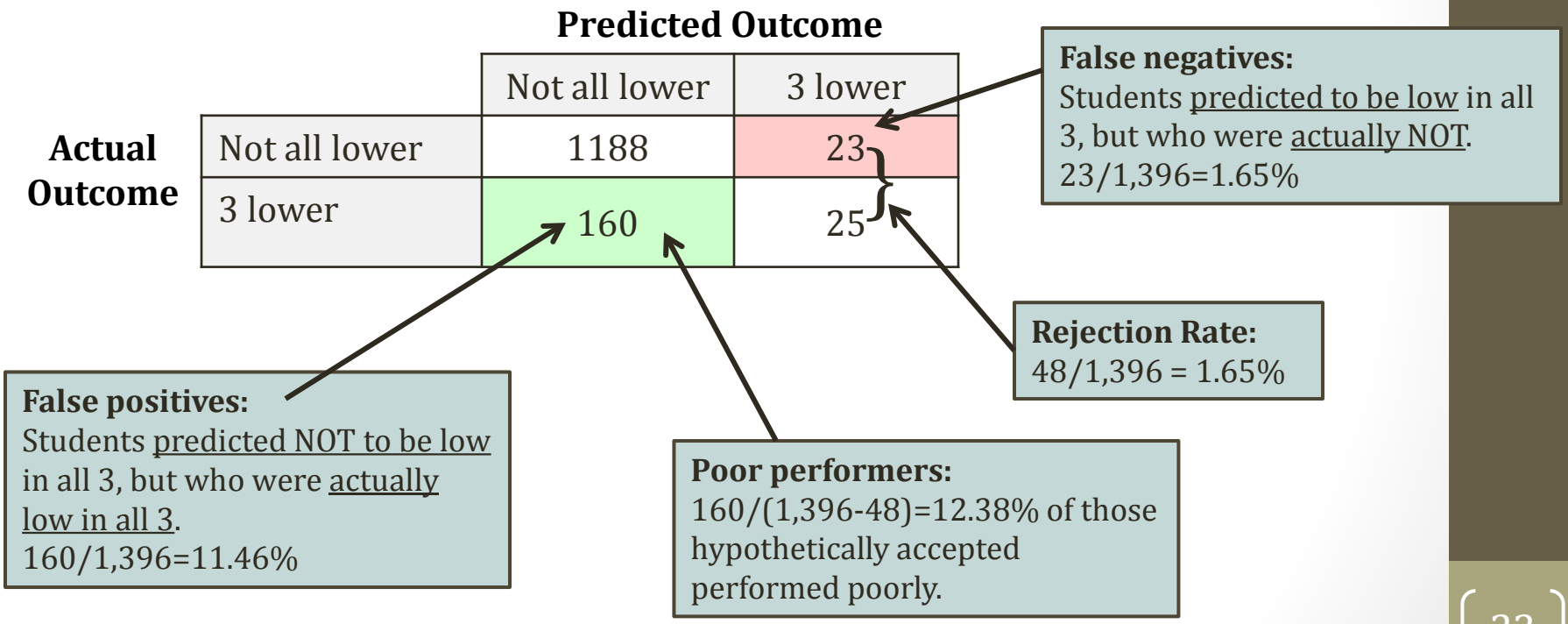
- All 3 Measures below threshold vs. All 3 NOT below
  - GPA < 2.00
  - Academic Capacity < 3.75
  - Credits Earned to Credits Attempted Ratio <0.75


Model new idex score, new threshold AIRUM - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Window Help
Source
Classify variables
Derive new variable
Logistic Regression Model
New Dataset with Model Data
merged 1110 and 1210..
Type
All_3_lower
Model 2010 and 2011
Table
allyn2010 and 2011
Analysis
Complete Model
Model Analysis
Streams Outputs Models
Stream1
Model new idex score, new thre
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics
All
Automated
Classification
Association
Segmentation
C&R Tree Quest CHAID Decision List Linear Regression PCA/Factor Neural Net C5.0 Feature Selection Discriminant Logistic GenLin Cox SVM Bayes Net
Server: Local Server
317MB / 453MB

