




## Jumping into Data Mining: The Triumphs and Tragedies of Our Second Year

---

Douglas K. Anderson, Director, *Enrollment Planning & Research*  
Bridgett J. Milner, Assistant Director, *Enrollment Planning & Research*  
Michael J. Sauer, Senior Research Analyst, *Office of the Registrar*  
Linda Shepard, Senior Associate Registrar, *Office of the Registrar*


Indiana University – Bloomington

[www.indiana.edu/~oem/pages/presentations.php](http://www.indiana.edu/~oem/pages/presentations.php)



## Data Mining – definitions

- Luan, Jing and Zhao, Chun-Mei (2006) *Data Mining in Action: Case studies of enrollment management. New Directions for Institutional Research. Josey-Bass*
  - Data rich and information poor
  - Data fishing- data exploration
  - Discovery of hidden patterns
  - If you torture data long enough it will confess to anything
  - An iterative process of finding trends and patterns in data (Groth, 2000) - Scientific method?
  - Manages the entire data process as a data stream (Bailey, 2006)



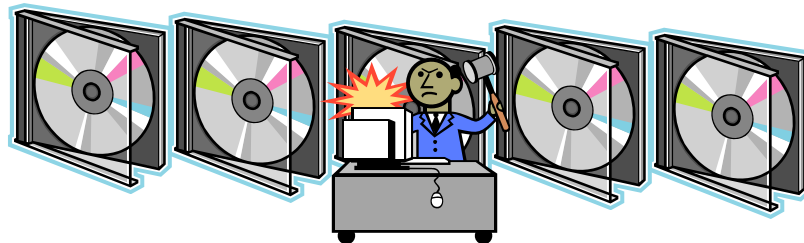
## Data Mining – Interest

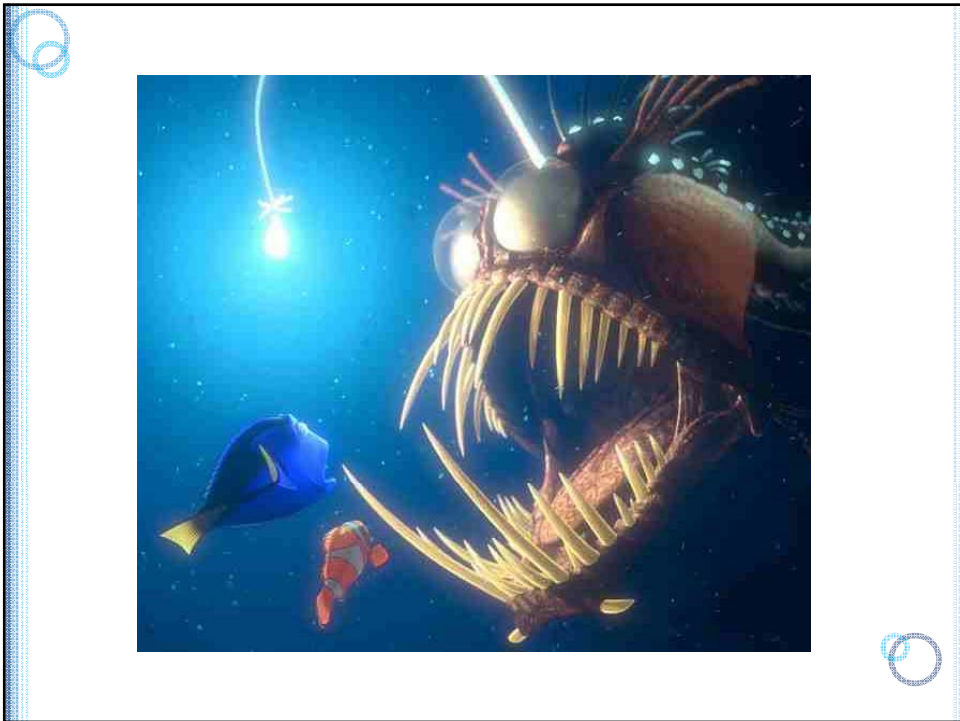
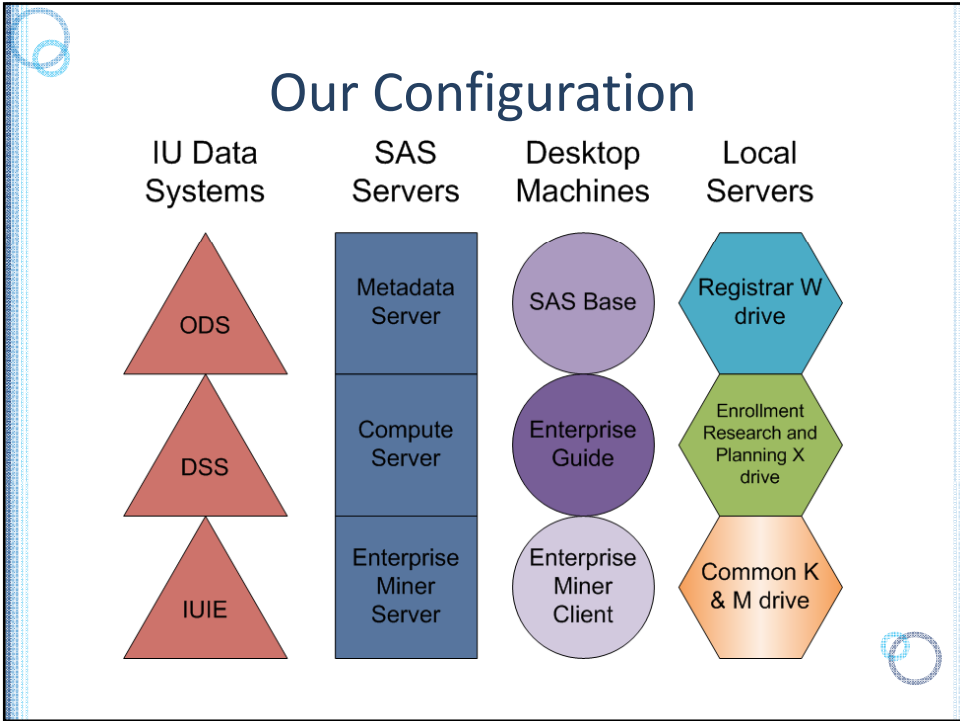


