

Information Discovery with Decision Trees (May 9, 2005)

Morgan C. Wang

Department of Statistics and Actuarial
Science

University of Central Florida

Orlando, FL 32816-2370

Lecture Outlines

- What is Data Mining?
- What is a Tree?
- Identification Rules with Tree
- Rule Based Regression with Tree
- Variable Reduction with Tree
- Conclusions

What is Data Mining?

- *Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of observational data in order to discover meaningful patterns and models to the data owner.* (David Hand, Heikki Mannila, and Padharic Smyth, 2001 and Michael J. A. Berry and Gordon S. Linoff, 2000)

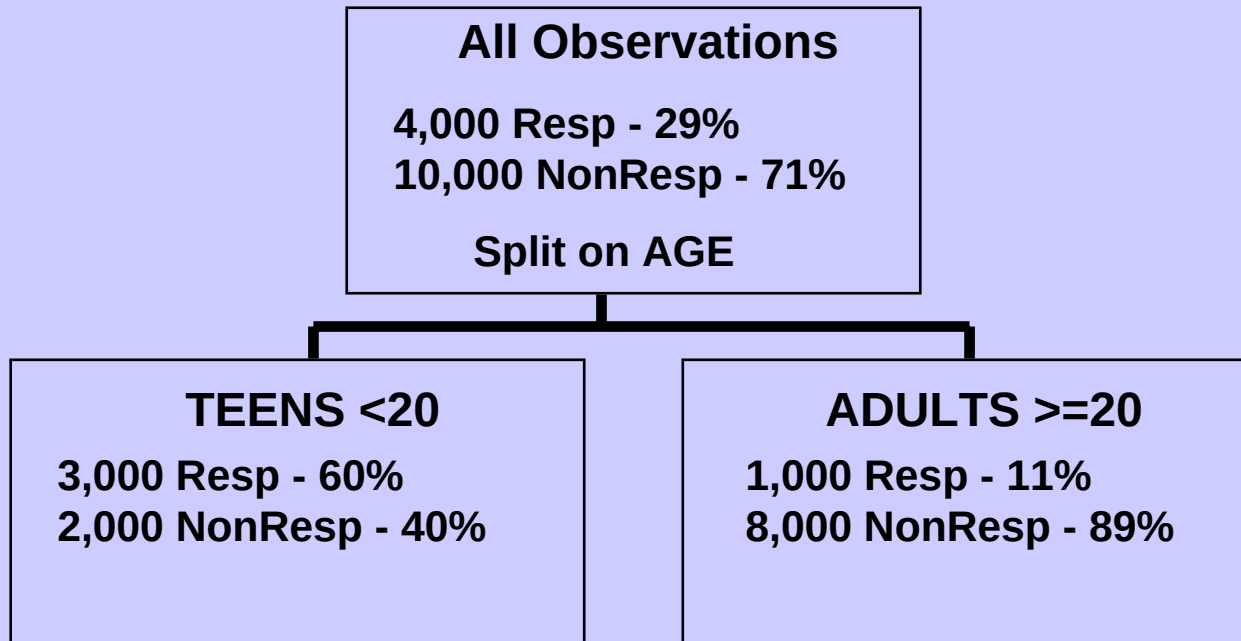
Data Mining: Why now?

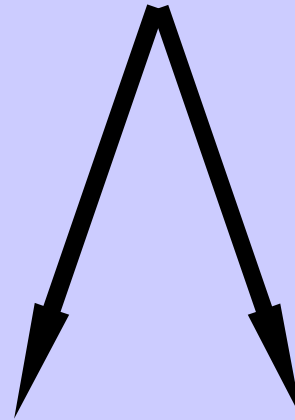
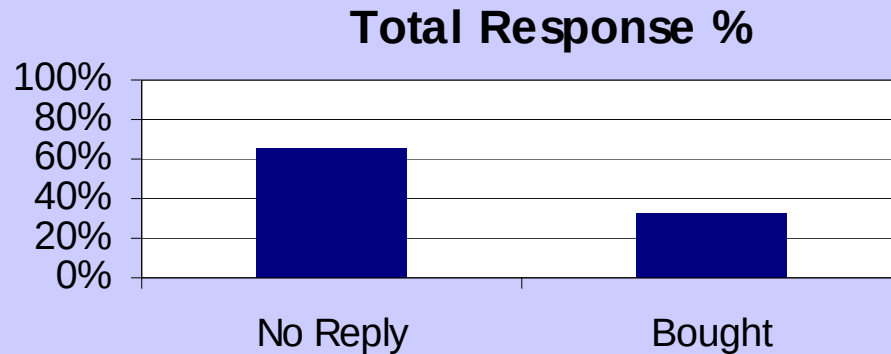
- Proliferation of high volume OLTP systems
- Proliferation of data warehouses
- Availability of scalable information technology
- Recognition of untapped value in large databases
- Increased competition

What is a Tree?

- Partitions data into subsets based on categories of input variables.
- Partitioning done to maximize predictability of target variable.
- Uses statistical test to determine best partitioning/grouping of categories.
- Target can be either categorical variable or continuous variable

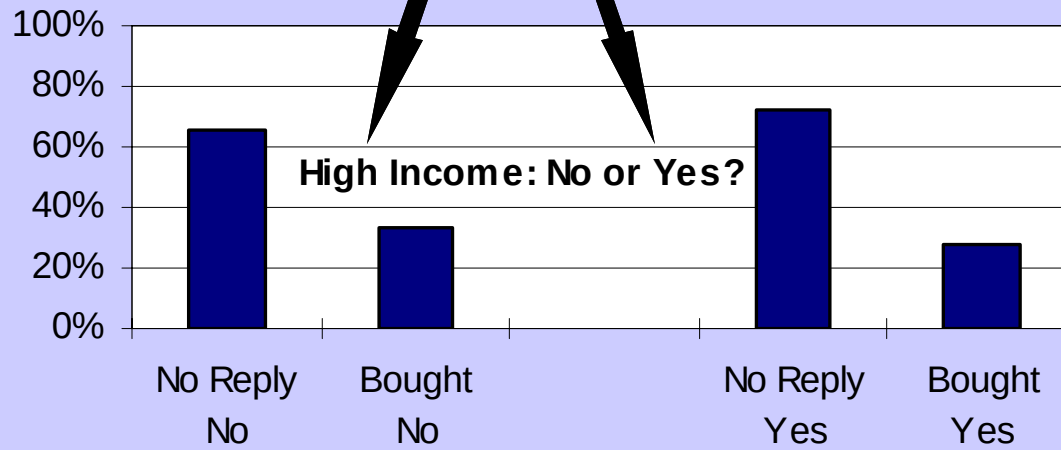
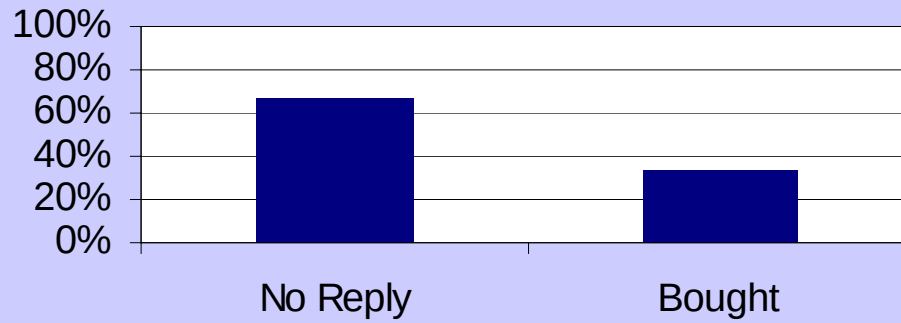
How Do Trees Work?



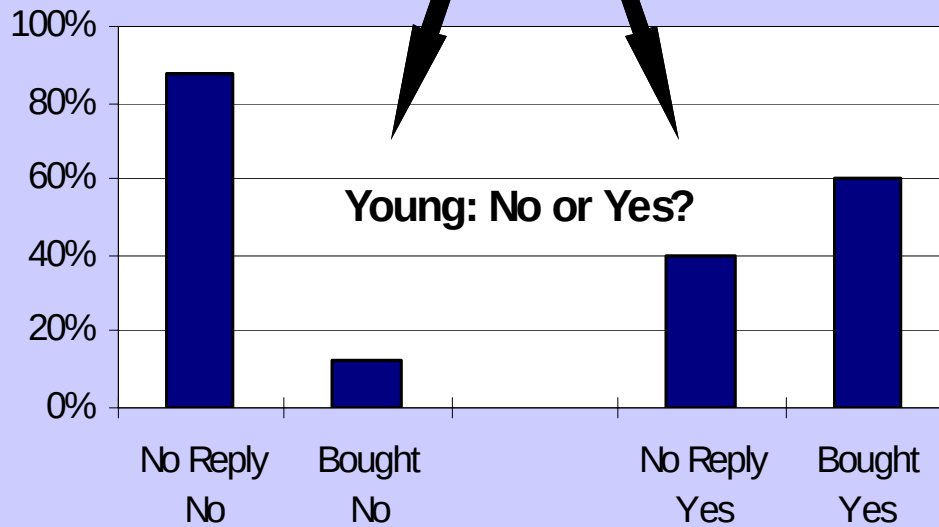
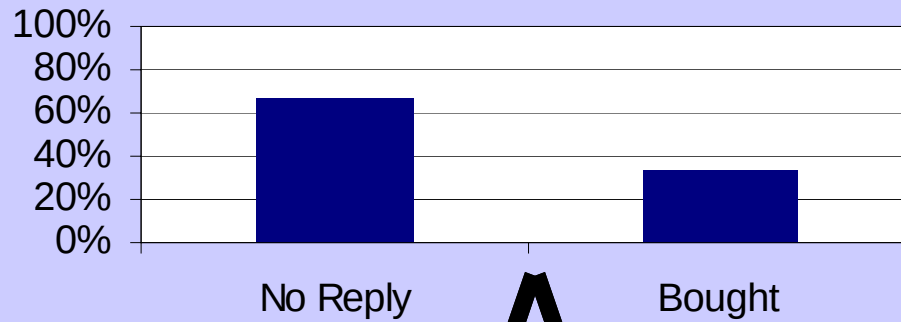


**Search for predictor with large difference
in response rate over its categories.**

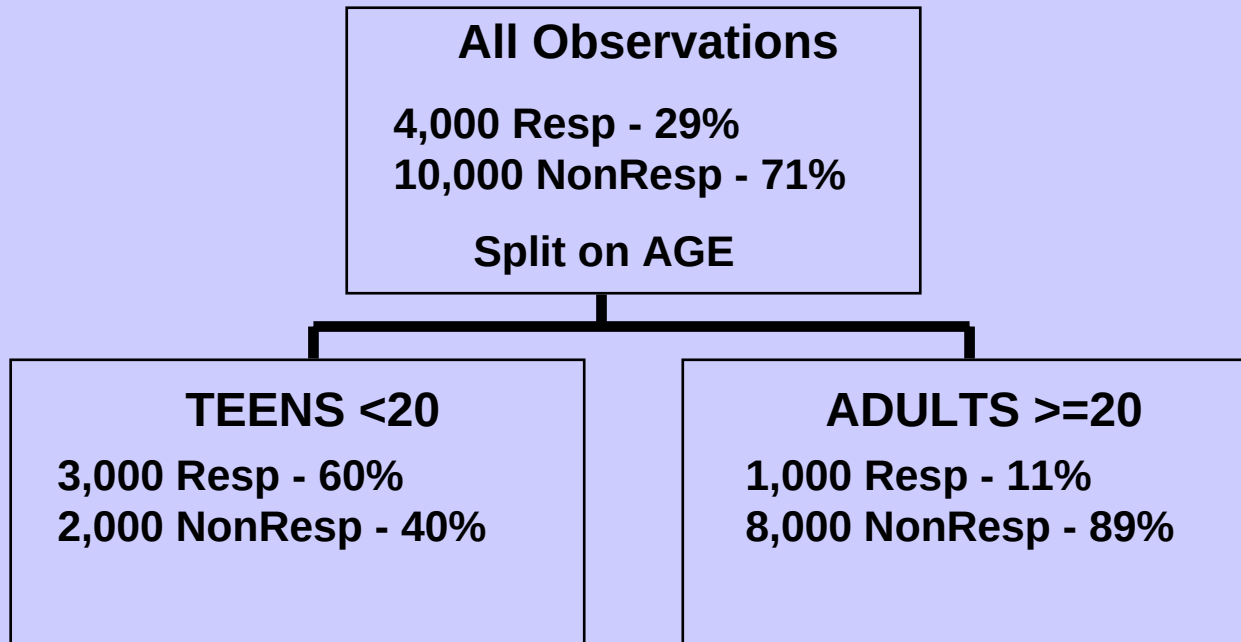
Total Response %



Total Response %



Tree Results



Why Are Trees Popular?

- **Easy to Understand ...**
 - Segmentation defined without formulae
 - Express rules using a single variable
 - User can create and modify rules
- **Easy to Use ...**
 - Accepts ordered and unordered variables
 - No transformation of predictors
 - Uses missing data

Case Study 1:

Identify Rules with Tree



Background

- Teaching evaluation data for a 3-year period were available to allow a comparison of two different types of measures (UCF/BOR).
- Responding to faculty interest, the UCF Faculty Senate requested that an evaluation of the Student Evaluation of Instruction measures be performed. (FS 1995-96-11)

The Instrument:

Board of Regent Items

- Description of course objectives and assignments
- **Communication of ideas and information**
- **Expression of expectations for performance**
- Availability to assist students in or outside of class
- **Respect and concern for students**
- Stimulation and interest in the course
- **Facilitation of learning**
- Overall assessment of instructor

Instrument: UCF Items

- Feedback concerning your performance in this course
- **The instructor's interest in your learning**
- **Use of class time**
- **The instructor's overall organization of the course**
- Continuity from one class meeting to the next
- The pace of the course
- The instructor's assessment of your progress
- The text and supplemental learning materials used
in this course

Traditional Methods

Exploratory Factor Analysis (16 Items)



Principal Components
Alpha Factor Analysis
Image Analysis
Transformation (Rotation)
Varimax
Promax

Traditional Methods

- Structure Equation Modeling through programs such as LISREL[®] (Jöreskog & Sörbom, 1993) that operate on a model that implies the covariance matrix:
 - This approach permits one to test hypotheses regarding the data and to judge the model fit. Often, this takes on the form of the measurement, confirmatory factor and full LISREL[®] models suggested by Jöreskog.

