Wharton

OPIM-410/672: Decision Support Systems
Spring 2008

Data Mining Project Report: Check List

Student(s): _____
Submission Date: _____
Total (out of 50 marks) ___ / 50

The example in this tick sheet is under the assumption that you are mining a data set to determine which customers are likely to purchase caravan insurance. You will be mining a different data set, but caravan insurance is used as an example for the purposes of explanation.

Business Objective:	____ / 4
- Objective of the mining study specified.	
- Justification of the business value given.	
- Novelty - what has been done toward this objective in the past	
- Novelty - why my approach is novel	

Data Preparation:	____ / 5
- Possible sources of the data used specified (what internal and external databases may it have come from?)	
- Explained how the data has been prepared by others (e.g. was the data *pre-binned*), and what preparation steps you took (if any). Described the implications of this preparation (e.g. in terms of biases it may have introduced).	
- Specified, in an appendix, the independent (input / predictor) variables and their types (e.g. numeric, categorical, binary, etc.).	
- Mentioned what other variables you may have used (and their sources), had they been available.	
- Specified which variable is the dependent (output / predicted) variable.	
- Specified how you divided the data into training and test sets, and why this separation is important.	
- For numeric output data: Performed an exploratory data analysis: o produce scatter plots for every combination of input-variable with the output variable (e.g. if trying to predict income, always have income on Y axis, and then produce different scatter plots with each input variables on your X-axis: e.g. age, number of children, etc.) o for each input attribute: found min, max, average, standard deviation, 25th and 75th percentiles (e.g. use `Min()`, `Max()`, `Avg()`, `Stdev()`, `Percentile()` function in Excel), and found outliers.	

Data Mining:	**___ / 15**
- Used a data mining technique/algorithm (such as Decision Tree induction, rule induction, or genetic algorithm) to induce a set of predictive rules or formulae. Specified which data mining technique (algorithm) was used (and what parameter values were used, if applicable).	**/2**
- Described the relevance of the data mining technique employed, and the advantages and disadvantages as compared to other data mining techniques.	**/2**
- Described what other data mining techniques might have been applied to this data, and what benefits could have been achieved by using them.	**/1**
- Showed the results of the data mining – specific patterns discovered and interpreted (e.g. showed the decision tree induced, and listed the IF ... THEN ... rules induced; or for neural nets showed the prediction formula that was suggested by the algorithm). E.g. if trying to determine who purchases caravan insurance, did you discover that hedonists or people with fire insurance policies are more likely to buy caravan insurance ?	**/2**
- Showed a confusion matrix (giving the various error rates) for each model.	**/2**
- Construct a cost matrix with hypothetical costs for your various classification decisions (true and false positives, true and false negatives), being sure to give row and column headings meaningful to your application (like "Mail / Do not mail" rather than "positive / negative").	**/2**
- Using your cost matrix, calculate the threshold probability (i.e. 'cut-off response rate') which gives Expected Cost(Predict Caravan Buyer) > Expected Cost(Predict Not Buyer).	**/2**
- If you used an entropy-based decision tree algorithm (e.g. C4.5 as in the free CTree tool on the internet): Use the threshold probability to re-label your decision tree - i.e. indicate how the most *profitable* prediction differs from the most *probable* prediction for each node in the tree. Note that decision tree built in a profit-driven way do not need to be re-labeled (re-labeling will not change the output predictions decision trees built in a profit-driven way).	**/2**

Analysis of Data Mining Results This section is predominantly for a categorical output variable (e.g. predicting whether a person is a customer or not): adapt as appropriate for numeric output variables.	**___ / 12**
- For each model, showed a Lift Chart. Explained briefly how the lift chart was created and what the meaning of the Lift Chart is.	
- Estimated some costs and revenues for each of your prediction types (true-positive, true-negative, false-positive, false-negative), and showed a Gain Chart depicting how profit varies with a change in the number of prospects contacted.	
- Explained the Gain Chart (how it was created and what it means) and specified how you would decide how many prospects to contact.	
- Explained what other ways of determining cut-off (i.e. number of prospects to mail) are possible ?	
- Showed and explained a Classification Errors Graph (see slides on webcafe)	
- Explain what threshold values are and why you needed to adjust them (improves recall).	
- Did a Sensitivity Analysis to show how the optimal number of prospects to contact varies with changes in costs and revenues.	
- Perform a ROC analysis	
- Plot learning curves	
- Explained why overall accuracy (overall error) is not the best metric for model quality. Described what other factors influence model quality and give an assessment of your model(s).	

Knowledge Exploitation:	**____ / 6**
- Described how you propose to exploit the knowledge gained.	
- Described what systems you would integrate with.	
- Drew a diagram showing their data mining process from preparation to mining to exploitation.	
- Provided a list of recommendations (i.e. specific actions the business should take such as particular system development projects, further data mining studies, data warehousing projects, etc.). Briefly described specific goals and justified each recommendation.	

Post-Implementation Review: WEKA or other	**____ / 3**
- Provide an assessment of the software tool used for the project	
- Mentioned bugs and weaknesses found in software tool.	
- Suggested desirable improvements and new features for software tool.	
- Suggested alternatives to software tool (e.g. other data mining tools like Clementine, which may provide better sampling facilities).	

Report Structuring and Formatting:	**____ / 5**
- headings, bulleting, bolding, and diagrams used.	
- diagrams labeled and numbered and appropriately referred to in text.	
- title page, table of contents, and list of references included.	
- pages numbered and with header giving report title and student name.	
- professional presentation.	

Extra Credit:	**____ / 6**
- created a new data set that oversampled the Response (i.e. Buyer) rows from your original data set, using the threshold probability to decide the correct proportion of Buyers-vs-NonBuyers in your new data set. (That is, constructed a new data set that is a stratified sample from the original data set, but with a greater proportion of Buyers-to-NonBuyers - the exact proportion of Buyers-to-NonBuyers can be decided using the threshold probability and the formula shown in lecture on Cost Sensitive Learning). Reconstructed the decision tree using the new data set and compare the results to the decision tree you obtained by re-labeling your original decision tree.	
- performed additional analyses on the data beyond what was required.	
- specified how various mining techniques might be combined to improve results.	
- evidence of massive effort.	
- ultra-professional presentation.	