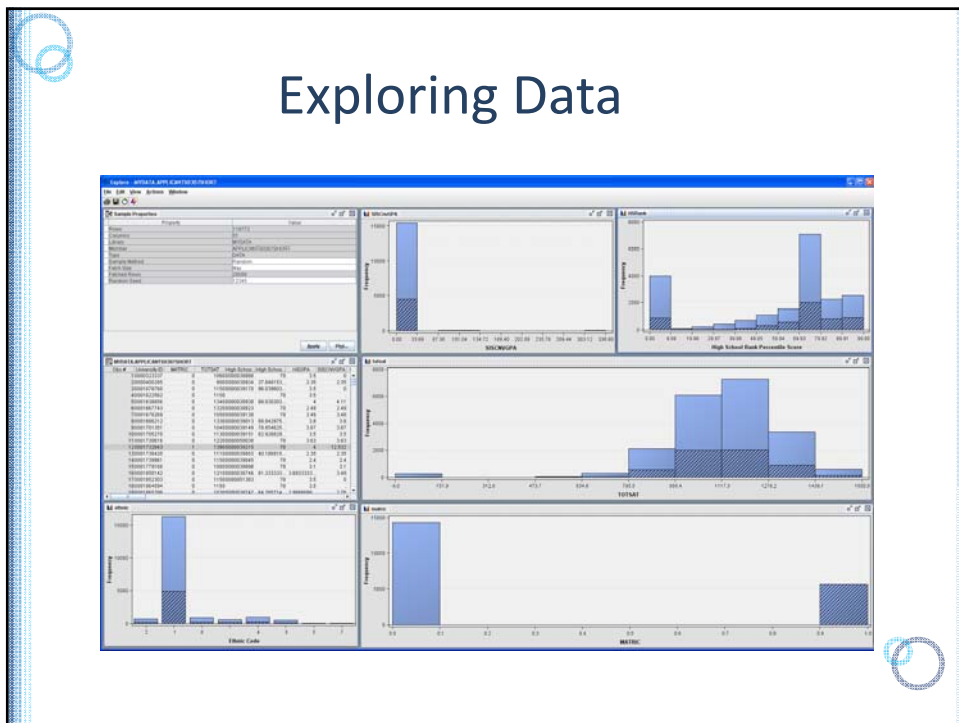
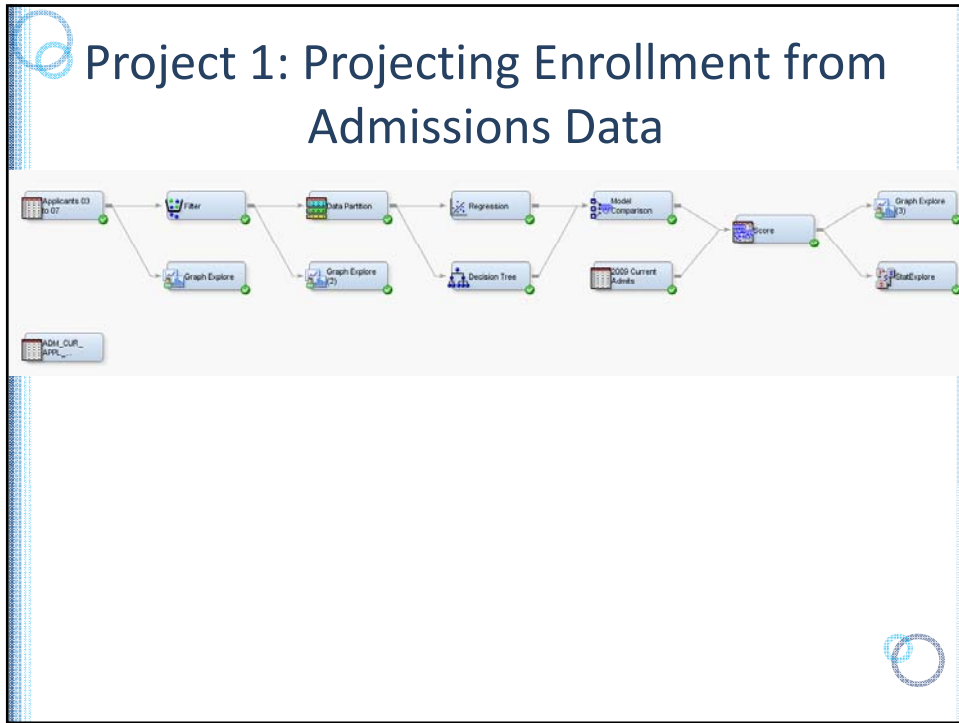
- Customer Relations Management system (CRM)
- Targeted marketing
- Enrollment projection models
- Course management systems
- Space utilization
- Retention/Attrition – operational programs