Traditional Methods

- Path Analysis:
 - Another popular procedure for identifying the causal relationships among variables is the path analysis model that finds the direct and indirect contributions of dependent variables to the outcome on dependent measures. Very often, however, path models are abstract and provide little that relates directly to process.

Traditional Methods

- Generalizability Theory
 - Generalizability theory (G-theory) identifies random variance components associated with pre-specified facets in a model. The outcome of G-Theory permits the investigator to gauge the relative contribution of a combination of factors to overall variance in a system. This approach has proved useful in many situations where the assumptions of classical reliability theory are restrictive. Although G-Theory is very flexible it does little in the way of developing rule-based decision making.

Why these approaches failed?

- All items are highly correlated
- All items considered are nominal level variables
- These procedures do not provide rule-based result
- Results produced by these methods are difficult to understand

Eigenvalues of the Correlation Matrix for the Sixteen Item Form (N=441,758)

Component (Factor)	Eigenvalue
1	11.36
2	.71
3	.57
4	.47
5	.38
6	.35
7	.33
8	.26
9	.25
10	.22
11	.21
12	.21
13	.19
14	.18
15	.18
16	.13

Decision Tree Based Approach

Binary or Multiple Decision Tree
SAS Enterprise Miner

Predictors (Target)

Remaining
15 Items

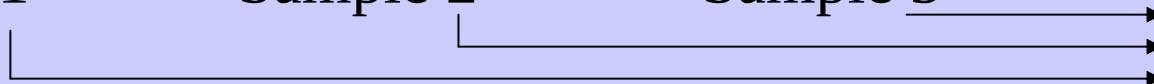
- Level
- Year
- College

Developmental
Sample 1

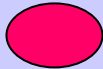
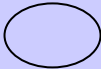
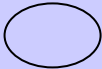
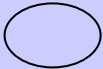
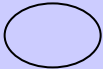
Developmental
Sample 2

Developmental
Sample 3

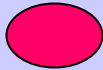
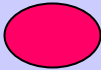
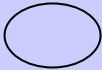
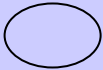
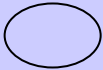
Validation
Sample



Rule #1: If...

	Excellent	Very Good	Good	Fair	Poor
Facilitation of learning					

&

Communication of ideas					
------------------------	---	---	---	---	---

Then...

The probability of an overall rating of *Excellent* = .93

&

The probability of an overall rating of *Fair* or *Poor* = .00

A college comparison of excellent ratings
(percentage)--unadjusted and adjusted
by Rule #1

<u>College</u>	<u>Unadjusted %</u>	<u>Adjusted %</u>
Arts & Sciences	41.6	92.4
Business	34.9	90.9
Education	56.8	94.8
Engineering	36.2	91.3
H&PA	46.1	93.9

(N=441,758) (N=147,544)

Rule #2: If...

	Excellent	Very Good	Good	Fair	Poor
Instructor is interested in your learning	☒	☐	☐	☐	☐
Organization of the course	☒	☐	☐	☐	☐
Respect & concern for students	☒	☐	☐	☐	☐
Communication of ideas	☒	☒	☐	☐	☐
Facilitation of learning	☐	☒	☒	☒	☒

Then...

The probability of an overall rating of *Excellent* = .84

&

The probability of an overall rating of *Fair* or *Poor* = .00

A college comparison of excellent ratings
(percentage)--unadjusted and adjusted
by Rule #2

<u>College</u>	<u>Unadjusted %</u>	<u>Adjusted %</u>
Arts & Sciences	41.6	86.8
Business	34.9	81.6
Education	56.8	81.9
Engineering	36.2	84.4
H&PA	46.1	84.5
	N=441,758	N=15,207

Rule #3: If...

	Excellent	Very Good	Good	Fair	Poor
Instructor is interested in your learning	☒	☐	☐	☐	☐
Respect and concern for students	☒	☐	☐	☐	☐
Expresses expectations for student performance	☒	☐	☐	☐	☐
Communication of ideas	☒	☒	☐	☐	☐
Organization of the course	☐	☒	☒	☐	☐
Facilitation of learning	☐	☒	☒	☒	☒
Use of class time	☒	☒	☐	☐	☐

Then...

The probability of an overall rating of *Excellent* = .78

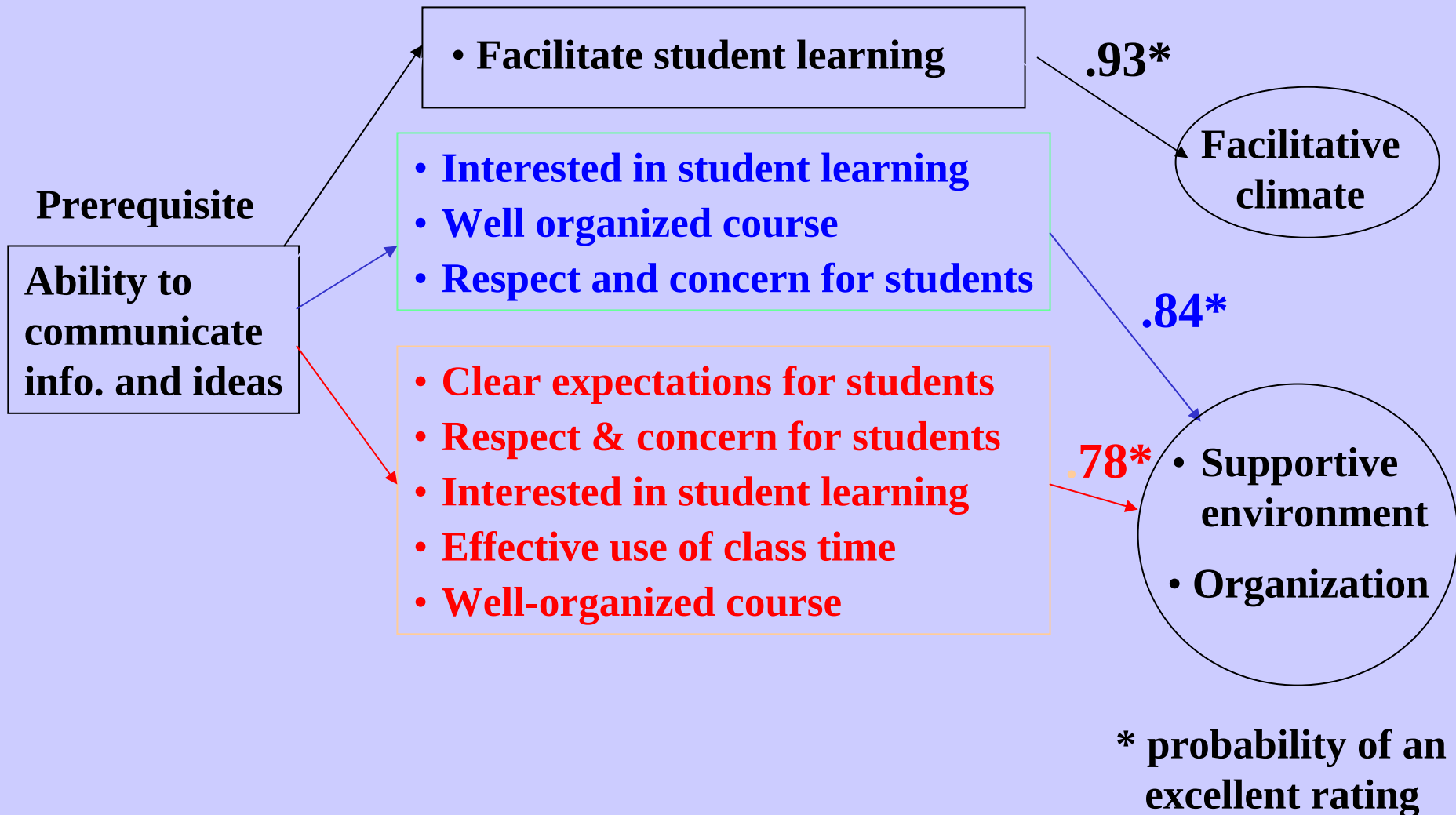
&

The probability of an overall rating of *Fair* or *Poor* = .00

A college comparison of excellent ratings
(percentage)--unadjusted and adjusted
by Rule #3

<u>College</u>	<u>Unadjusted %</u>	<u>Adjusted %</u>
Arts & Sciences	41.6	81.6
Business	34.9	78.0
Education	56.8	74.5
Engineering	36.2	80.0
H&PA	46.1	77.6
	N=441,758	N=15,060

A conceptual path diagram for an instructor receiving an overall rating of Excellent at UCF



Case Study 2:

Student Retention using Decision Trees



Background

- Student retention has been studied for about 70 years
- Overall, retention rates have changed very little during the past 70 years
- Generally report student retention as the percentage of a freshman class cohort that return to begin their sophomore year
- Now, as in 1985, most of those who drop-out do so during the freshman year
- Half of all students who don't graduate drop out before their second year
- Freshman Retention Rate:
 - National Average: 74% (US News & World Report)
 - UCF: 84% in 2004

Top 10 Public Doctoral University Freshman Retention Rate

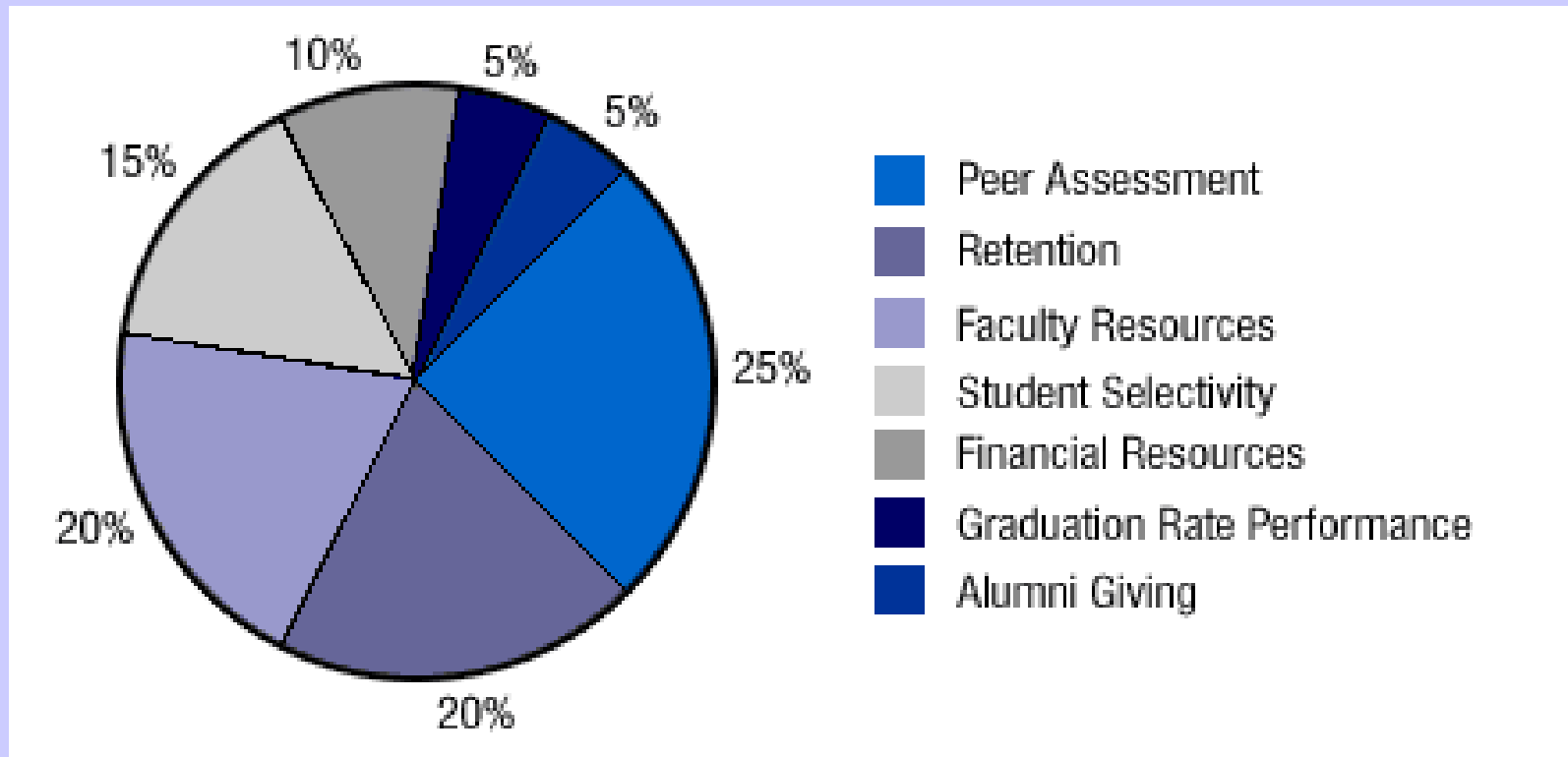
Rank	University	Retention Rate
1	University of California, Los Angeles	97%
1	University of Virginia	97%
3	University of Michigan, Ann Arbor	96%
4	University of North Carolina, Chapel Hill	95%
4	University of California, Berkeley	95%
6	University of California, San Diego	94%
7	University of California, Irvine	93%
8	University of Florida	92%
8	Pennsylvania State University	92%
8	University of Illinois, Urbana-Champaign	92%

Source: U.S. News & World Report , 2004

Why retention is very important?