Model 1: 1,396 North Dakota students

Score = 2.854*GPA - 0.08575*ACT - 0.03446*Courses - 5.021

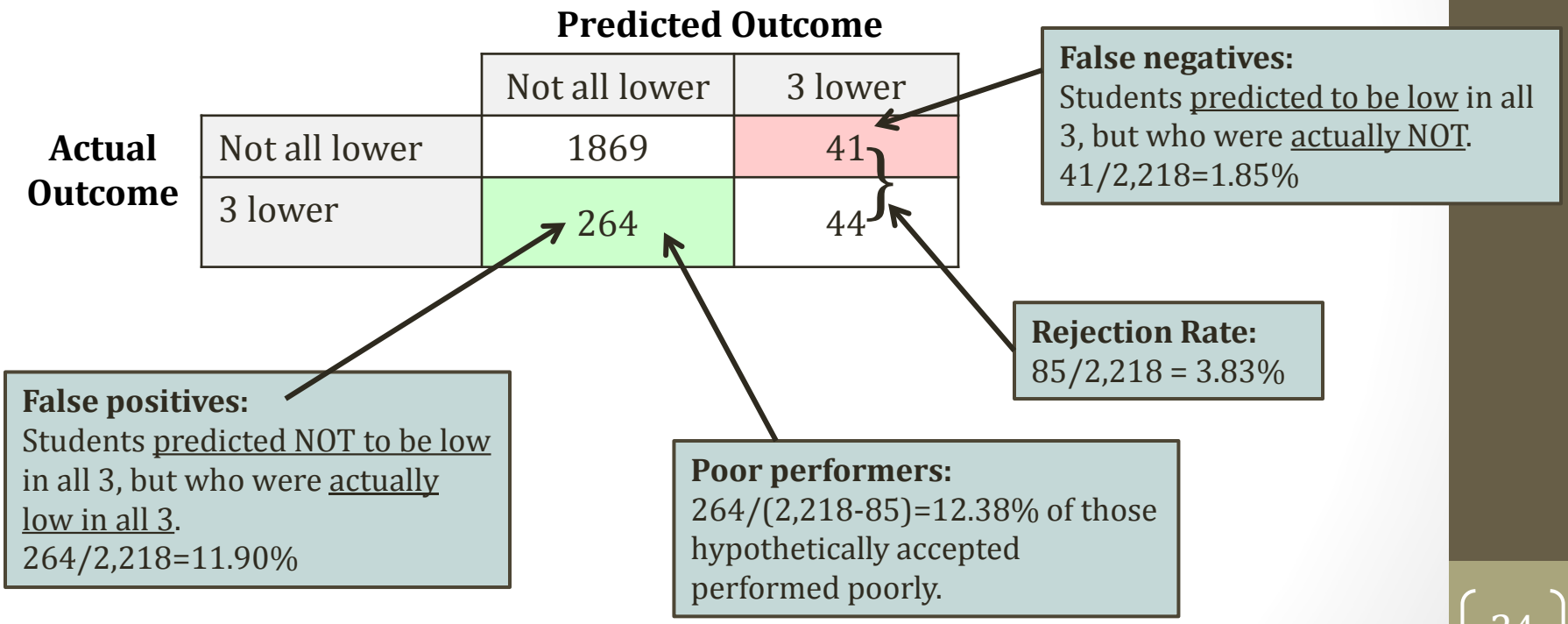
Predictive Accuracy = 86.89% (1,213 of 1,396 outcomes)



Model 2: 2,218 Non-North Dakota students

Score = 2.705*GPA - 0.02054*ACT + 1.19*Courses – 21.74

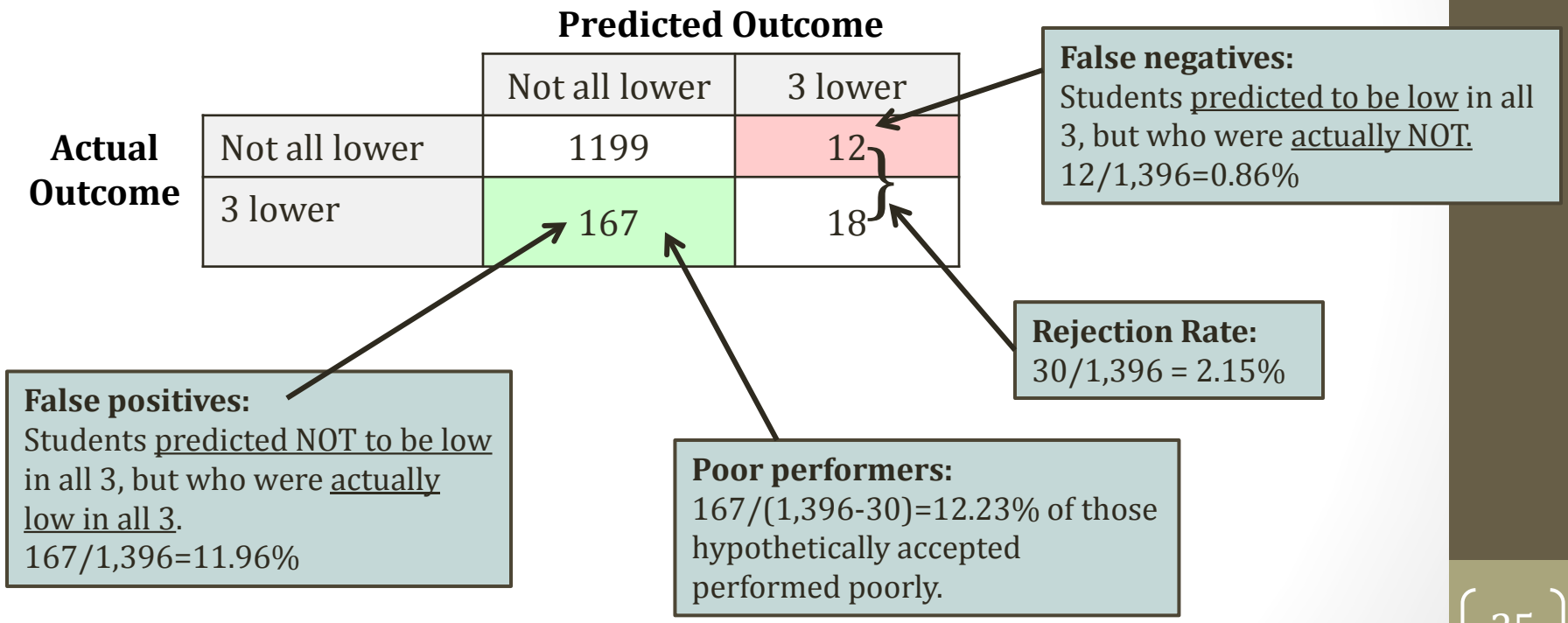
Predictive Accuracy = 86.25% (1,913 of 2,218 outcomes)