## Technical Implementation Issues

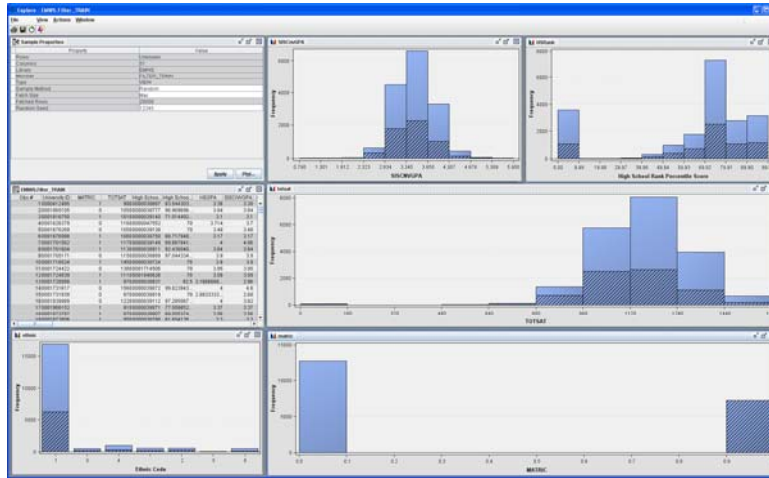
- Diverse Technical Expertise Required
  - Server -Database -Statistical -Desktop
  - Programming
  - SAS tech support connection
  - Translator to facilitate communication







# Clean Data



# Regression Results

Output

```

340 totsat      1  -0.00031  0.000022  -14.22  <.0001
341
342
343 NOTE: No (additional) effects met the 0.05 significance level for entry into the model.
344
345
346          Summary of Stepwise Selection
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361 The selected model, based on the CHOOSE-CRITERION, is the model trained in Step 4. It consists of the following effect
362
363 Intercept appao female reside totsat
364
365
366          Analysis of Variance
367
368          Sum of