- Social Perspective:
 - Life time earning power (Day and Newburger, 2002) :
 - High school graduates: \$1.2 million
 - Associate degree: \$1.6 million
 - College degree: \$2.1 million
- University perspective:
 - Quality (measure of academic excellence)
 - Funding (Federal, State, Local)
 - Loss of revenue (CRM, university loss of revenue)

Measuring Academic Excellence in Higher Education



This graph shows the relative weights assigned to each category of indicator for national universities and liberal arts colleges.

(Retention is 20%)

Why retention is very important?

(continued)

- University Funding
 - Retention is a *priority* for the Federal Government and State Government
- Retention is a national issue
 - **“The Education Department will examine retention policies and look at ways to use federal money in an incentive fashion to reward programs that work”**
- Retention is a CRM issue
 - In times of “tight” state education budgets, saving money by keeping existing, enrolled students, can be critical for universities

Retention Theories

- Astin's (1977, 1985) Theory of Involvement
- Bean's (1980, 1983) Model of Work Turnover to Student Attrition
- Bean and Metzner's (1985) Nontraditional Student Attrition
 - Environmental factors have a greater impact on departure decisions of adult students than academic variables.
- Kamens (1971, 1974) "multi-institutional data"
- McNeely (1937) "College Student Mortality"
 - Examined many factors in college student retention including time to degree, when attrition was most prevalent in a student's education, impact of college size etc.
- Spady Model (1971)
 - Interaction between student characteristics and campus environment
- Summerskill (1962)
 - Personality attributes of students is the main reasons for persistence and leaving.
- Tinto Model (1975, 1993)
 - Academic and social integration with the formal and informal academic and social systems of a college.

Retention Theories (continued)

- These models are not based on data that are collected on a routine basis by universities, thereby prohibiting student retention offices from using them on a regular basis
- Factors identified by these theoretical models cannot be used to prevent a particular student from dropping out since they simply document the circumstances for a student who has already dropped out
- These approaches do not provide a “risk score” for each student. It is not easy or cost effective to develop prevention programs because the number of students in the at risk group can be very large and variable.

Why use data mining?

- No additional data collection needed
- Treat each student as an individual
- Assign each student a “risk score” before entering college
- A set of easy to understand rules and to guide the administration to develop prevention program to target the students at risk
- Emphasize on prediction, i.e., on prevention in stead of documentation

Retention Goal

- At the present time, UCF retention studies have been limited to simple year-by-year demographic summaries which do not fully explain student progression patterns or trends. Academic Development and Retention has been gathering data on program attendance, attitudes, and opinions from various sources, and believes that student behavior can be explained with a more sophisticated method of data analysis.
- The development of a dynamic retention modeling tool will enable the university to predict and thereby improve student success by identifying factors, or combinations of factors, that positively or negatively influence retention.
- A retention data model will also provide a powerful management tool for administrators to use when planning new programs, assessing current ones, and when making budgetary decisions.

Data

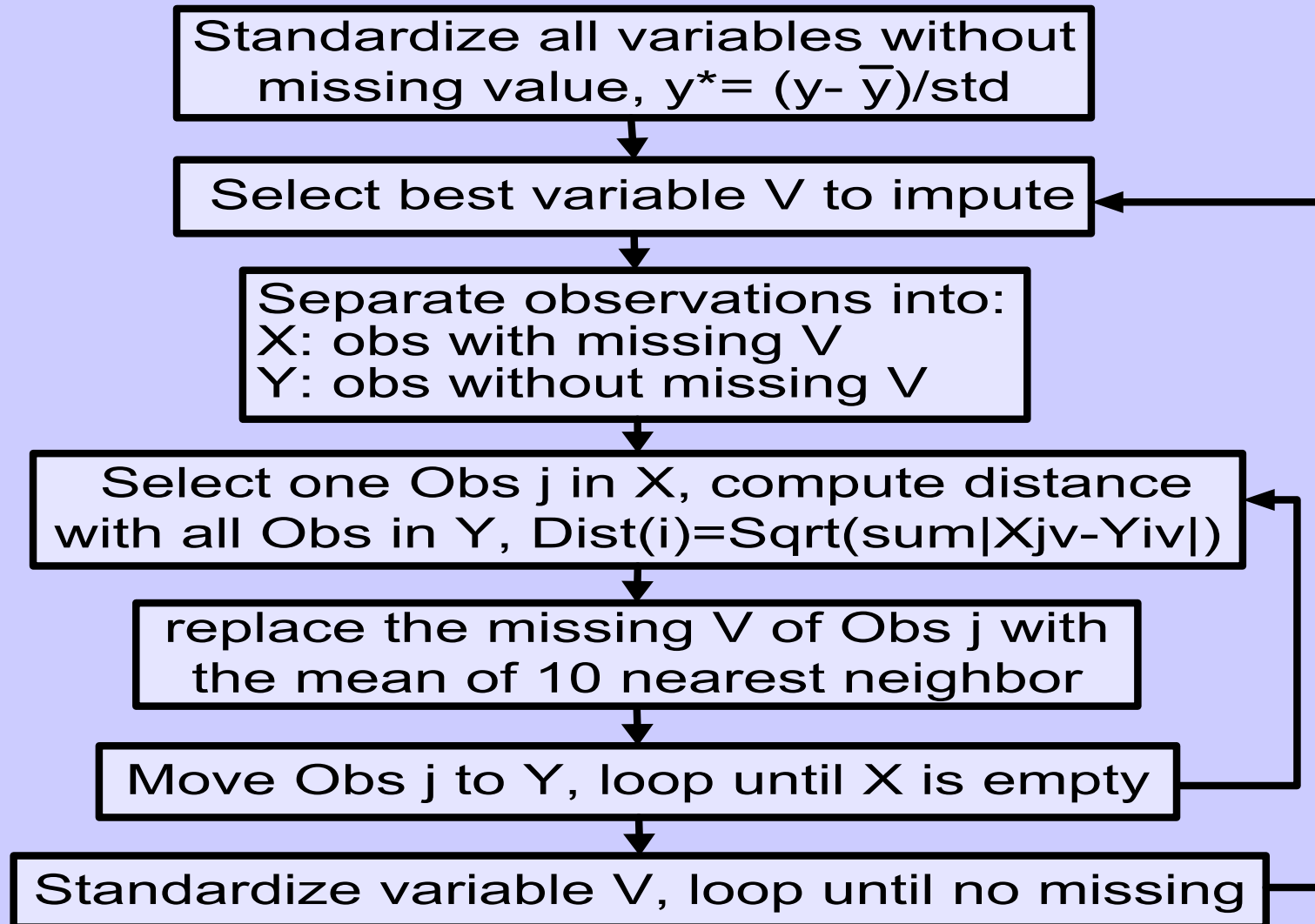
- **CIRP** (Cooperative Institutional Research Project) Survey in 2002
- **6945** observations
- **316** Variables:
 - 20 numerical variables: SAT_Verb, SAT_Math, Income
 - 213 nominal variables: Ethnic, Student_status, Goal
 - 36 ordinal variables: HSGPA, Age, ...
 - 47 binary variables: Gender, Full_status, Churn...
- **Binary Target:**
 - To predict who will drop-out in the upcoming semester

Data Problems

- **Missing Values:** More than 64% observations have one or more variables that have missing values
 - **ACT_Composite_Score** 46%
 - **Highest_Degree_Plan** 39%
 - **Finance_AID_From_Other** 53%
 - **Finance_AID_Must_Repay** 30%
- **More than 300 Variables**
- **Nominal variables with many levels**
- **Variable mis-scale problem**

Continuous Variable Imputation

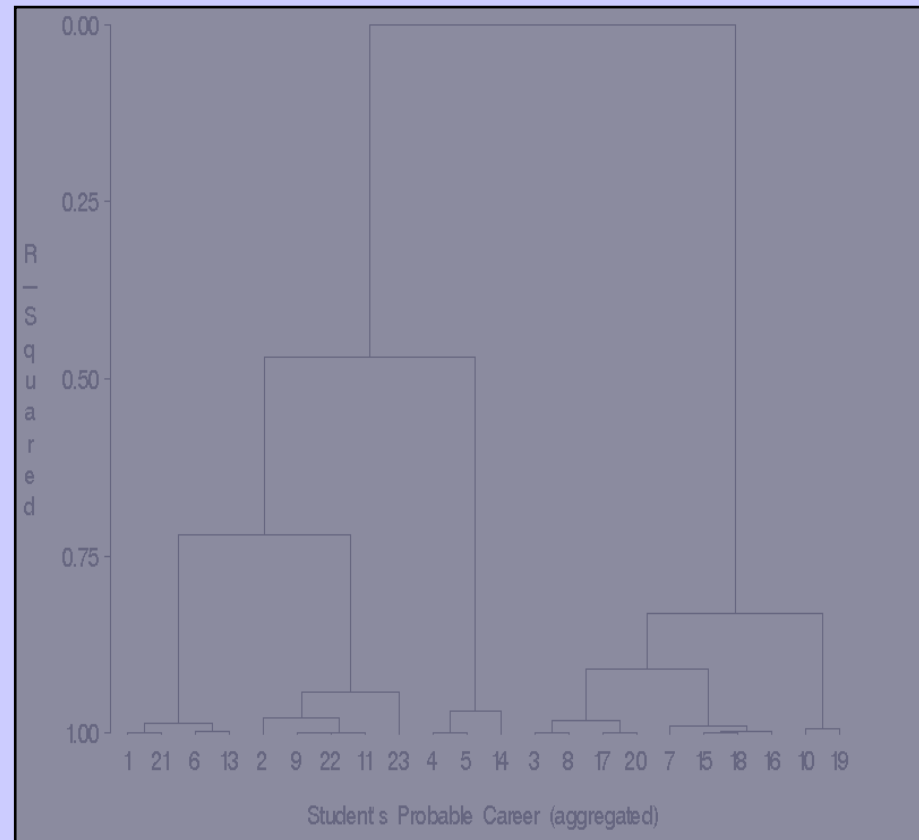
--- Nearest Neighbor Algorithm



Nominal Level Collapsing

Clustering rule

1. Get proportion of the target variable for each level
2. Create the output dataset cluster that include clustering procedure
3. Compute the chi-square statistics
4. Find the number of cluster associated with lowest p-value

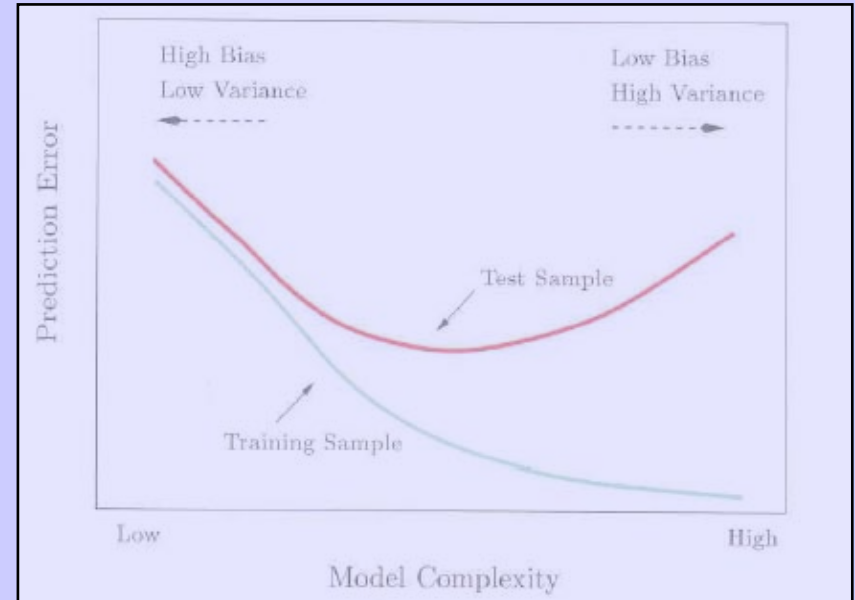


23 levels in Student_Prob_Career are reduced to 4 levels

Variable Selection

Goal:

- Look for the best tradeoff between bias of the model and variance of the model



Ways:

- Building Regression Tree to find the important variables
- Spearman correlation and Hoeffding' D statistics
Remove variables with large p-value
- Variable Clustering
choose one variable from each cluster & Select appropriate number of clusters

