DUKE UNIVERSITY
The Fuqua School of Business

MGRECON 491 Data Mining — Gallant
Homework 2 — Term 1
Due October 3, 2008

The goal of this assignment is to determine a model that can be used by the same Czech bank considered in Homework 1 to score customers for the purpose of soliciting loan applications. The model will only use transactions data for two reasons: (1) We learned in Homework 1 that transactions data probably contain the best information regarding loan defaults of the data currently available to the bank. (2) While a teller or bank representative is talking to a customer, transactions data is all that will be available on the computer monitor.

This task has much in common with the Charity Case discussed in lecture. The main difference is that the number of features is much smaller (about 30 instead of 400) so that exhaustive model screening techniques such as best subset regression using Mallows C_p are feasible after, perhaps, deletion of some features on a logical basis.

The Excel workbook trans.xls, which can be downloaded from the course web site, contains data on customer transactions and profits and losses from loans for the Czech bank. The workbook has two worksheets: tran_data and tran_score. The customers in tran_data are loan customers; those in tran_score are not. You will use tran_data to train and validate your model. The target variable in tran_data is the one labeled loanprofit. Then you will apply your model to the accounts in tran_score to predict loan profits for each of the bank's customers who do not already have loans. Bank staff will be asked to solicit loan applications from customers with high predicted profits when opportunities arise. Offhand, one would think that any customer with positive predicted profit should be solicited. You may be able to come up with a better decision rule, but be prepared to defend its logic if you do.

To save you time I have added a partition variable named "partition" to the worksheet tran_data. You can add derived variables to worksheet tran_data and re-partition on the variable "partition" and thereby be able to accurately compare validation "RMS Error" from runs on an old partition to runs on a new partition. The data were properly randomized before assigning t's and v's using the procedure described in Homework 1. If you add a derived variable to tran_data then, of course, you must add the same variable to tran_score in order for the model you select to be able to score the data in tran_score. Here is the data dictionary:

account_id	account identifier and key for merges with other data bases
first	date of the first transaction in days since December 31, 1992
last	date of last transaction in days since December 31, 1992
numtrans	number of transactions of all types over the period first to last
avbal	average balance over the period in Czech currency
dposits	number of deposits of all types over the period
avdposit	average deposit of all types over the period
wdraws	number of withdrawals of all types over the period
avwdraw	average withdrawal of all types over the period
ccardwdraws	number of withdrawals by credit card (used as debit card)
avccardwdraw	average withdrawal by credit card
cashdposits	number of cash deposits
avcashdposit	average cash deposit
wiredposits	number wire transfer deposits
avwiredposit	average wire transfer deposit
cashwdraws	number of cash withdrawals
avcashwdraw	average cash withdrawal
wirewdraws	number of wire withdrawals
avwirewdraw	average wire withdrawal
inspments	number of automatic insurance payments
avinspment	average insurance payments
stmntpments	number of service charges for a statement
avstmntpment	average statement service charge
intcredits	number of interest credits
avintcredit	average interest credit
odraftfees	number of overdraft interest charges
avodraftfee	average overdraft interest charge
mortpments	number of automatic mortgage or rent payments
avmortpment	average number of mortgage or rent payments
pensndposits	number of electronic deposits from a retirement pension
avpensndposit	average pension deposit
loanpments	number of automatic loan payments
avloanpment	average loan payment
loanprofit	total interest for a good loan, write-off (1/2 loan amount) for a bad

Some things to notice about these variables when building a model are that there appear to be no loans to customers with pensions and that a customer with no loan does not have any automatic loan payments. It would seem that these variables would be therefore be useless to predict profits of customers without loans. In fact, XLMiner will not even allow you to use pensndposits or avpensndposit as an "Input Variable" because all entries in tran_data are zero. There may be other anomalies like this in the data that I did not notice. In all cases a total can be gotten by multiplying the average feature by the count for that feature; e.g. total deposits is deposits*avdposit. An obvious derived feature to consider is the duration of the customer relationship, which is computed as last minus first. Give some thought to curved relationships by either using derived variables such as numtrans squared as a feature or plotting profits against features. Another technique is to plot model residuals against each feature in hopes of finding curvature. In general, try all the ideas suggested by the analysis of the Charity Case in lecture.

The procedures that you will follow are basically the same as in Homework 1 with the main exception being that you will select all of your tools from the "Prediction" menu rather than the "Classification" menu.

Build the best prediction model that you can for at least these three tools: regression, regression tree, neural net. Best means that it generalizes well and has a low "RMS Error" in the validation sample.

After you have built the best model you can, score the data in worksheet trans_score. The scores will end up in a worksheet with a name of the form "TOOL_NewScore#." Sort the data descending on "Predicted Value." You can do this by "Copy," "Paste special," "values" to a new worksheet or sort the predictions in "TOOL_NewScore#" in place.

As with Homework 1, the best way to organize your thoughts for presenting your work is to think of yourself as a consultant advising this Czech bank. This carries with it the presumption that the technical level of your audience is lower than yours so you will have to explain the ideas behind the tools that you use and your results. Present your results as a PowerPoint presentation that takes no more than twenty minutes to present. Teams will present their work to the class. Here are the main points to address in your presentation.

- Present your results on identifying customers to solicit for loans.
    - Describe the tool that you used to obtain these these results.

  - Describe the entire analysis.
  - Include a lift chart for your tool and explain what information it conveys.
  - Be sure you discuss the notions of loss function and validation.
  - Make sure that a table of "RMS Error" in the validation sample for each model you fitted is included.
  - Make sure that your tool generalizes well and present the evidence that it does.

- Explain your search for derived variables.
  - Explain how you looked for them and present relevant evidence.
    * Explain how you looked for curvature and present your evidence that curvature is or is not present.
    * Explain how you looked for interactions and present your evidence that interactions are or are not present.

What to turn in: Please submit your PowerPoint presentation and your Excel workbooks on a CD. The Excel worksheet files trans01.xls, etc. will provide an audit trail where I can trace the results shown in your presentation, if necessary. Your presentation should describe: (a) the goal of the analysis and what you learned about the bank's consumer loan business from your analysis, (b) why you chose the prediction model that you did in terms that a layman would understand. Try to follow good principles of statistical presentation i.e., describe the data and try to make its message as clear as possible. Feel free to comment on what you feel might be the underlying causes for the patterns you have observed.

Your presentation should begin with one or two slides that highlight your key conclusions and recommendations especially the bottom-line prediction model. Next, it should include some slides that address the background questions addressed. Be sure to include a slide or two describing the data variables and the units in which they were measured. The presentation should include a few slides that illustrate your solution and the process followed in reaching it. You should describe exactly how your tool predicts a good loan prospect.