```

| Step | Entered       | DF | Number | In | F Value | Pr > F |
|------|---------------|----|--------|----|---------|--------|
| 1    | reside        | 1  | 1      | 1  | 2996.72 | <.0001 |
| 2    | totsat        | 1  | 2      | 2  | 254.24  | <.0001 |
| 3    | female        | 1  | 3      | 3  | 120.70  | <.0001 |
| 4    | appao         | 1  | 4      | 4  | 90.61   | <.0001 |
| 5    | SISKWAPA      | 1  | 5      | 5  | 24.99   | <.0001 |
| 6    | White         | 1  | 6      | 6  | 16.00   | <.0001 |
| 7    | international | 1  | 7      | 7  | 14.38   | 0.0001 |
| 8    | EthnicNA      | 1  | 8      | 8  | 6.87    | 0.0088 |

# Regression Results

363  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394

| Source          | DF    | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|-------|----------------|-------------|---------|--------|
| Model           | 4     | 717.681187     | 179.420297  | 878.50  | <.0001 |
| Error           | 28176 | 5754.546484    | 0.204236    |         |        |
| Corrected Total | 28180 | 6472.227671    |             |         |        |

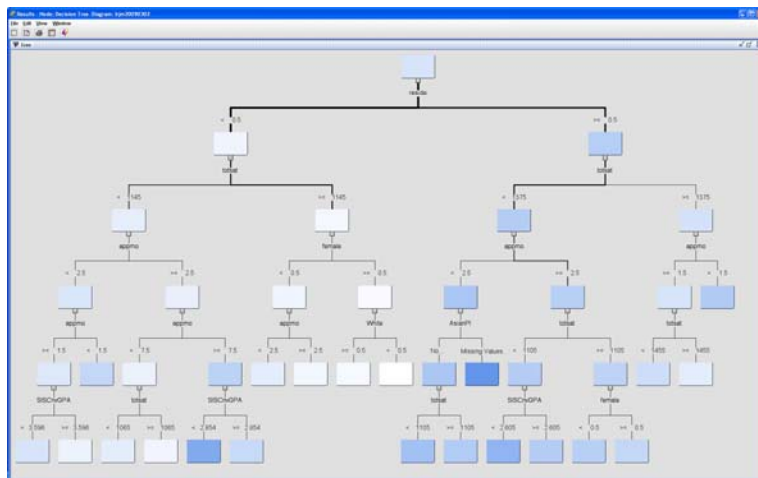
  

|          |             |          |             |
|----------|-------------|----------|-------------|
| R-Square | 0.1109      | Adj R-Sq | 0.1108      |
| AIC      | -44759.9642 | BIC      | -44757.9057 |
| SBC      | -44718.7322 | C(p)     | 70.4709     |

| Parameter | DF | Estimate | Standard Error | t Value | Pr >  t |
|-----------|----|----------|----------------|---------|---------|
| Intercept | 1  | 0.7530   | 0.0255         | 29.52   | <.0001  |
| ageao     | 1  | -0.0163  | 0.00171        | -9.52   | <.0001  |
| female    | 1  | -0.0646  | 0.00551        | -11.73  | <.0001  |
| reside    | 1  | 0.2772   | 0.00559        | 49.70   | <.0001  |
| totstat   | 1  | -0.00036 | 0.000019       | -18.52  | <.0001  |

# Decision Tree Results



# Decision Tree Results

EMWS.Tree\_BROWSETREE[EMWS.PART\_TRAIN] - SAS Enterprise Miner Tree Desktop Application

File Edit Model Train View Options Window Help

| Variable                          | Nodes | Training | Validation | Importance |
|-----------------------------------|-------|----------|------------|------------|
| reside                            | 1     | 1.000    | 1.000      |            |
| totsat                            | 6     | 0.314    | 0.310      |            |
| appmo                             | 6     | 0.264    | 0.241      |            |
| female                            | 2     | 0.123    | 0.126      |            |
| SISCnvGPA                         | 3     | 0.123    | 0.044      |            |
| AsianPI                           | 1     | 0.118    | 0.116      |            |
| White                             | 1     | 0.049    | 0.035      |            |
| Black                             | 0     | 0.000    | 0.000      |            |
| international                     | 0     | 0.000    | 0.000      |            |
| AmerInd                           | 0     | 0.000    | 0.000      |            |
| Hispanic                          | 0     | 0.000    | 0.000      |            |
| High School Rank Percentile Score | 0     | 0.000    | 0.000      |            |