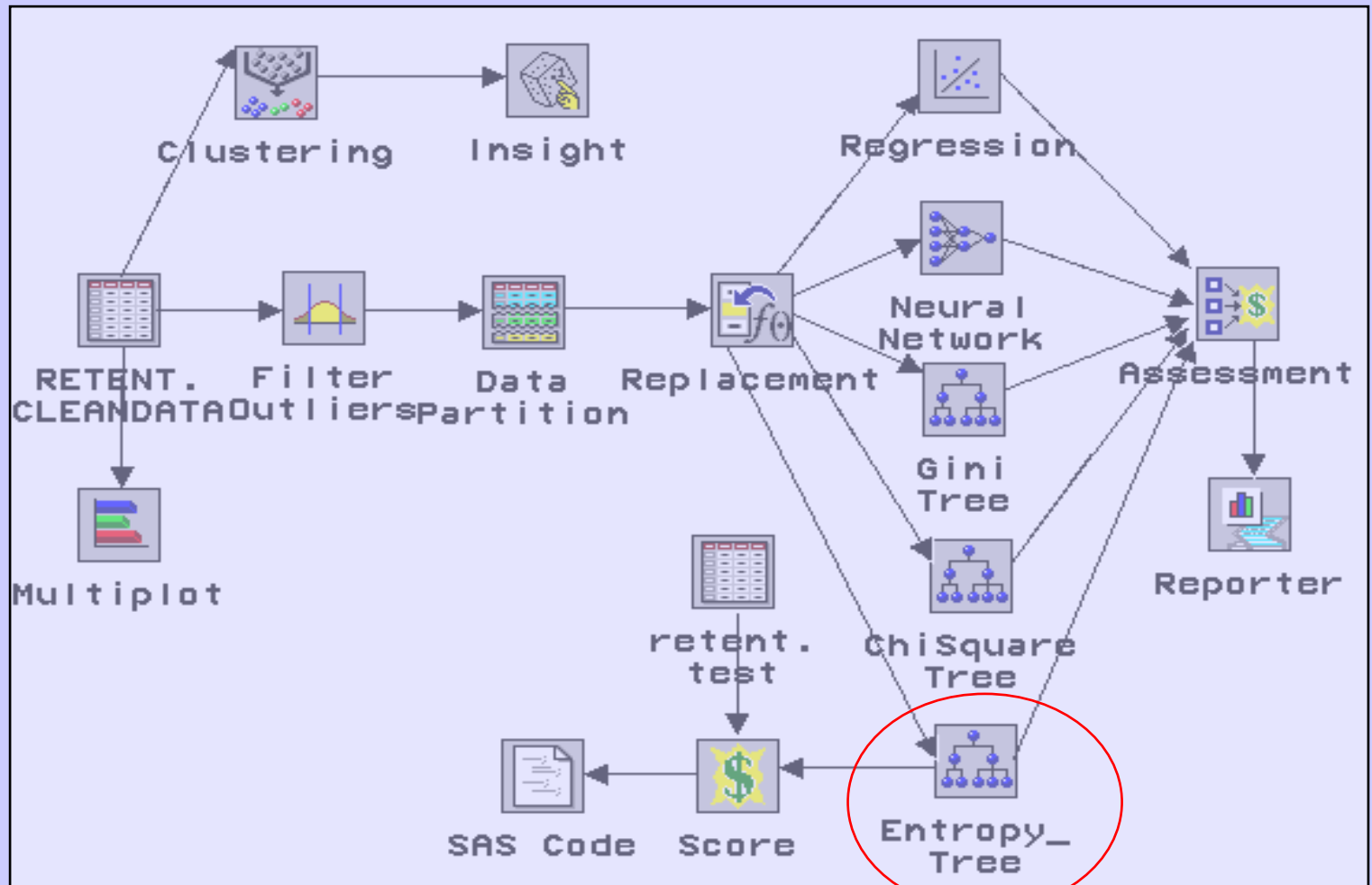
Model Building

- Data Partition:
 - 40% Training
 - 30% Validation
 - 30% Testing
- Models are constructed using both training and validation data sets, and be tested using the “testing” data set
- Several modeling techniques are used, e.g., logistic regression, neural network, decision trees, and clustering

Predictive Models

--- (SAS/Enterprise Miner)

Process Flow Diagram

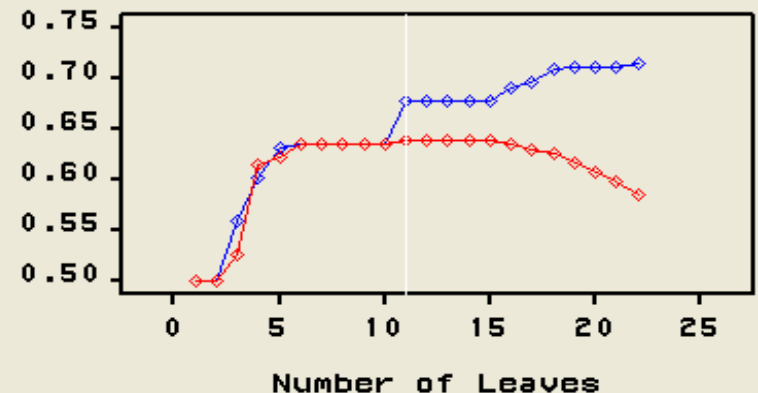
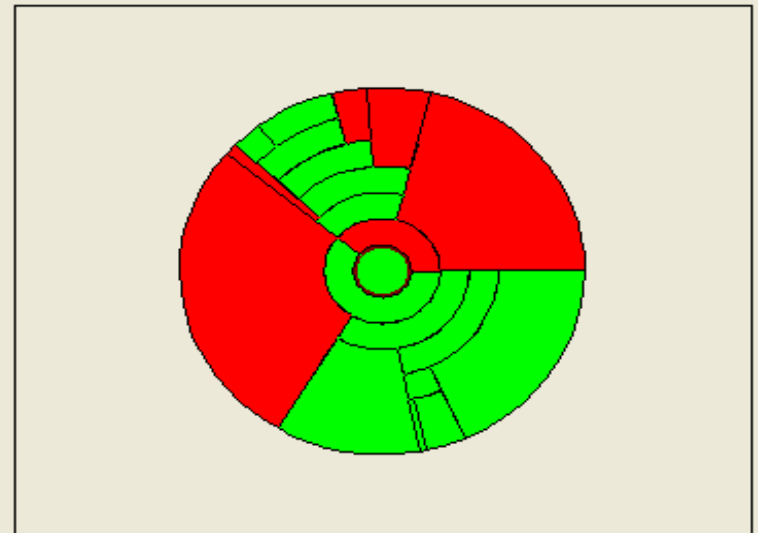


Entropy Decision Tree Summary

Model | All | Summary | Tree Ring | Table | Plot | Score | Log | Notes |

PRIORS	SOURCE	STAT	CHURN	==> 1	==> 0	
NOT APPLIED	TRAIN	N	1	447	8	*
NOT APPLIED	TRAIN	N	0	1374	949	*
NOT APPLIED	TRAIN	N	+	1821	957	*
NOT APPLIED	TRAIN	Row%	1	98	2	*
NOT APPLIED	TRAIN	Row%	0	59	41	*
NOT APPLIED	TRAIN	Row%	+	66	34	*
NOT APPLIED	TRAIN	Col%	1	25	1	*
NOT APPLIED	TRAIN	Col%	0	75	99	*
NOT APPLIED	TRAIN	Col%	+	100	100	*
NOT APPLIED	TRAIN	%	1	16	0	*
NOT APPLIED	TRAIN	%	0	49	34	*
NOT APPLIED	TRAIN	%	+	66	34	*

Leaves	Training	Validation
11	0.6779	0.6396
12	0.6779	0.6396
13	0.6779	0.6396
14	0.6779	0.6396
15	0.6779	0.6396
16	0.6920	0.6358
17	0.6960	0.6295
18	0.7101	0.6258
19	0.7112	0.6168
20	0.7115	0.6071



Important variables of Decision Tree

Decision tree produces **11** leaves and splits according to **9** important variables.

Model | All | Summary | Tree Ring | Table | Plot | Score | Log | Notes |

☒ Export role as indicated in the table

Name	Importance	Role	Rules	Variable Label
HSGPA	1.0000	input	2	Average High School Grade
NS_GPA	0.9479	input	1	
APEXAM	0.6426	input	1	# of Advanced Placement Exams Taken
HPW0208	0.3578	input	1	Student Clubs or Groups
EXT_CAREE	0.3216	input	1	EXT_CAREER
ACT0201	0.2653	input	1	Attended Religious Services
FUTACT19	0.2503	input	1	Strengthen Religious Beliefs/Convictions
OPTQ0219	0.2486	input	1	Factors Influencing Leaving University
ACT0205	0.1378	input	1	Studied with Other Students
NATENGSP	0.0000	rejected	0	Student Native English Speaker?
SEX	0.0000	rejected	0	Student's Gender

Data | Variable Selection | New Variables | Code

11 Rules found by Decision Tree

Rule1 : If Average High School Grade is lower than B+
and Natural Science GPA < 2.85

Then

N: 586

Drop-out : 100.0%

Retain: 0.0%

Rule2: If Choose Number of Advanced Placement Exams
Taken equals None
and Natural Science GPA ≥ 2.85

Then

N: 760

Drop-out: 53.0%

Retain: 47.0%

Rule11:

Drop-out Student Distribution in each Rule

Rule	Drop-out Student	Drop-out % in this Rule	Drop-out % in Overall Data	Total Student in this Rule
Rule 1	482	100.0%	40.88%	482
Rule 2	345	18.77%	29.26%	1838
Rule 3	26	100.0%	2.21%	26
Rule 4	7	0.51%	0.59%	1378
Rule 5	67	24.45%	5.68%	274
Rule 6	159	10.64%	13.49%	1495
Rule 7	27	16.46%	2.29%	164
Rule 8	11	18.64%	0.93%	59
Rule 9	3	0.84%	0.25%	359
Rule 10	30	4.41%	2.54%	681
Rule 11	22	11.64%	1.87%	189
Total	1179		16.98%	6945

Retain Student Distribution in each Rule

Rule	Retain Student	Retain % in this Rule	Retain % in Overall Data	Total Student in this rule
Rule 1	0	0.00%	0.00%	482
Rule 2	1493	81.23%	25.89%	1838
Rule 3	0	0.00%	0.00%	26
Rule 4	1371	99.49%	23.78%	1378
Rule 5	207	75.55%	3.59%	274
Rule 6	1336	89.36%	23.17%	1495
Rule 7	137	83.54%	2.38%	164
Rule 8	48	81.36%	0.83%	59
Rule 9	356	99.16%	6.17%	359
Rule 10	651	95.59%	11.29%	681
Rule 11	167	88.36%	2.90%	189
Total	5766		83.02%	6945

What is Hit Rate ?

- Definition: Drop out Hit Rate

Predicted Value

True
Value

	Drop-out	Retain	Total
Drop-out	N_{11}	N_{12}	$N_{1.}$
Retain	N_{21}	N_{22}	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	N

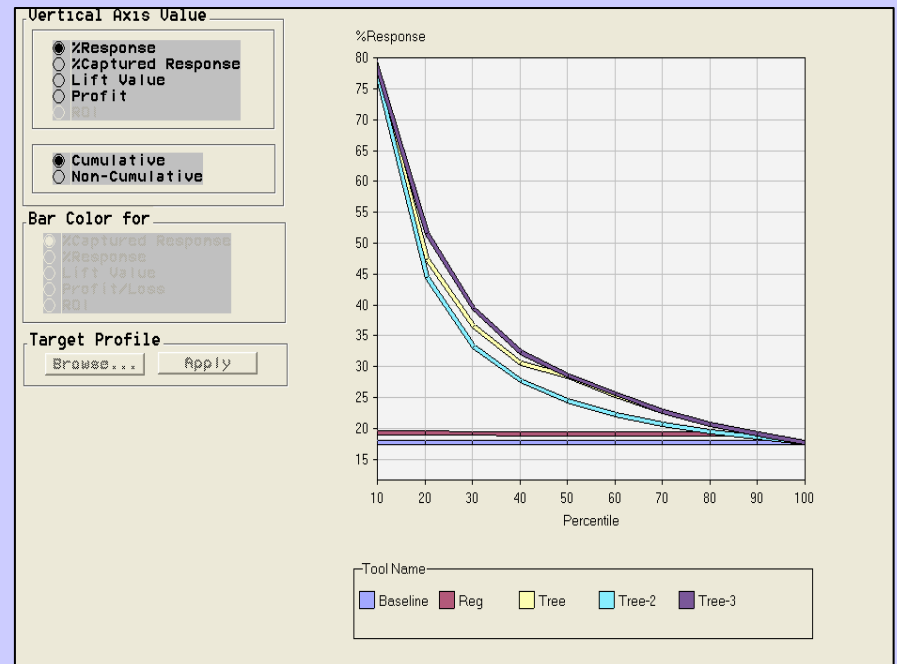
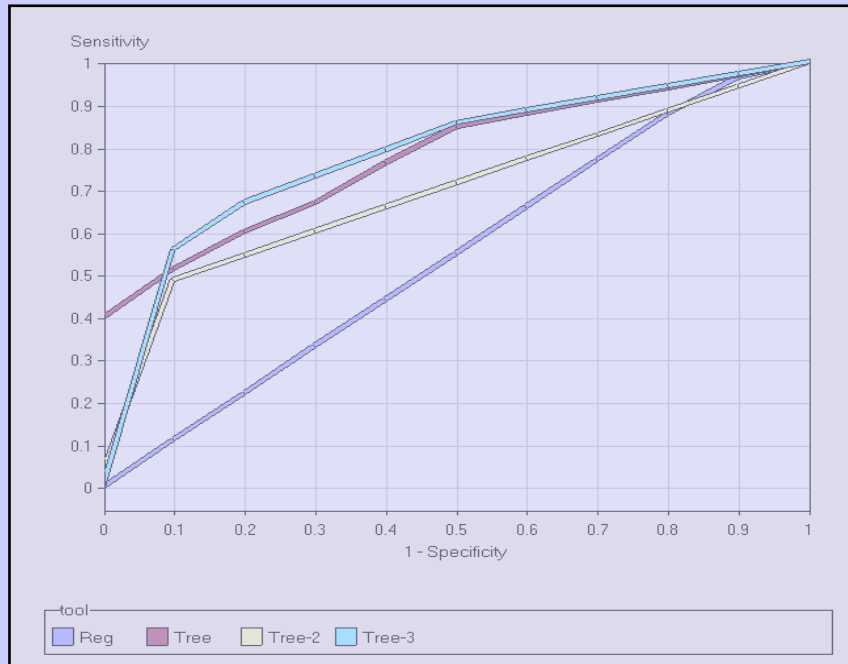
$$\text{Hit Rate} = N_{11} / N_{1.}$$

- Hit Rate is a powerful measurement in model fitting.
- Hit Rate represents the prediction accuracy in our retention model.