Model 3: 1,396 North Dakota students + 0.25

Score = 2.854*GPA - 0.08575*ACT - 0.03446*Courses - 5.021 +0.25

Predictive Accuracy = 87.18% (1,217 of 1,396 outcomes)



Stage 2: 1 or 2 Low Measures

- What about students with one or two measures low?

Probability range	Prediction	Admitted	Rejected	Action
0-0.39	Not all lower	100%	0%	Automatic
0.4-0.449	Not all lower	67%	33%	1 reviewer
0.45-0.499	Not all lower	33%	67%	2 reviewers
0.5-1.0	All lower	0%	100%	Automatic

- Probability means probability of being in the “All lower” category (calculated by Modeler)

2-Stage Admission Proposal Results: North Dakota students

1,321 students admitted and predicted to not have all 3 lower

1,173 (88.80%) of 1,321 students admitted performed well

Admission Status

	Admitted		Denied	
Outcome	Predicted	Actual	Predicted	Actual
Not all lower	1,321	1,173	27	38
3 lower	0	148	48	37
Total	1,321	1,321	75	75

None of the students predicted to have all 3 lower were admitted

148 (11.20%) of 1,321 students admitted actually performed poorly

2-Stage Admission Proposal Results: North Dakota students

27 (36.00%) of 75 students denied were predicted not to have all 3 lower

38 (50.67%) of 75 students denied actually would have had 1 or 2 lower, but not all 3

Admission Status

	Admitted		Denied	
Outcome	Predicted	Actual	Predicted	Actual
Not all lower	1,321	1,173	27	38
3 lower	0	148	48	37
Total	1,321	1,321	75	75

48 (64.00%) of 75 students denied were predicted to have all 3 lower

37 (49.33%) of 75 students denied actually would have had all 3 lower

2-Stage Admission Proposal Results: Non-North Dakota students

2,092 students admitted and predicted to not have all 3 lower

1,845 (88.19%) of 2,092 students admitted performed well

Admission Status

	Admitted		Denied	
Outcome	Predicted	Actual	Predicted	Actual
Not all lower	2,092	1,845	41	65
3 lower	0	247	85	61
Total	2,092	2,092	126	126

None of the students predicted to have all 3 lower were admitted

247 (11.81%) of 2,092 students admitted actually performed poorly

2-Stage Admission Proposal Results: Non-North Dakota students

41 (32.54%) of 126 students denied were predicted not to have all 3 lower

65 (51.59%) of 126 students denied actually would have had 1 or 2 lower, but not all 3

Admission Status

	Admitted		Denied	
Outcome	Predicted	Actual	Predicted	Actual
Not all lower	2,092	1,845	41	65
3 lower	0	247	85	61
Total	2,092	2,092	126	126

85 (67.46%) of 126 students denied were predicted to have all 3 lower

61 (48.41%) of 126 students denied actually would have had all 3 lower

Grades of Admission Index Students

Communications 101

Group	Percent with D or F
Whole Class	7%
Predicted low in all 3	40%
Actually low in all 3	54%
Predicted and actually low in all 3	71%

Grades of Admission Index Students

Math 101 (Remedial)

Group	Percent with D or F
Whole Class	38%
Predicted low in all 3	85%
Actually low in all 3	90%
Predicted and actually low in all 3	100%

Grades of Admission Index Students

English 110

Group	Percent with D or F
Whole Class	15%
Predicted low in all 3	35%
Actually low in all 3	46%
Predicted and actually low in all 3	64%

Grades of Admission Index Students

English 120

Group	Percent with D or F
Whole Class	10%
Predicted low in all 3	38%
Actually low in all 3	56%
Predicted and actually low in all 3	70%

Admissions Index Revision: Main Points

We would sacrifice some people, all of whom are less likely to do well, to increase the overall success of the population.

Students who are predicted not to do well, and actually don't do well, have a higher incidence of low grades in basic general education courses.

Data Mining: Main Points

Easy to use interface is encouraging for a beginning data miner.

Increased use in IR provides more opportunities for support and for sharing of projects among researchers.

Predictive capabilities can help us meet the goals set forth by today's administrations.

"What will happen in the future?" vs. "What has happened in the past?"