MIS 6334 Advanced BI
Project 1: Data and Model Analytics using SAS Enterprise Miner

Project halfway check: Week of 9/21
Report due (before class starts): 10/5 for Section 003; 10/6 for Section 501; 10/7 for Section 001

In this group project you will walk through the whole SEMMA process in SAS Enterprise Miner with the goal of improving your prediction results. You will start with a real dataset and finish with managerial-relevant findings.

Submit one report per group in eLearning under Assignments -- Project 1. Due time is right before class on the due date. eLearning will stop accepting submissions right at the due time, and **no late submission is acceptable**. The report should be **a single Microsoft Word document with your group number and all group member names.** The report should addresses all following project components. Messy reports will receive penalty. Many questions are open-ended, and within-group discussions on these questions are strongly encouraged.

A pair of datasets are provided: a training dataset (expediatrain.sas7bdat) and an evaluation dataset (expediaevaluation.sas7bdat). The end of this project handout provides data description.

Part I. Basic Data Preprocessing

1. Use dataset expediatrain.sas7bdat. Provide a brief summary of the variables in the data set (you can choose to use StatExplore for that purpose).
2. Explore the statistical properties of the variables in the input data set. The results that are generated in this step will give you an idea of which variables are most useful in predicting the target response. Unless you see anything interesting, no need to report the details of this step.
3. Check the **Class Variable Summary Statistics** and the **Interval Variable Summary Statistics** sections of the output.
    a. Are there any missing values for any of the variables? Use **imputation** to fill in all missing data (describe how you did imputation in the report).
4. Partition dataset expediatrain.sas7bdat. Use 55% of the data for training and 45% for validation.

Part II. Building Decision Trees

Now that you have familiarized yourself with the input data, it is time to build predictive models. First try nonparametric decision trees. In particular, perform the following tasks:

1. Enable SAS Enterprise Miner to automatically train a full decision tree and to automatically prune the tree to an optimal size. When training the tree, you select split rules at each step to maximize the split decision logworth. Split decision logworth is a statistic that measures the effectiveness of a particular split decision at differentiating values of the target variable. For more information about logworth, see SAS Enterprise Miner Help. Report the results.
2. Then, interactively train a decision tree. At each step, you select from a list of candidate rules to define the split rule that you deem to be the best. Report the results.

Part III. Building Neural Networks and a Regression Model

1. Transform input variables to make the usual assumptions of regression more appropriate for the input data. Explain the transformations you did.
2. Model the input data using logistic regression. Report the results.

3. Model the input data using neural networks, which are more flexible than logistic regression (and more complicated). Report the results.

Part IV. Model Comparison and Champion Model Evaluation

1. Compare the above four models you tried, and select a champion model. When evaluating the model performance, try to use confusion matrix as the main evaluation criterion. And let's use a cost 5 for misclassifying 1 as 0, and a cost of 1 for misclassifying 0 as 1.
2. Score the new evaluation dataset -- expediaevaluation.sas7bdat -- using the champion model.

(I expect you to reach at least the end of Part IV before our halfway-check meeting.)

Part V. Improve Your Model Performance

This part is an **open question**. Suppose you are not happy with the performance of your champion model from Part IV. Try to use what you learned in our ABI class to further improve your model performance. Below are a number of possible ways:

- Should we do imputation on all variables with missing data? And use imputed data for all classifiers?
- Skewed data? Data with high variance?
- Do we need all 40 attributes for prediction? If no, how about removing some variables?
- Is the dataset too large in terms of number of records? If you think so, how about sampling for a smaller size? Will this actually help with the prediction performance?
- Can ensemble help? How will you do it?

Please pick at least 3 ways from the above list (or any idea you have that is not on the above list) and see whether it improves the performance of your project. In doing so, please create new flows/nodes in your diagram (and don't remove or modify the four classifiers already done in Parts II and III). Describe the improvements you tried and the results in the report, if any.

Part VI. Summary

1. Include the complete final diagram you get. Make sure it is legible, and if needed use several pages in print.
2. Summarize what you learned from this project. Be concise. Again, the whole report should be nicely presented in a single Microsoft Word document (submitted through eLearning).

Need Help?

- Review your knowledge of SAS Enterprise Miner from last semester's MIS 6324 course.
- Utilize the detailed AAEM e-book for technical issues:
- Use help menu in SAS Enterprise Miner (perhaps the best help files among BA tools).
- Come to TA or the professor's office hours as a last resort.

What Will the Professor Ask at Project 1 Halfway Check?

- Have you done basic data preprocessing? Any findings/thoughts?
- Any plan on improving the classifier performance (that is beyond the four required ones)?
- This will be verbal and quick. Try to impress your professor so your group can present on this project (rather than the considerably harder Project 2).

Appendix -- Data from “An Empirical Analysis of the Value of Complete Information for eCRM Models”, Padmanabhan, B., Z. Zheng, and S. Kimbrough. *MIS Quarterly*, 30(2), 2006. Variables 1-15 are site-centric variables; 16-40 are additional user-centric variables and the last is the dependent variable. At the end of some variables, “g” means all sites (global) and “l” means only this site (local); “c” means only the current session and “h” means all past sessions.

No.	Variable	Description
1	gender	“1”—Male, “0” – Female
2	age	Age of the user
3	income	Income of the user
4	edu	“0” – high school or less, “1”-- college, “2” – post college
5	hhsize	Size of house hold
6	child	“1” – have, “0” – not have
7	booklh	No. of bookings the user made at this site in the past
8	sesslh	No. of sessions to this site so far
9	minutelh	Time spent in this site so far in minutes
10	hpsesslh	Average hits per session to this site
11	mpsesslh	Average time spent per sessions to this site
12	booklc	Dummy variable, indicating if the user has booked at this site up to this point in the current session
13	httlc	No. of hits to this site up to this point in this session
14	minutelc	Time spent up to this point in this session
15	weekend	Indicating if this session occurs on weekend
16	bookgh	No. of past bookings of all sites so far
17	sespsite	Average sessions per site so far
18	sessgh	Total no. of sessions visited of all sites so far
19	minutegh	Total minutes of all sites
20	hpsessgh	Average hits per session
21	mpsessgh	Average minute per session
22	awareset	Total no. of unique shopping sites visited
23	basket	Average no. of shopping sites visited per session
24	single	Percentage of single-site sessions
25	booksh	Percentage of total bookings are to this site
26	hitsh	Percentage of total hits are to this site
27	sessh	Percentage of total sessions are to this site
28	minutesh	Percentage of total minutes are to this site
29	entrate	No. of sessions start with this site/total sessions of this site
30	peakrate	No. of sessions the user spend the most time within this site/total sessions of this site
31	exitrate	No. of sessions end with this site/total sessions of this site
32	SErate	No. of sessions coming from search engines/total sessions of this site
33	*bookgc*	Binary variable, indicating if this user has booked at any sites up to this point in the current session
34	*hitgc*	Total hits of all sites in the current session
35	*basketgc*	No. of shopping sites in this session
36	*minutegc*	Time spent of all sites in this session
37	*SEgc*	Indicating if this session uses search engines
38	*path*	Indicating if this site is an entry/peak
39	*hitshc*	Hits to this site/ hits to all sites in this session
40	*minutshc*	Minutes to this site/total minutes in this session
41	bookfut	Binary dependent variable, indicating if this user is going to book in the remainder of the session (after the clipping point)