Model Comparison by Hit Rate

Model	Model Description	Hit % in Training data	Hit % in Validation data	Hit % in Testing data
Decision Tree 1	Entropy split criterion	98%	92%	90%
Decision Tree 2	Chi-square split criterion	83%	78%	80%
Decision Tree 3	Gini Index split criterion	90%	84%	82%
Logistic Regression	Stepwise regression	87%	84%	83%

Model Assessment



Comparing ROC for all models, **Entropy Tree model** has higher performance than others in validation data set.

By counting the percentage of actual responses in each bin, and comparing the overall performance, the **Entropy Tree model** perform the best.

Conclusion

- Data Mining is a powerful tool for analyzing student retention.
- These model can identify more than **90%** students who dropped out in the test data.
- These models can be used to predict students retention started from freshman year.

Conclusion (continued)

- The quality of student learning experience (such as High School GPA, SAT) is the most significant factor in retention rate.
- The number of taken advance placement exams play a important role in retention rate.
- Student retention rate is also affected by recognition of student sociability.
- Career motivation also affects retention rate

Strategies of Student Retention

- To provide a higher level of learning support (especially Science and Math) to minimize drop-out rate
- Develop retention program to form a more comprehensive, supportive environment
- Enhance the communication between the students and faculty
- Keep student study interest and motivation alive

Further Research

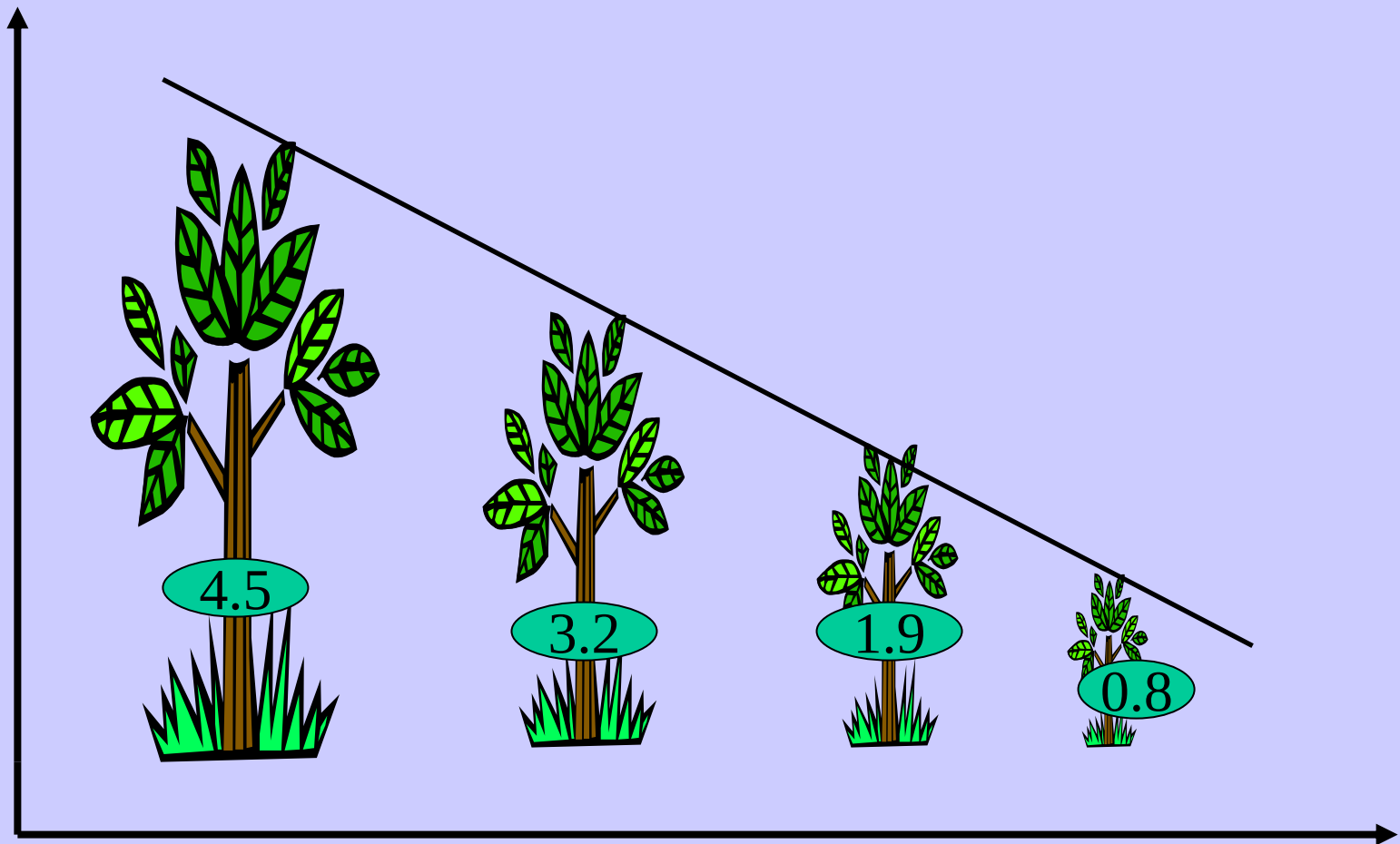
Our approach is not a solution to all the problems that exist with retention

- Enlarge the data source to look for other significant factors
- Determine the most appropriate threshold for Decision tree model
- Check accuracy of predictions on new data source
- Develop integrated models of student retention

.....

Case Study 3:

Rules Based Regression with Tree



Background

- Mortality tables currently in use have used a variety of mathematical techniques to generate mortality rates for advanced ages as a smooth extension of the patterns of mortality rates of septuagenarians and octogenarians. The most popular model is the Gompertz law that assumes that the force of mortality continues to increase exponentially with age is used.

Background (cont.)

- Other existing studies on advanced-age mortality
 - Smooth extension of the patterns
 - Families of curves - Gompertz law, etc.
 - All these approaches aim at explaining the age pattern of mortality.

Background (cont.)

- Mortality distribution varies among seniors with different backgrounds
- Variables considered:
 - Continuous Variable: Age
 - Categorical Variables: Gender, Pay Type, Annularity Size, Collar, Union, Participation Status

Note: The data availability limited the variables considered in this study

Rule-Based Regression

Step 1: Fit a regression model with only age and age*age

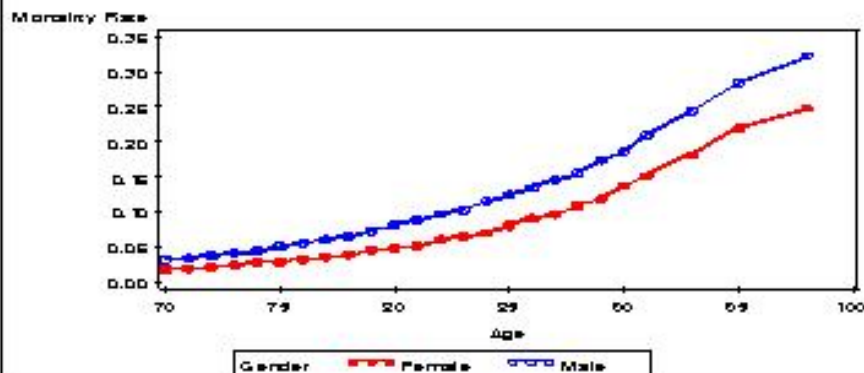
Step 2: Fit a CHAID like tree (Kass, 1980) with the residuals from the previous regression model

Step 3: Partition the data into six segments

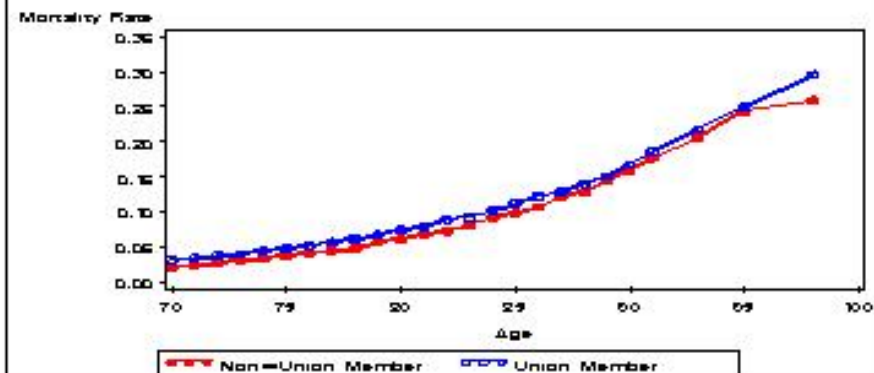
Step 4: Fit a regression model in each segment

Figure 1 Advanced Age Mortality Trend

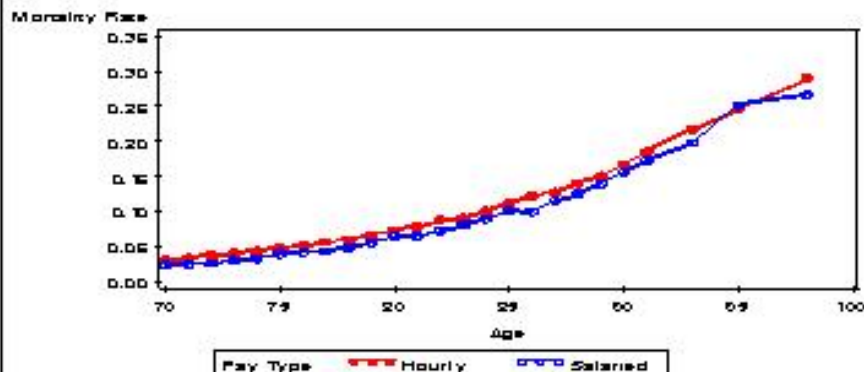
Gender: Female v.s. Male



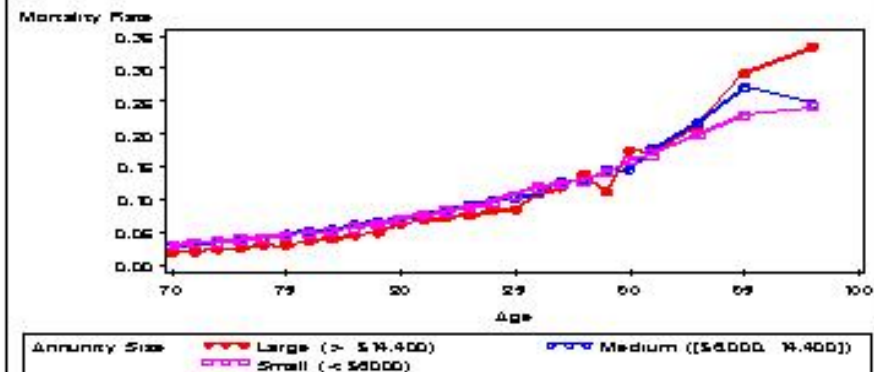
Union v.s. Non-Union



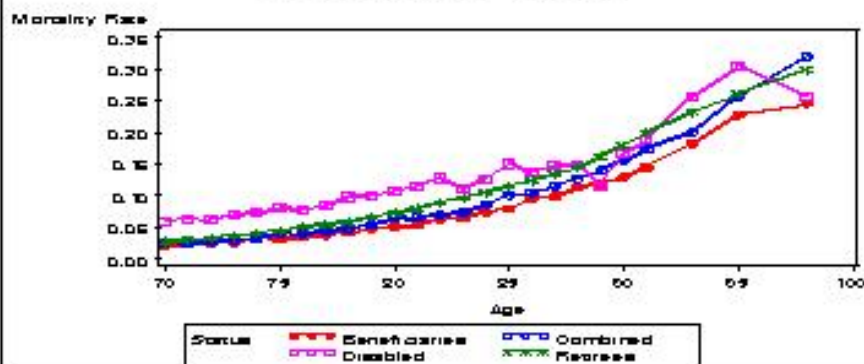
Pay Type: Hourly v.s. Salaried



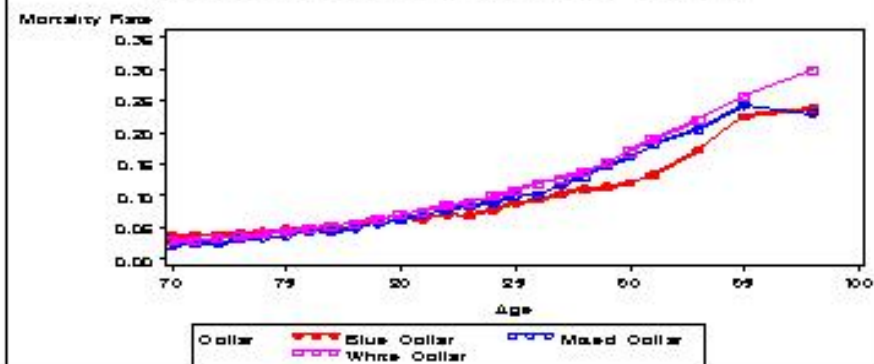
Annuity Size: Large, Medium, or Small



Participation Status



Collar: Blue, White, and Mixed



Rule-Based Regression (cont.)

Step 1: Fit a regression model with only age and age*age

- Age is the most important variable in study mortality rate
- The growth trend for mortality rate is not linear
- Tree is not good at handle “continuous variable”

Rule-Based Regression (cont.)