# Model Comparison

| Selected Model | PARENT | MODEL | MODEL DESCRIPTION   | TARGET | Test Average Squared Error | Train: Sum of Frequencies | Train: Sum of Case Weights Times Freq | Train: Maximum Absolute Error | Train: Sum of Squared Errors | Train: Average Squared Error | Train: Root Average Squared Error | Train: Divisor ASE |
|----------------|--------|-------|---------------------|--------|----------------------------|---------------------------|---------------------------------------|-------------------------------|------------------------------|------------------------------|-----------------------------------|--------------------|
| Y              | Tree   | Tree  | Decision Tr. matrix | matrix | 0.203655                   | 36048                     | 36048                                 | 0.861538                      | 7313.468                     | 0.202854                     | 0.450393                          |                    |
|                | Reg    | Reg   | Regression matrix   | matrix | 0.209398                   | 36048                     | 36048                                 | 2.285731                      | 7692.213                     | 0.210337                     | 0.458625                          |                    |

Fit Statistics  
Model selection based on \_TASE\_

| Selected Model | MODEL | Test: Average Squared Error | Train: Average Squared Error |
|----------------|-------|-----------------------------|------------------------------|
| Y              | Tree  | 0.20365                     | 0.20285                      |
|                | Reg   | 0.20939                     | 0.21034                      |

# Scoring New Data

Interval Variable Summary Statistics

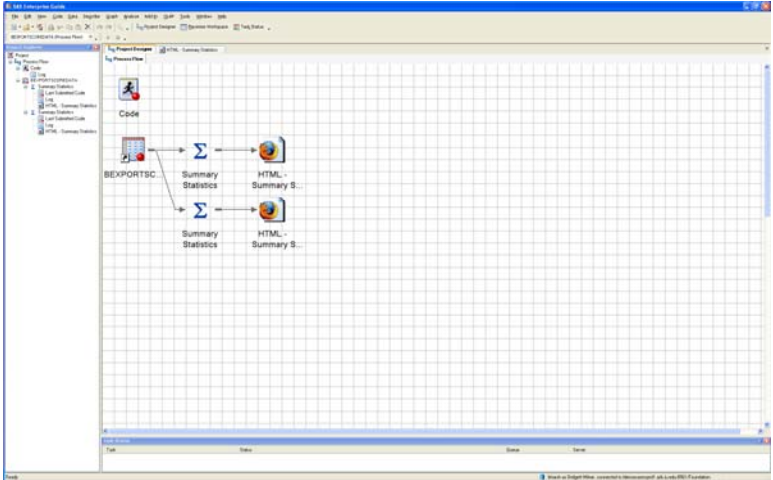
Variable Name=P\_matric

| Statistics | Label           | TRAIN    | VALIDATE | TEST     | SCORE    |
|------------|-----------------|----------|----------|----------|----------|
| MEAN       | Mean            | 0.36     | 0.36     | 0.36     | 0.34     |
| STD        | Std. Deviation  | 0.17     | 0.17     | 0.17     | 0.17     |
| N          |                 | 36048.00 | 27036.00 | 27037.00 | 20788.00 |
| MIN        | Minimum         | 0.14     | 0.14     | 0.14     | 0.14     |
| P25        | 25th Percentile | 0.21     | 0.21     | 0.21     | 0.19     |
| MEDIAN     | Median          | 0.29     | 0.29     | 0.29     | 0.24     |
| P75        | 75th Percentile | 0.55     | 0.55     | 0.55     | 0.53     |
| MAX        | Maximum         | 1.00     | 1.00     | 1.00     | 0.75     |

# Project 1: Extract Scored Data


```
SAS Code:  
Data MyData.BExportScoreData;  
Set &EM_IMPORT_SCORE;  
Run;
```

## Open & Explore Scored Data in Enterprise Guide



The screenshot shows the SAS Enterprise Guide interface. On the left is a project tree with folders for 'Data', 'Code', and 'Results'. The main workspace contains a workflow diagram. It starts with a 'BEXPORTSIC' node, which branches into two 'Summary Statistics' nodes. Each 'Summary Statistics' node is connected to an 'HTML Summary S...' node. The background is a large grid.

