Use Data Mining Techniques to Assist Institutions in Achieving Enrollment Goals: A Case Study

Tongshan Chang
The University of California Office of the President
CAIR Conference in Pasadena
11/13/2008

Email: Tongshan.Chang@ucop.edu
Data Mining: Concepts

- **SAS**: “the process of sampling, exploring, modifying, modeling, and assessing (SEMMA) large amounts of data to uncover previously unknown patterns, which can be utilized as a business advantage.” (Applying Data Mining, 2005, p. 1-3)


- **Berry and Linoff**: “Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. ...the goal of data mining is to allow a corporation to improve its marketing, sales, and customers.” (Data Mining Techniques, p.7).
Data Mining: What Can We Do with It?

- **Classification**: discrete outcomes: yes or no
- **Estimation**: continuous values outcomes
- **Prediction**: the same as classification or estimation, but classifying according to some predicted future behavior or estimated future value
- **Association Rules**: determine which things go together
- **Clustering**: segment a heterogeneous population into a number of more homogeneous subgroups or clusters
- **Description and Profiling**: simply describe what is going on in a complicated database
Data Mining: Techniques—Decision Tree

- **Decision Tree**
  - Divide up a large collection of records into smaller sets of records using decision rules
  - Process: Record $\rightarrow$ Root Node $\rightarrow$ Child Node $\rightarrow$ Leaf Node
  - The PATH is an expression of the rules used to classify the records.
    - 3 paths in this tree
      - $\text{GAP} \geq 3.40$
      - GPA $< 2.40 \rightarrow$ SAT $\geq 1100$
      - GPA $< 3.40 \rightarrow$ SAT $< 1100$
Data Mining: Techniques—Regression (Logistic Regression)

Regression

Logistic Regression

Data Mining: Techniques—Neural Network

- **Neural Network**
  - Similar property of biological neurons
  - Interconnected artificial neurons
  - Inputs $\rightarrow$ Hidden Layer $\rightarrow$ Output(s)
  - Weights
    - Inputs and Hidden Layer
    - Hidden Layer and Output
Data Mining: Techniques—Ensemble

- **Ensemble**: Averaging the posterior probabilities for class targets or the predicted values for interval targets from multiple models

- **Methods**:
  - Different models from the same modeling method based on separate samples of training data set
Data Mining: Techniques—Ensemble

- **Ensemble**: Averaging the posterior probabilities for class targets or the predicted values for interval targets from multiple models

- **Methods**:
  - Different models from the same modeling method based on three separate samples of training data set
  - Different models from the different modeling methods based on the same training data set
Data Mining: Applications in Institutional Research

- **College admissions yield** (Chang, 2006)
- **Retention** (Herzop, 2006; Sujitparapitaya, 2006)
- **Time to degree** (Eykamp, 2006; Herzop, 2006)
- **Enrollment management** (Aksenova, Zhang, & Lu, 2006)
- **Course offerings** (Luan, 2006; Dai, Yeh, & Lu, 2007)
- **Student performance** (Dede & Clarke, 2007; Heathcote & Dawson, 2005; Minaei-Bidgoli, 2004; Ogor, 2007)
- **Graduation rate** (Baily, 2006)
- **Student experience survey study** (Yu, et. al, 2007)
A Case Study Using SAS Enterprise Miner

Assist Institutions in Achieving Enrollment Goals
A Case Study Using SAS Enterprise Miner—Background

- **Paths to Eligibility for CA Residents at UC**
  - Eligibility in the Statewide Context
  - Eligibility in the Local Context (ELC)
  - Eligibility by Examination Alone

- **Admissions:**
  - UC guarantees to admit all CA eligible applicants, but does not guarantee to admit everyone in terms of the campus or the program he/she applied to.
A Case Study Using SAS Enterprise Miner—Background

- **Referral Pool:**
  - Eligible, not admitted
  - To the referral pool
  - Two UC campuses: Riverside and Merced
  - Don’t know until April, too late, so the yield rate is low

- **Early Referral Pool:**
  - A letter to those who may be in the referral pool
  - Admit those who would like to consider these two campuses

- **Question:** Who do we send a letter to?
A Case Study Using SAS Enterprise Miner—Purpose

- Predict UC applicants who are qualified to UC admissions systemwide, but not admitted to the campus they applied to
- Two campuses use the information to make Early Referral Pool admissions offers and try to enroll more students.
A Case Study Using SAS Enterprise Miner—Data Description

- **UC Freshman Application Data**
  - **Data Sets:**
    - Fall 2007 data, training data
    - Fall 2008 data, target data
  - **Observations (Eligible Applicants):**
    - Fall 2007: 45,393
    - Fall 2008: 48,356
  - **Elements**
    - Student demographic and academic information
    - Family information
    - Application information (campuses, major, etc.)