Step 2: Fit a CHAID like tree (Kass, 1980)
with the residuals from the previous
regression model

Rule-Based Regression (cont.)

Step 3: Partition the data into six segments

- Participation status is “Beneficiaries”
- Participation status is “Combined”
- Participation status is “Disabled”
- Participation status is “Employee”
- Participation status is “Retiree” and Gender is “Female”
- Participation status is “Retiree” and Gender is “Male”

Rule-Based Regression (cont.)

Step 4: Fit a regression model in each segment

- The data available in each segment are different
- Variables selected in the regression model in each segment can be very different
- Tree diagram provides information on selecting variables

Model for “Beneficiaries” Segment

- Fit a regression with variables age, age*age, gender, “annuity size”, gender*“annuity size”, union, and “pay type”

Figure 2 Tree Diagram for Beneficiaries

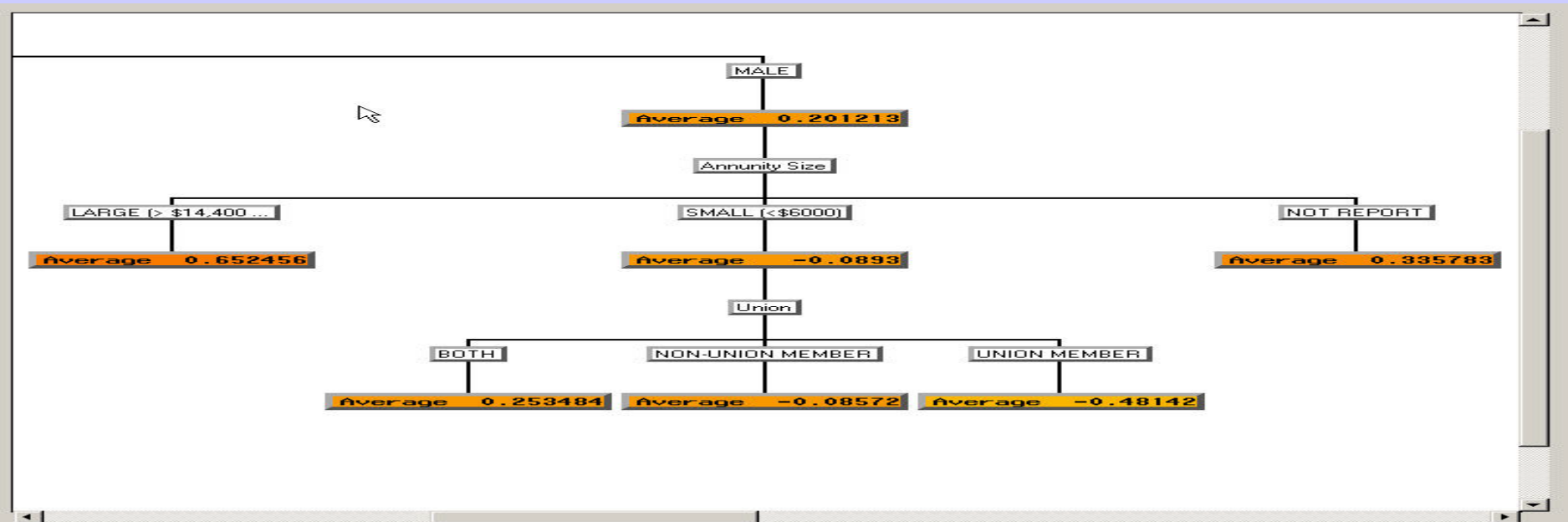
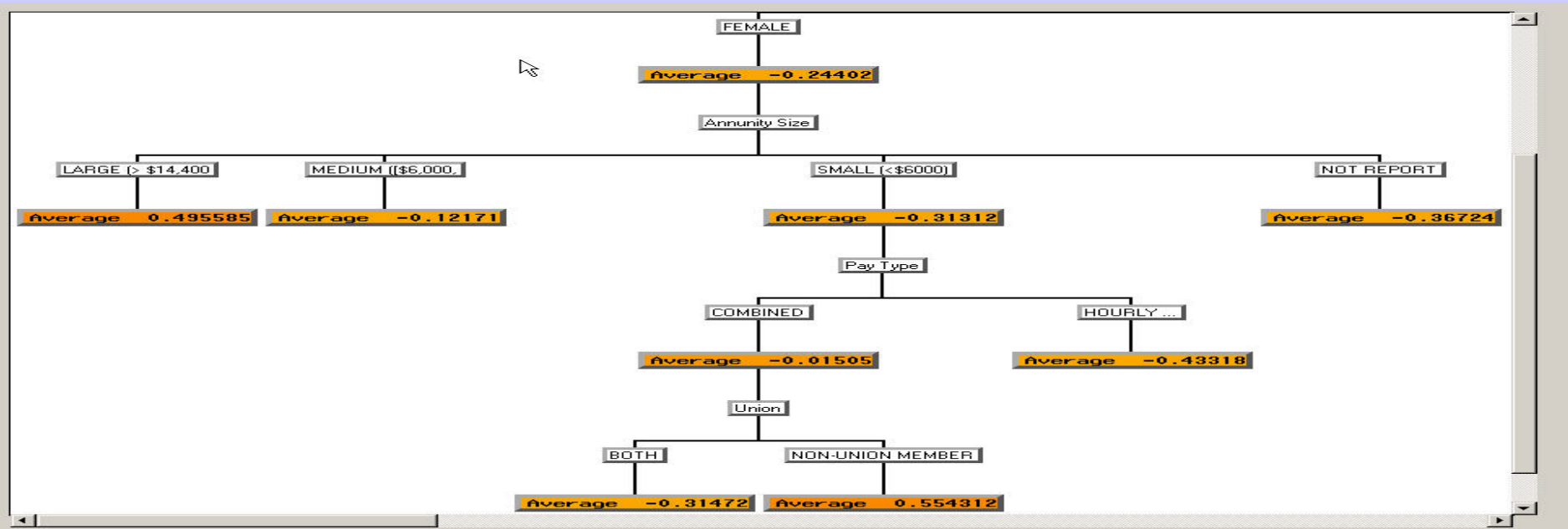
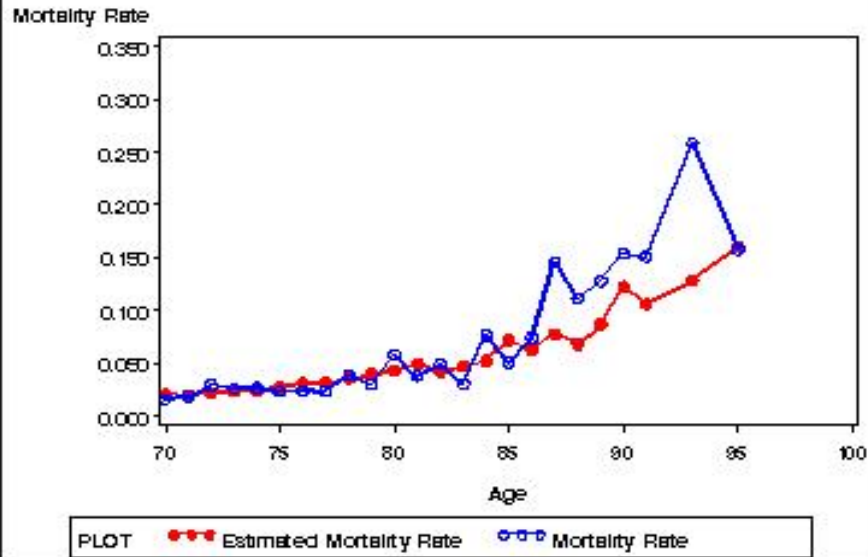
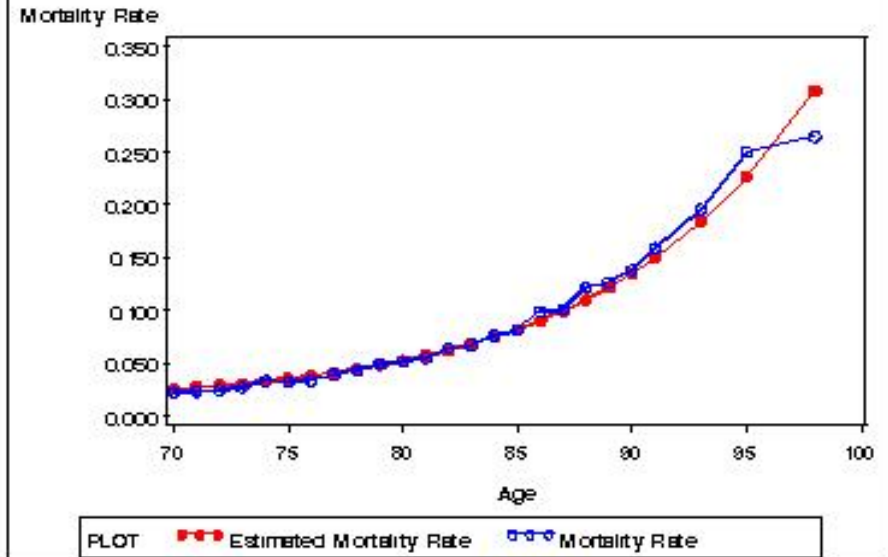


Figure 3 Mortality Model for **Female** (Beneficiaries)

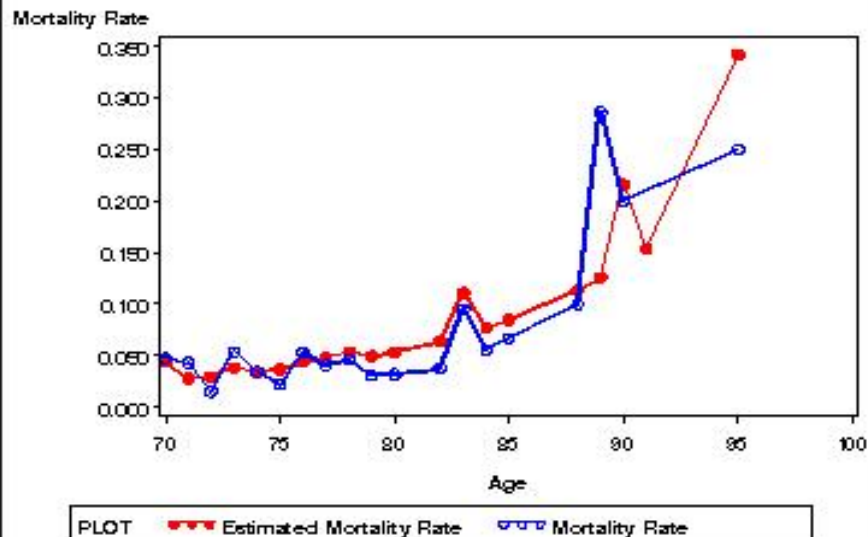
Female: Medium Annuity Size



Female: Annuity Size Not Report



Female: Large Annuity Size



Female: Small Annuity Size

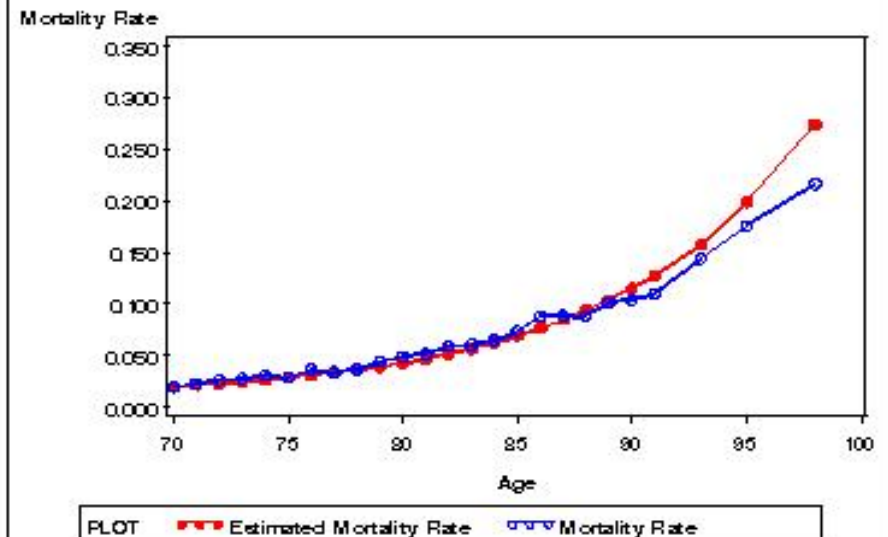
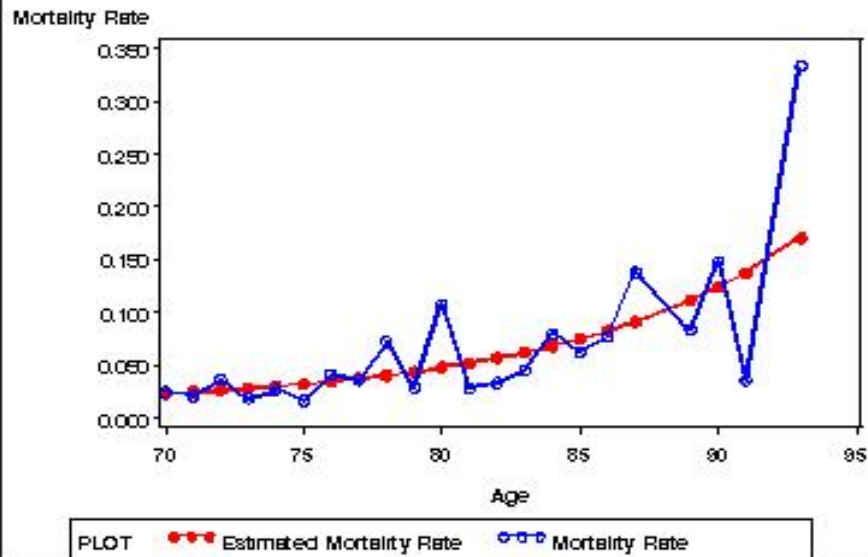
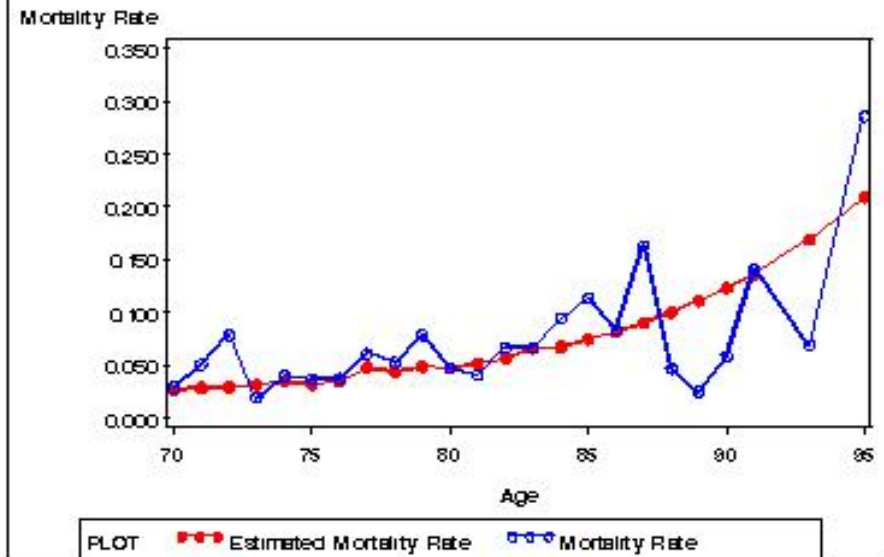


