**ORIE 474 — Exam 2**

**Professor David Ruppert**

**In-Class**

**Mon, Oct 28, 2002, 1:25 – 2:15, Thurston 205**

The in-class exam is closed book. You may use three sheets (six pages) of notes. Time yourself carefully. Questions 1–5 require only short answers.

You may keep this exam. Just turn in your blue book or books. Indicate on the cover of each blue book which problems are contained inside and in what order. Also indicate if problems are split. For example, write 1-3-2-4-6-5-1 on the cover if you did the problems in that order, with problem 1 started first but finished at the end, and all problems done in one blue book.

The point value of each subquestion is in square brackets, for example, [8].

During the in-class and take-home exams, you may not communicate about any aspect of OR474 with anyone except the TA and instructor. During the take-home exam, the TA is allowed to answer your questions about the use of SAS EM but not about conceptual issues.

Good luck.

1. What is the difference between a regression tree and a classification tree? [8]

2. Why does one grow a large decision tree by node splitting and then prune it to get a smaller tree? Why not just grow a right-sized tree from the start and avoid the need to prune? Also, why not just use the largest possible tree? [8]

3. Give two advantages of regression trees compared to linear or logistic regression. State a disadvantage of decision trees compared to linear or logistic regression. [8]

4. What is meant by a “weakest-link” and how does “weakest-link pruning” work? [8]

5. In his guest lecture, Kurt Holstein discussed how Rosetta Marketing uses NAB segmentation. What does the acronym NAB mean? How does Rosetta use NAB segmentation to improve the marketing of its clients? [8]

6. A stratified sample was taken of a population. The response rate, that is, the probability that \( Y = 1 \), is 5% in the population, but was fixed by stratification at 40% of the sample. The sample proportions were:

<table>
<thead>
<tr>
<th>( Y )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.25</td>
<td>.05</td>
</tr>
<tr>
<td>2</td>
<td>.25</td>
<td>.1</td>
</tr>
<tr>
<td>( X )</td>
<td>.05</td>
<td>.15</td>
</tr>
<tr>
<td>4</td>
<td>.05</td>
<td>.1</td>
</tr>
</tbody>
</table>

Estimate \( P(Y = 1 \text{ and } X = 2) \), \( P(X = 2) \), and \( P(Y = 1|X = 2) \). [10]

**Take-home**

**Due 5pm, Wed, Oct 30, 2002**

The data sets for the the take-home segment are on the course web site under “Course Data Sets.” There are two data sets, “TargetKnown” and “Unclassified.” “TargetKnown” has ten variables, a binary target and nine predictor variables. “Unclassified” has the nine predictor variables, but target is missing in “Unclassified.” In the population, the response rate is only 15%. However, for this assignment you should use a stratified sample containing all responders and an equal number of nonresponders.

Using “TargetKnown,” find a good logistic regression model and a good decision tree model for predicting target. Write a short report describing your models and discussing which one is best. Use your favorite model to classify the observations in the data set “Unclassified.” Your report can be typed or hand-written. It does not need to be in a blue exam book. [50]

**Announcements**

Please make certain that you are on the class electronic mailing list. A message about the course project was sent on Oct 27. If you did not receive it, then you are not on the list.

On Wednesday I will be in my office from 4:30 to 5pm. If you are locked out of the building and cannot turn in your exam, you can call me at 5-9136. There is a free phone outside the main entrance to Rhodes — look for it on the wall to your left.