- **CDE School Performance Data**
  - **Academic Performance Index (API)**
## A Case Study Using SAS Enterprise Miner—Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Referral Pool</td>
<td>Dichotomous</td>
<td>Dependent variable: 1=in referral pool, 0=not in referral pool</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Categorical</td>
<td>7 categories</td>
</tr>
<tr>
<td>First Language</td>
<td>Categorical</td>
<td>3 categories: English Only, English and Another Language, and Another Language</td>
</tr>
<tr>
<td>Campus(es) Applied to</td>
<td>Categorical</td>
<td>7 variables, one for each campus: e.g. CAMP_BK: 1=applied to UC Berkeley, 0=not applied to UC Berkeley</td>
</tr>
<tr>
<td>Parent’s Educational Level</td>
<td>Categorical</td>
<td>5 Categories: HS or Less, 2 Year College, 4 Year College and Post Ed. Study, Missing</td>
</tr>
<tr>
<td>Family Income</td>
<td>Continuous</td>
<td></td>
</tr>
<tr>
<td>Home Location</td>
<td>Categorical</td>
<td>5 Categories: San Francisco Bay Areas, CA North, LA County, CA South, and Other</td>
</tr>
<tr>
<td>Discipline</td>
<td>Categorical</td>
<td>7 Variables, one for each campus: 5 categories for each variable: Engineering, Science, Social Science, Humanities, Others.</td>
</tr>
<tr>
<td>Outreach Programs</td>
<td>Dichotomous</td>
<td>Participated at least one or not participated in any one.</td>
</tr>
<tr>
<td>API Ranking</td>
<td>Categorical</td>
<td>1 to 10 for public schools, missing for private schools</td>
</tr>
<tr>
<td>High School GPA</td>
<td>Continuous</td>
<td>Weighted, Capped GPA</td>
</tr>
<tr>
<td>UC Score (SAT or ACT)</td>
<td>Continuous</td>
<td>Highest of converted SAT or ACT score, including 2 highest SAT subject tests</td>
</tr>
</tbody>
</table>
A Case Study Using SAS Enterprise Miner—Missing Value Imputation

- **Categorical Variable**: not necessary, “MISSING” is a category.
- **Continuous Variable**: 
  - **Discard vs. Impute**
    - For data accuracy, simply discard, but reduce data drastically
    - Scoring problem: records with missing values will not be scored
    - Decision tree modeling: not necessary
    - Logistic regression and neural network modeling: ignore all records with missing values
    - Compare models: on the same set of observations
  - **SAS Methods**: 11—mean, median, mid-range, tree, etc.
  - **Method for This Project**: median, tree, mean, etc. were used, but the best method is mean
A Case Study Using SAS Enterprise Miner—Data Transformation

- **Transformation**: highly skewed distribution, a great deal of influence
- **Decision tree and neural network modeling**: Flexible
- **Logistic regression modeling**: Transformation may yield a better fitting model
A Case Study Using SAS Enterprise Miner—Modeling Process

- **Data Partition:**
  - **Training Data Set:** Preliminary model fitting
  - **Validation Data Set:** Monitoring and tuning the model to improve its generalization
  - **Test Data Set:** Estimate of Generalization

- **Data Set Percentage:** User decides, but each observation is allowed to use only once, 40%, 30%, and 30%.