Figure 4 Mortality Model for **Male** (Beneficiaries)

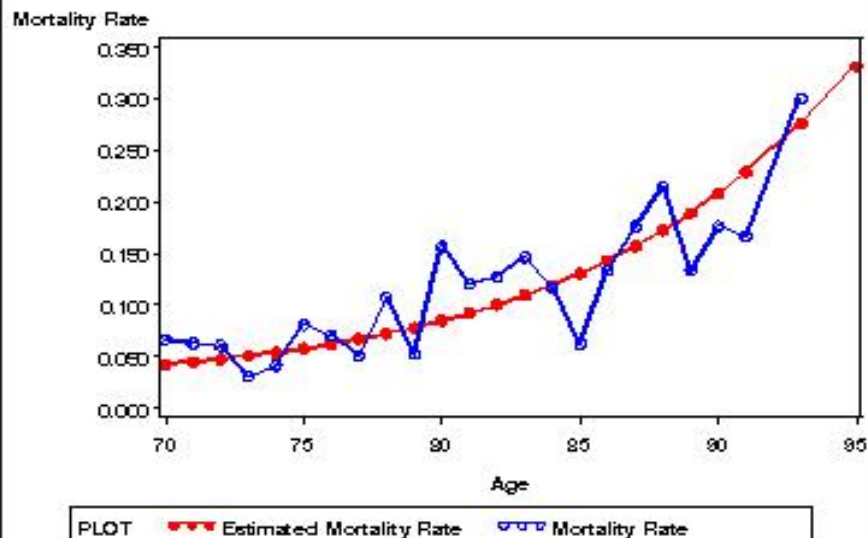
Male (Small Annuity): Union Member



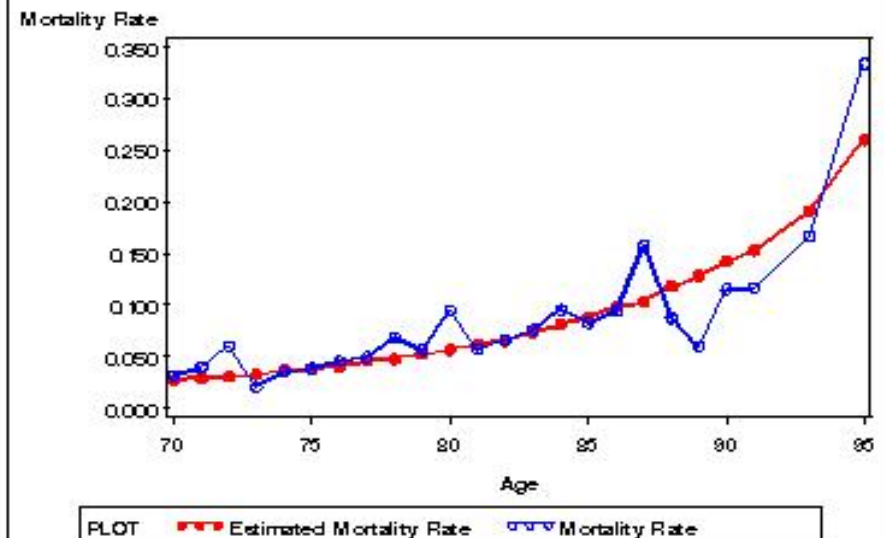
Male (Small Annuity): Non—Union Member



Male (Small Annuity): Both



Male (Small Annuity)



Model for “Disabled” Segment

- Fit a regression with variables age, age*age, annuity size, gender, and gender*annuity size

Figure 5

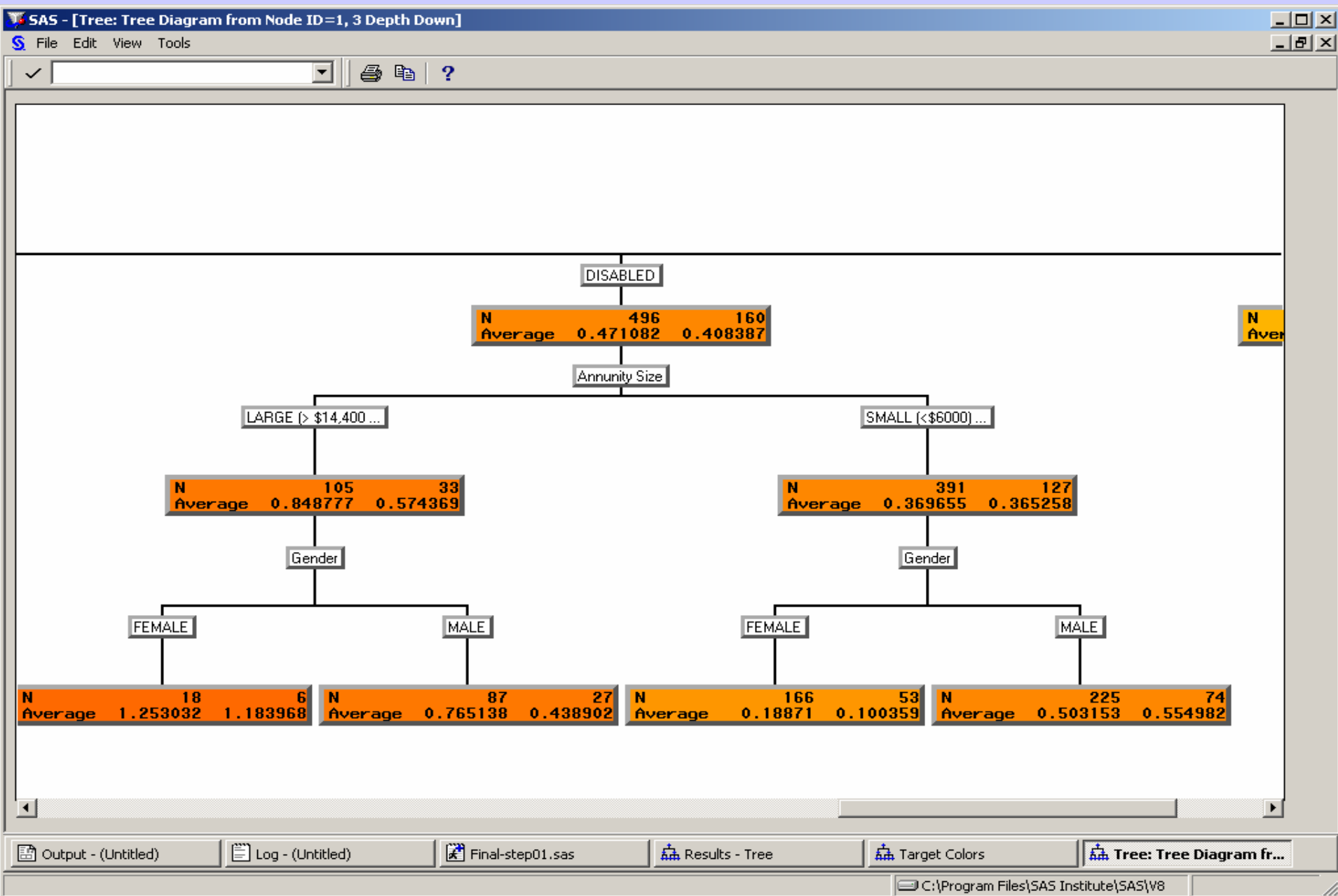
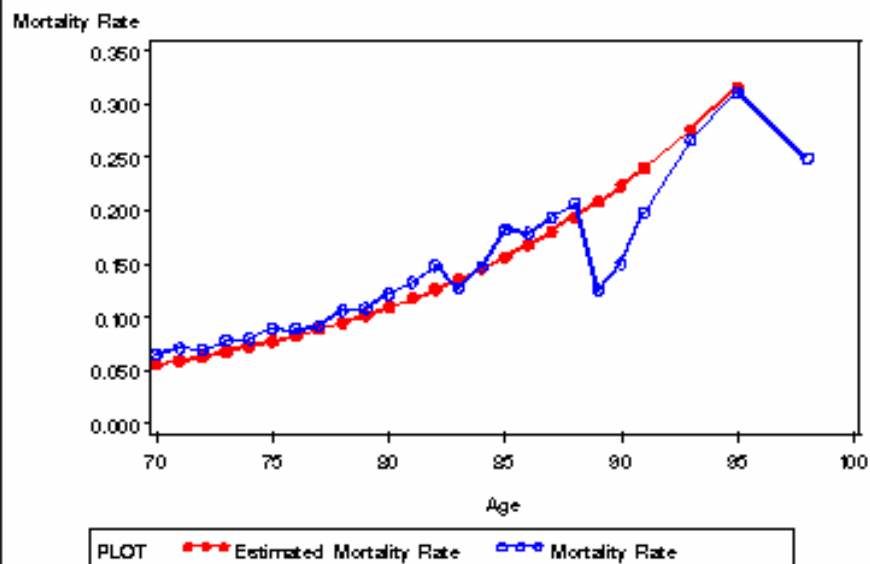
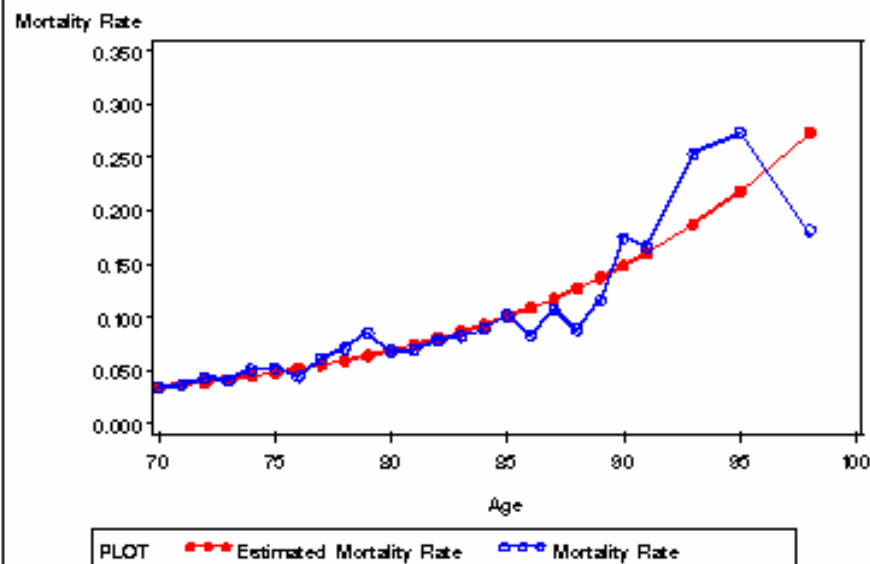


Figure 6 Mortality Trend (Disabled)