## Scored Data

 Enterprise Guide. *The Power to Know.*

**Summary Statistics Results**

*The MEANS Procedure*

reside=0

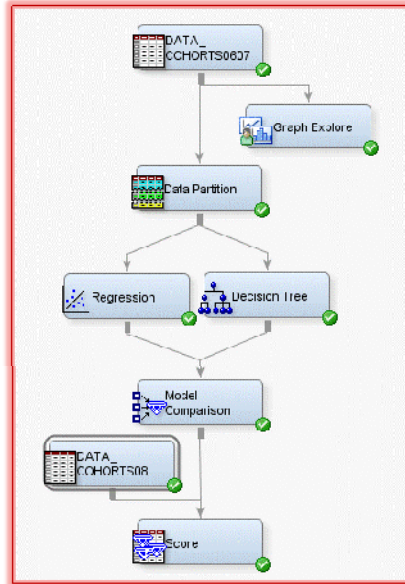
| Analysis Variable : EM_PREDICTION Prediction for matric |           |           |           |         |       |
|---|-----------|-----------|-----------|---------|-------|
| Mean  | Std Dev   | Minimum   | Maximum   | Sum     | N     |
| 0.2185044   | 0.0548304 | 0.1384615 | 0.4668675 | 2843.18 | 13012 |

reside=1

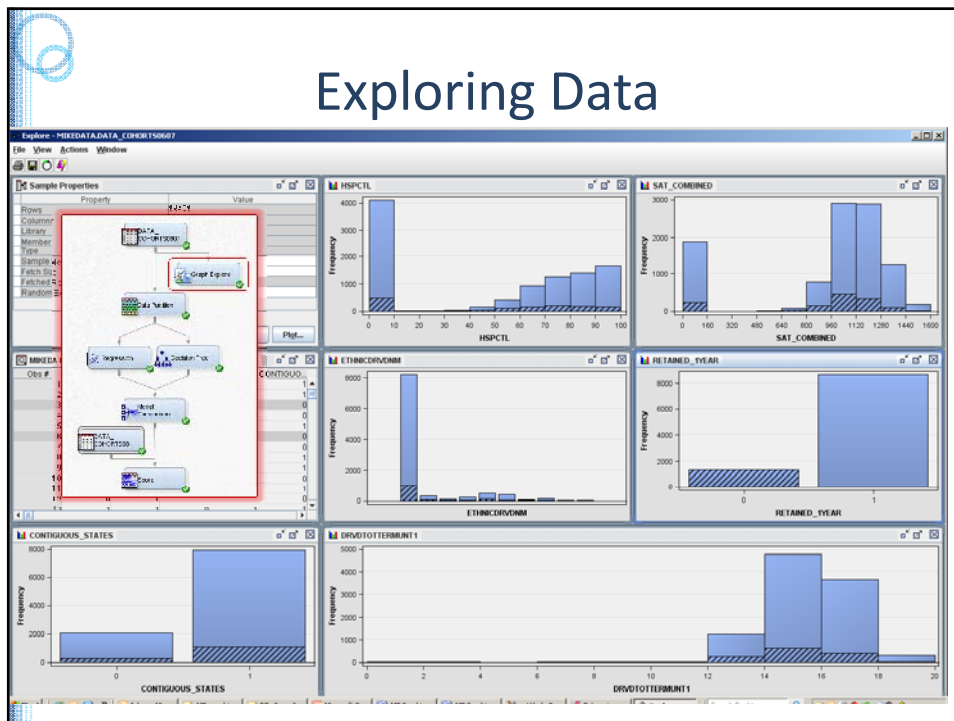
| Analysis Variable : EM_PREDICTION Prediction for matric |           |           |           |         |      |
|---|-----------|-----------|-----------|---------|------|
| Mean  | Std Dev   | Minimum   | Maximum   | Sum     | N    |
| 0.5383844   | 0.0733461 | 0.2899408 | 0.7522124 | 4186.48 | 7776 |

*Generated by the SAS System (Local, XP\_PRO) on 27MAY2009 at 2:28 PM*

## Project 2: Predicting 'at risk' students from enrollment data



## Exploring Data



# Partitioning Data

The screenshot shows the Enterprise Miner interface. On the left, a tree view shows 'AIR09mjs' with sub-items for 'Data Sources', 'Diagrams', 'Flow1', 'Model Packages', and 'Users'. Below this is a 'Property' window with the following details:

| Property                    | Value                    |
|-----------------------------|--------------------------|
| <b>General</b>              |                          |
| Node ID                     | Part                     |
| Imported Data               |                          |
| Exported Data               |                          |
| Notes                       |                          |
| <b>Train</b>                |                          |
| Variables                   |                          |
| Output Type                 | Data                     |
| Partitioning Method         | Simple Random            |
| Random Seed                 | 12345                    |
| <b>Data Set Allocations</b> |                          |
| Training                    | 40.0                     |
| Validation                  | 30.0                     |
| Test                        | 30.0                     |
| <b>Report</b>               |                          |
| Interval Targets            | Yes                      |
| Class Targets               | Yes                      |
| <b>Status</b>               |                          |
| Creates Time                | 5/14/09 12:05 PM         |
| Run Id                      | e8287f5f-a8f4-4aef-9db1- |
| Last Error                  |                          |
| Last Status                 | Complete                 |
| Last Run Time               | 5/14/09 12:00 PM         |