- **Four Models:** Decision Tree, Logistics Regression, Neural Network, and Ensemble
A Case Study Using SAS Enterprise Miner—Model Assessment

- Score Rankings Overlay
  - Lift
  - Cumulative Lift
  - Gain
  - % Response
  - Cumulative % Response
  - % Captured Response
  - Cumulative % Captured Responses

![Score Rankings Overlay: REFP](image)

- Decile = 40
- CAPC = 97.6
- Data Role = Validation
A Case Study Using SAS Enterprise Miner—Model Assessment

- **Fit Statistics**

<table>
<thead>
<tr>
<th>TARGET</th>
<th>Fit statistics</th>
<th>Statistics Label</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>REFP</td>
<td><em>AIC</em></td>
<td>Akaike's Information Criterion</td>
<td>8388.369</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFP</td>
<td><em>ASE</em></td>
<td>Average Squared Error</td>
<td>0.071947</td>
<td>0.073726</td>
<td>0.073328</td>
</tr>
<tr>
<td>REFP</td>
<td><em>AVERR</em></td>
<td>Average Error Function</td>
<td>0.228364</td>
<td>0.235384</td>
<td>0.235675</td>
</tr>
<tr>
<td>REFP</td>
<td><em>DFE</em></td>
<td>Degrees of Freedom for Error</td>
<td>18108</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFP</td>
<td><em>DFM</em></td>
<td>Model Degrees of Freedom</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFP</td>
<td><em>DFT</em></td>
<td>Total Degrees of Freedom</td>
<td>18156</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFP</td>
<td><em>DIV</em></td>
<td>Divisor for ASE</td>
<td>36312</td>
<td>27234</td>
<td>27240</td>
</tr>
<tr>
<td>REFP</td>
<td><em>ERR</em></td>
<td>Error Function</td>
<td>8292.369</td>
<td>6410.444</td>
<td>6419.8</td>
</tr>
<tr>
<td>REFP</td>
<td><em>FPE</em></td>
<td>Final Prediction Error</td>
<td>0.072329</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFP</td>
<td><em>MAX</em></td>
<td>Maximum Absolute Error</td>
<td>0.990743</td>
<td>0.999043</td>
<td>0.997095</td>
</tr>
<tr>
<td>REFP</td>
<td><em>MSE</em></td>
<td>Mean Square Error</td>
<td>0.072138</td>
<td>0.073726</td>
<td>0.073328</td>
</tr>
<tr>
<td>REFP</td>
<td><em>NOBS</em></td>
<td>Sum of Frequencies</td>
<td>18156</td>
<td>13617</td>
<td>13620</td>
</tr>
<tr>
<td>REFP</td>
<td><em>NW</em></td>
<td>Number of Estimate Weights</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFP</td>
<td><em>RASE</em></td>
<td>Root Average Sum of Squares</td>
<td>0.26823</td>
<td>0.271526</td>
<td>0.270791</td>
</tr>
<tr>
<td>REFP</td>
<td><em>RFPE</em></td>
<td>Root Final Prediction Error</td>
<td>0.26894</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFP</td>
<td><em>RMSE</em></td>
<td>Root Mean Squared Error</td>
<td>0.268585</td>
<td>0.271526</td>
<td>0.270791</td>
</tr>
<tr>
<td>REFP</td>
<td><em>SBC</em></td>
<td>Schwarz's Bayesian Criterion</td>
<td>8763.093</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFP</td>
<td><em>SSE</em></td>
<td>Sum of Squared Errors</td>
<td>2612.545</td>
<td>2007.863</td>
<td>1997.446</td>
</tr>
<tr>
<td>REFP</td>
<td><em>SUMMV</em></td>
<td>Sum of Case Weights Times F...</td>
<td>36312</td>
<td>27234</td>
<td>27240</td>
</tr>
<tr>
<td>REFP</td>
<td><em>MISC</em></td>
<td>Misclassification Rate</td>
<td>0.103602</td>
<td>0.104502</td>
<td>0.105874</td>
</tr>
</tbody>
</table>
A Case Study Using SAS Enterprise Miner—Model Assessment

- Importance of a variable in modeling: Tree Map
A Case Study Using SAS Enterprise Miner—Model Assessment

- **Tree:** The closer a variable is to the root node, the more prominent in the model.
- **Regression Effects Plot:** Displays a ranked plot of the effect scores most prominent in the model.