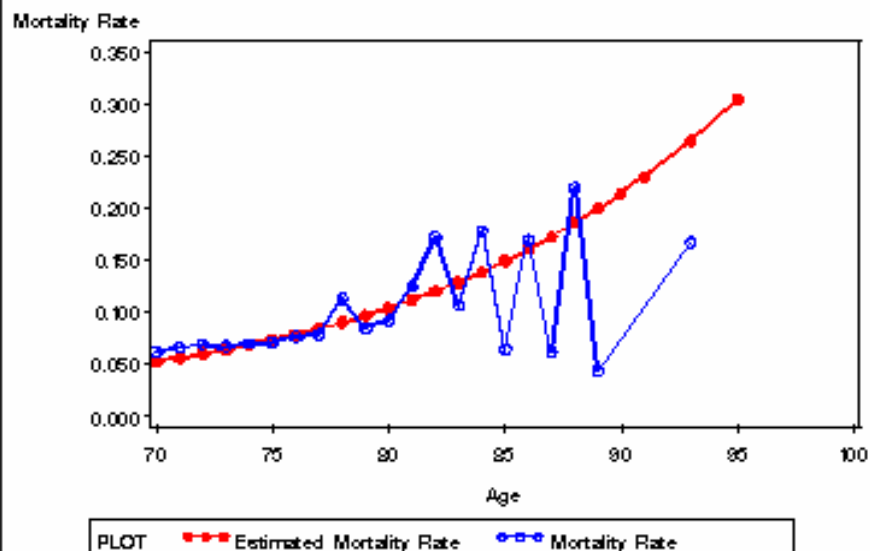
Male: Small Annuity Size



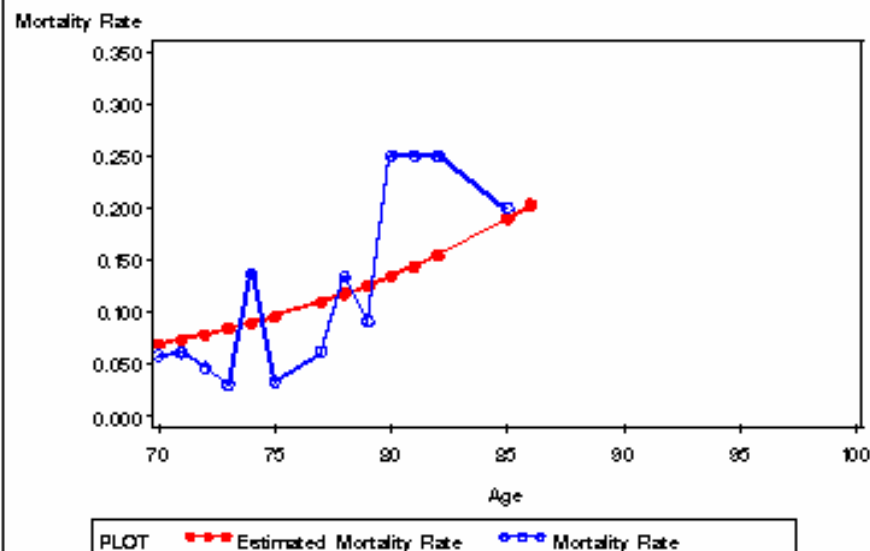
Female: Small Annuity Size



Male: Large and Midium Annunity



Female: Large and Medium Annunity



Model for “Combined” Segment

- Fit a regression with age, age*age, annuity size, and gender

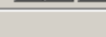
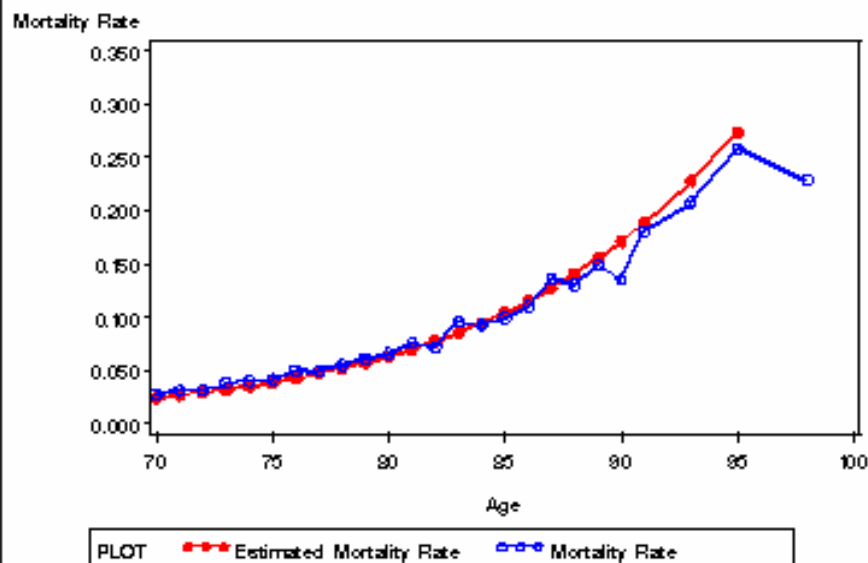
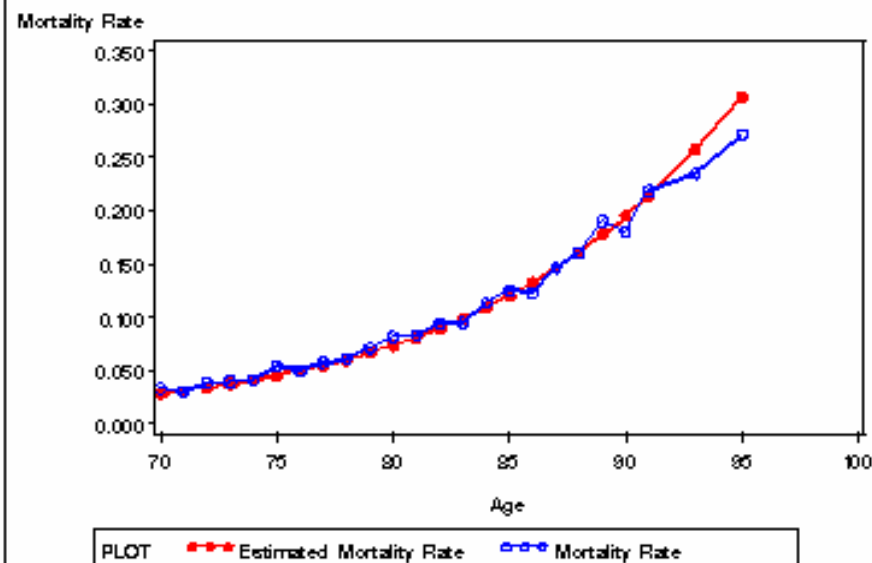


Figure 4 Mortality Trend (Combined)

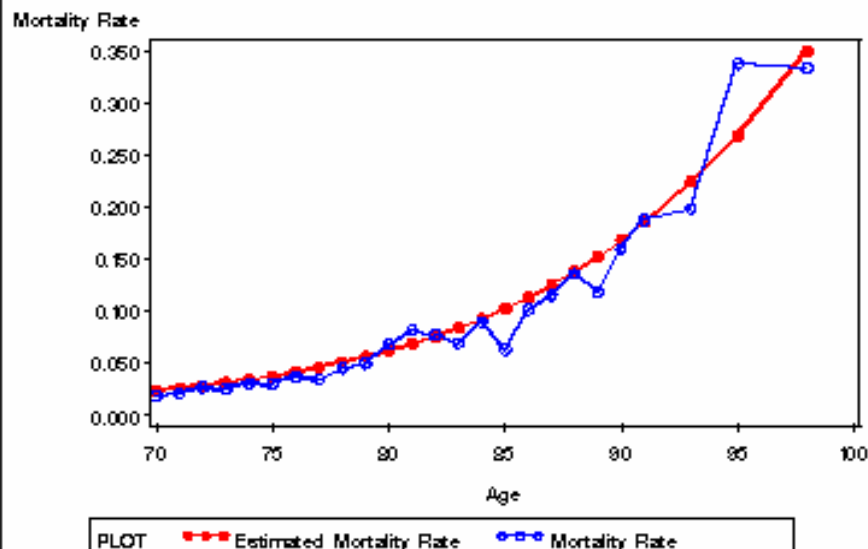
Male: Medium Annuity Size



Male: Annuity Size Not Report



Male: Large Annuity Size



Male: Small Annuity Size

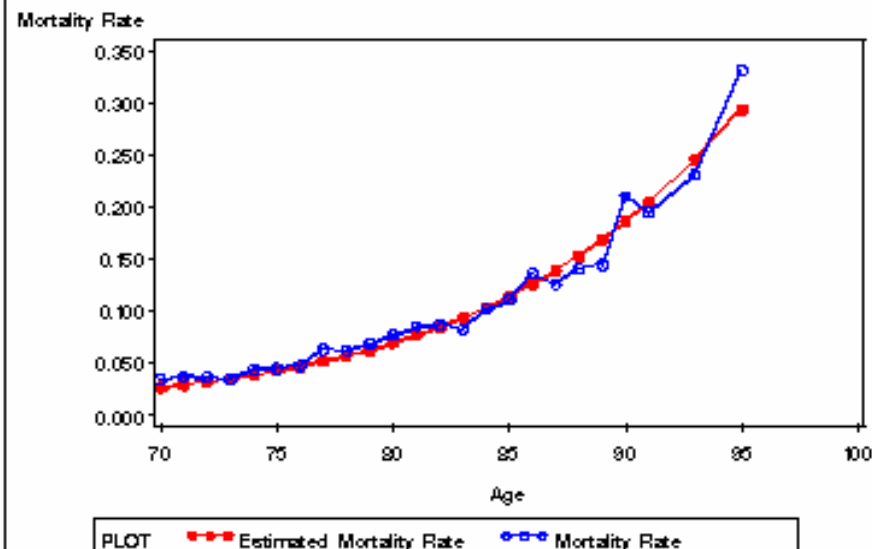
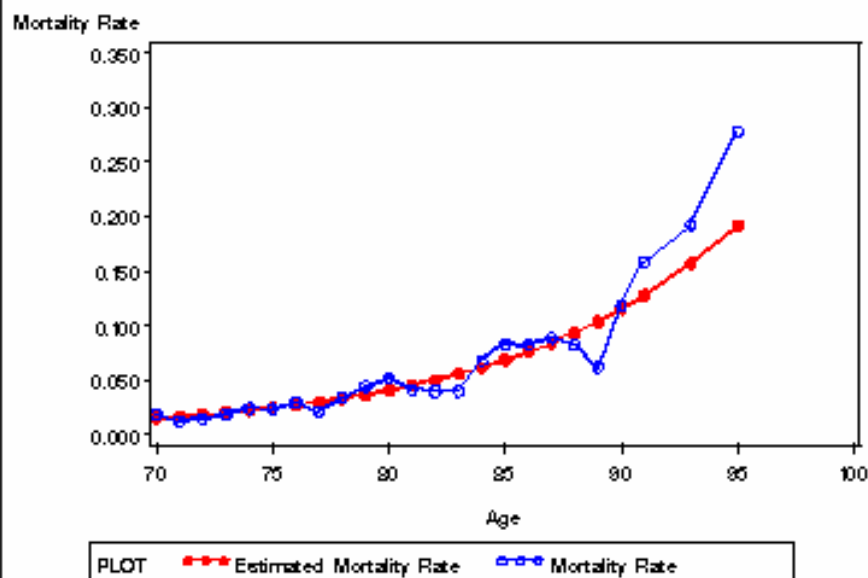
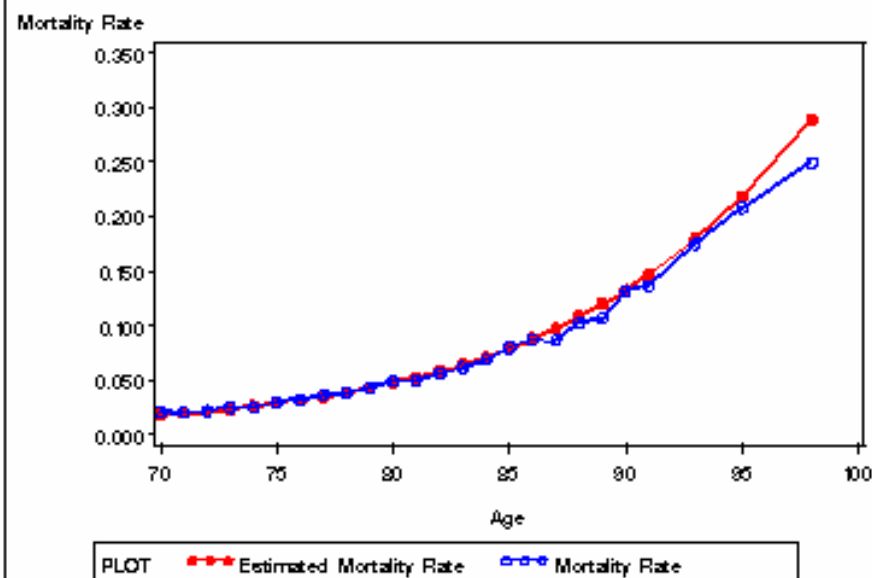


Figure 5 Mortality Trend (Combined)

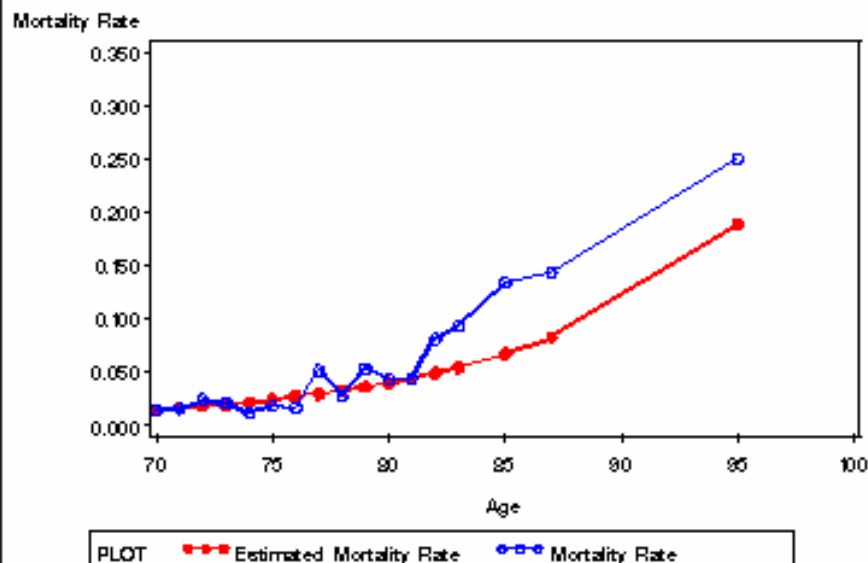
Female: Medium Annunity Size



Female: Annunity Size Not Report



Female: Large Annunity Size



Female: Small Annunity Size