The main workspace shows a workflow diagram for 'Flow1'. It starts with a 'DATA\_COHORTS0607' node, followed by a 'Data Partition' node. The partitioned data then flows into a 'Graph Explorer' node. Below the 'Graph Explorer' is a decision tree model, which is highlighted with a red box. The decision tree has a root node 'DRVDTOTTERMUNT1' with a split on the variable 'RESIDENT'. The left branch is for 'RESIDENT = 0' and the right branch is for 'RESIDENT = 1'. Each branch further splits on the variable 'SAT\_COMBINED'.

# Decision Tree

The screenshot displays a detailed decision tree structure. The root node is 'DRVDTOTTERMUNT1', which splits on the variable 'RESIDENT'. The left branch is for 'RESIDENT = 0' and the right branch is for 'RESIDENT = 1'. Each branch further splits on the variable 'SAT\_COMBINED'.

Statistical data for each node:

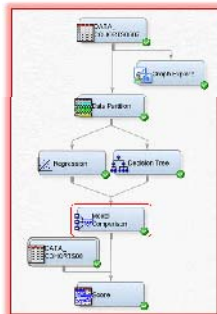
- Root Node: DRVDTOTTERMUNT1**
  - Statistic: 1: 86.4%, 0: 13.6%
  - Training: 87.2%
  - Validation: 12.0%
  - N in Node: 5780
- Left Branch: RESIDENT = 0**
  - Statistic: 1: 78.7%, 0: 21.3%
  - Training: 80.2%
  - Validation: 19.8%
  - N in Node: 774
- Right Branch: RESIDENT = 1**
  - Statistic: 1: 87.6%, 0: 12.4%
  - Training: 88.2%
  - Validation: 11.8%
  - N in Node: 5006
- Left-Left Branch: RESIDENT = 0, SAT\_COMBINED < 805**
  - Statistic: 1: 54.8%, 0: 45.2%
  - Training: 59.5%
  - Validation: 40.5%
  - N in Node: 62
- Left-Right Branch: RESIDENT = 0, SAT\_COMBINED >= 805**
  - Statistic: 1: 75.4%, 0: 24.6%
  - Training: 78.1%
  - Validation: 21.9%
  - N in Node: 451
- Right-Left Branch: RESIDENT = 1, SAT\_COMBINED < 805**
  - Statistic: 1: 78.3%, 0: 21.7%
  - Training: 81.0%
  - Validation: 19.0%
  - N in Node: 425
- Right-Right Branch: RESIDENT = 1, SAT\_COMBINED >= 805**
  - Statistic: 1: 84.5%, 0: 15.5%
  - Training: 88.5%
  - Validation: 16.5%
  - N in Node: 3826