![Regression Effects Plot](image)
**A Case Study Using SAS Enterprise Miner—Model Comparison**

- **Receiver Operating Characteristics (ROC) Chart**: Measure of the predictive accuracy of a model.
- **Fit Statistics, Score Rankings Overlay, Output**
A Case Study Using SAS Enterprise Miner—Scoring and Deployment

- **Scoring:** Process to apply the model to new cases
  - Generate SAS Code
  - Cleaning target data
  - Calculate probability

- **Deployment:**
  - A list of students with a probability equal to or above 40% to two campuses
  - Campuses sent a letter to selected students
  - Campus made offers to those students who responded, allowed campuses to review their applications (Early Referral Pool)
A Case Study Using SAS Enterprise Miner—Results

- **Results**: Comparison with the actual referral pool
  - **Accuracy**
    - In terms of the number, accuracy rate: 93%
    - In terms of individual students,

<table>
<thead>
<tr>
<th>Predicted Probability</th>
<th>Predicted Referral Pool</th>
<th>Actual Referral Pool</th>
<th>Cumulative Accuracy Rate</th>
<th>Predicted Referral Pool as Cumulative % of Total Population</th>
<th>Actual Referral Pool as Cumulative % of the Entire Referral Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 - 100%</td>
<td>65</td>
<td>52</td>
<td>80.0%</td>
<td>0.1%</td>
<td>0.6%</td>
</tr>
<tr>
<td>80 - 89%</td>
<td>353</td>
<td>275</td>
<td>77.9%</td>
<td>0.7%</td>
<td>3.3%</td>
</tr>
<tr>
<td>70 - 79%</td>
<td>2,732</td>
<td>2,018</td>
<td>73.9%</td>
<td>5.6%</td>
<td>23.9%</td>
</tr>
<tr>
<td>60 - 69%</td>
<td>4,986</td>
<td>3,555</td>
<td>71.3%</td>
<td>10.3%</td>
<td>42.0%</td>
</tr>
<tr>
<td>50 - 59%</td>
<td>6,659</td>
<td>4,597</td>
<td>69.0%</td>
<td>13.8%</td>
<td>54.3%</td>
</tr>
<tr>
<td>40 - 49%</td>
<td>8,209</td>
<td>5,518</td>
<td>67.2%</td>
<td>17.0%</td>
<td>65.2%</td>
</tr>
</tbody>
</table>
A Case Study Using SAS Enterprise Miner—Results

- Results: Comparison with the actual referral pool
  - **Accuracy**
    - In terms of the number, accuracy rate: 93%
    - In terms of individual students,
  - **Yield**

<table>
<thead>
<tr>
<th></th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>Total</th>
<th>Early Referral Pool</th>
<th>Traditional Referral Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Referral Pool</td>
<td>6,170</td>
<td>6,090</td>
<td>6,923</td>
<td>9,300</td>
<td>1,099</td>
<td>8,201</td>
</tr>
<tr>
<td>SIRs(^1) from Actual Referral Pool</td>
<td>392</td>
<td>398</td>
<td>465</td>
<td>769</td>
<td>241</td>
<td>528</td>
</tr>
<tr>
<td>Referral Pool Yield Rate</td>
<td>6.4%</td>
<td>6.5%</td>
<td>6.7%</td>
<td><strong>8.3%</strong></td>
<td><strong>21.9%</strong></td>
<td>6.4%</td>
</tr>
<tr>
<td>Total SIRs from All Admits</td>
<td>3,691</td>
<td>4,006</td>
<td>4,412</td>
<td>5,770</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Referral Pool SIRs as % of Total SIRs</td>
<td>10.6%</td>
<td>9.9%</td>
<td>10.5%</td>
<td><strong>13.1%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Mining Workshop Information

Summer Program for Educators Teaching Data Mining

- **Track 1**: Basic SAS programming; **Track 2**: SAS Enterprise Miner
- **Location**: CSU Long Beach, the SAS Campus in Cary, NC
- **Time**: Early August
- **Registration Fee**: No
- **Text Books**: Free
- **Breakfasts and Lunches**: Every day and free
- Invited people only
- Invitation letter is sent out early February
- Contact the SAS Institute in January
- Contact person: Susan Walsh, susan.walsh@sas.com