# Regression Results?

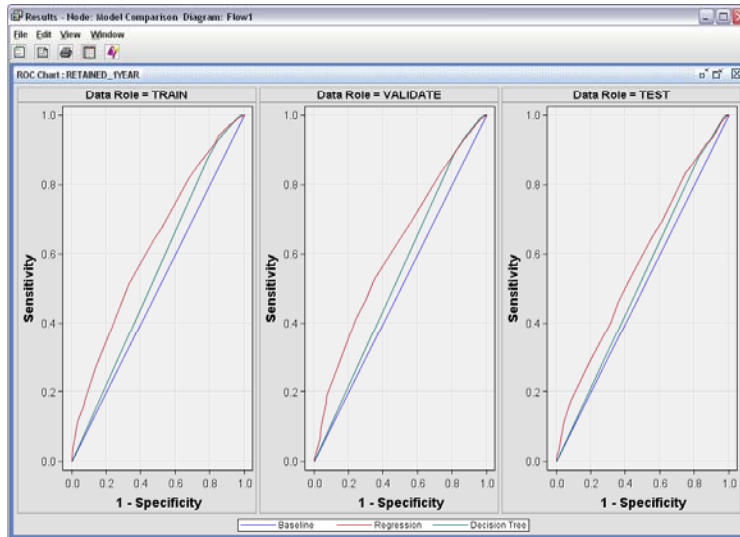
Fit Statistics

Model selection based on \_TMISC\_

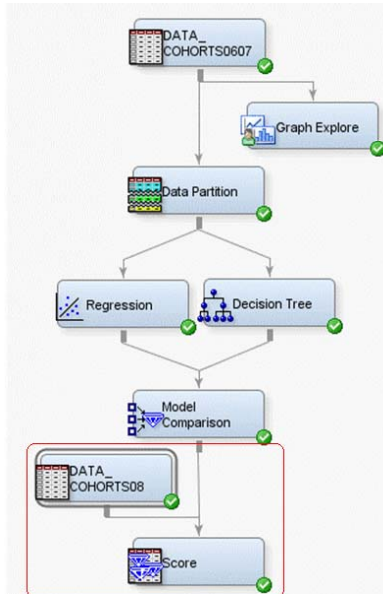
| Selected Model | Model Node | Test: Misclassification Rate | Train: Average Squared Error | Train: Misclassification Rate | Valid: Average Squared Error | Valid: Misclassification Rate |
|----------------|------------|------------------------------|------------------------------|-------------------------------|------------------------------|-------------------------------|
| Y              | Tree       | 0.12269                      | 0.11507                      | 0.13391                       | 0.11035                      | 0.12734                       |
|                | Reg        | 0.12292                      | 0.11459                      | 0.13633                       | 0.10972                      | 0.12826                       |



Hmmm....

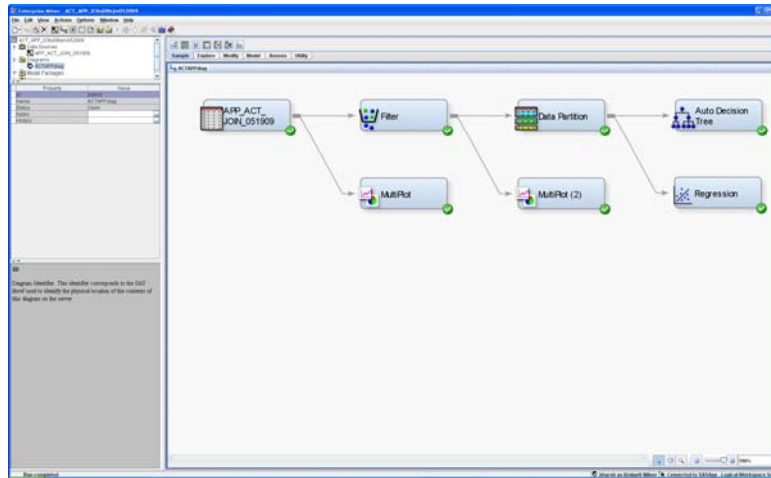


## Next Hurdle - Score New Data



## ACT Profile Data

- Revisit Project 1 and 2 and include ACT data



## Project 3: Market Segmentation

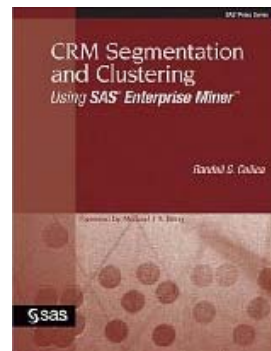
GOAL: Segment markets for targeted admissions and scholarship materials

Training:

- Customer Segmentation Using SAS Enterprise Miner, 2-day training

- Presented by Goutam Chakraborty, Ph.D., professor of marketing at Oklahoma State University

Documentation:





## Project 3: Market Segmentation

GOAL: Segment markets for targeted admissions and scholarship materials


– Still working on:

- Leaders ready for new initiative
- Operational people available to coordinate



## Overall “Triumphs”

- New framing of research questions
- Integrated workflow documentation
- Grew local support group
- Collaboration across units





## Future Directions

- Future plans:
  - More training - specialists
  - Local support group collaborations
  - Manage expectations
  - Support dedicated time in office just for exploration with tool
  - Develop technical expertise with an application user
  - Projects



## Jumping into Data Mining: The Triumphs and Tragedies of Our Second Year

---

Douglas K. Anderson, Director, *Enrollment Planning & Research*  
Bridgett J. Milner, Assistant Director, *Enrollment Planning & Research*  
Michael J. Sauer, Senior Research Analyst, *Office of the Registrar*  
Linda Shepard, Senior Associate Registrar, *Office of the Registrar*

Indiana University – Bloomington

[www.indiana.edu/~oem/pages/presentations.php](http://www.indiana.edu/~oem/pages/presentations.php)

