

Information Analysis for Managerial Decisions – DSC 433/533 Project

Analysis of Very Interesting Phenomena A using Data Mining Techniques B and C

by Nigel Networks, Cassie Ficayson, and R. Grayson Tree

Introduction

Briefly describe the substantive questions that you have attempted to address in your analysis, and how your findings might be used.

The introduction should be no longer than half a page.

Data

Describe the data you have used in your analysis, including descriptions of the variables, data sources, any issues with missing values (and how you have dealt with this), discussion of how the particular variables you have selected to study are relevant to the substantive questions (in the introduction), etc.

Also describe your findings from any exploratory data analysis, including any interesting patterns or relationships that could be important, any findings that prompted you to consider transformations, and any coding issues that you had to address (such as coding qualitative data with dummy indicator variables, etc.).

You might also like to include a few well-chosen graphs that reveal particularly interesting patterns in the data, or a table containing some relevant summary statistics (such as means, standard deviations, minimums, maximums, category proportions, etc.).

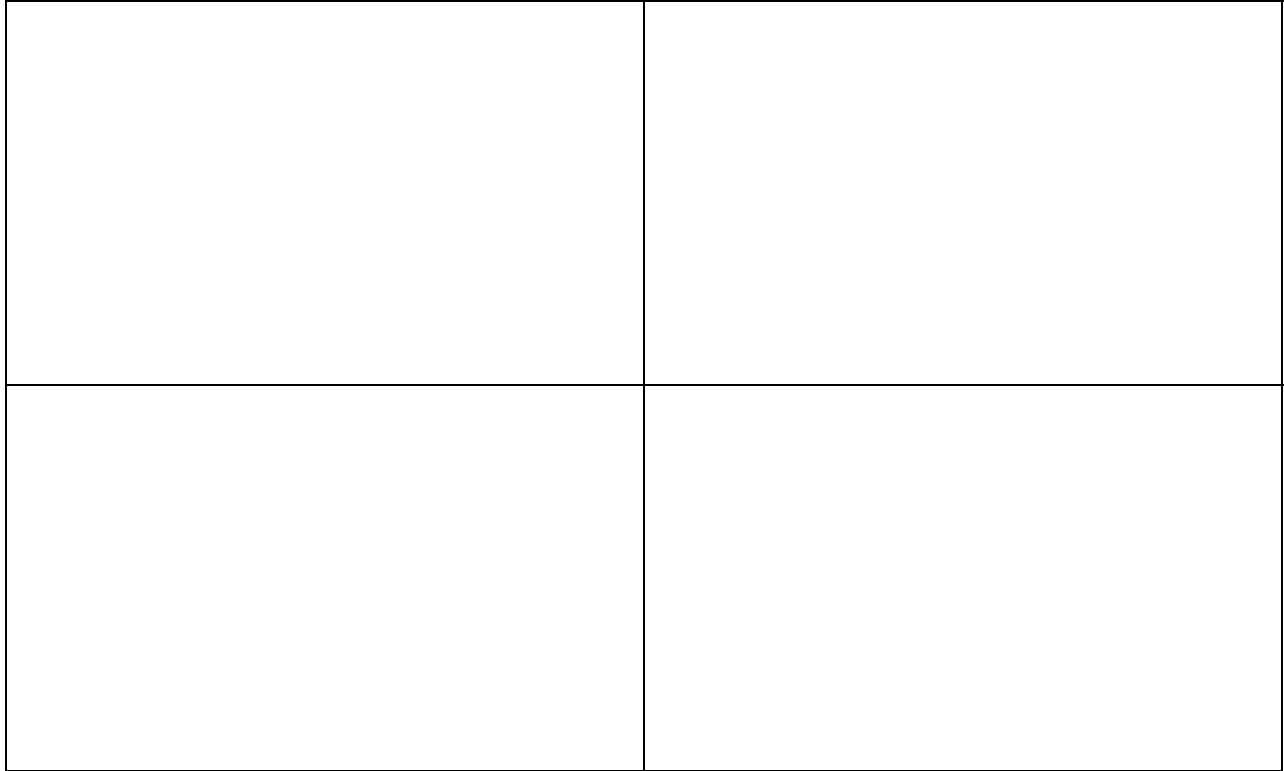


Figure 1. The upper left graph shows a histogram of some variable or other and shows The upper right graph shows boxplots of some variable for different categories represented by some other variable and shows The lower left graph shows a scatterplot of variable Y versus variable X and shows The lower right graph shows a bar chart of frequency counts of categorical variable Z and shows

Table 1. Summary statistics for variables in the dataset. Column 1 contains Column 2 contains Column 3 contains Column 4 contains. There were N missing values for variable X and M missing values for variable Y – see below for how we dealt with this in subsequent analysis.

Figure 1 contains graphs showing the most interesting and relevant patterns we observed in the data. Table 1 contains summary statistics for the variables. We dealt with the missing values in variables X and Y by This had such and such an impact on subsequent analysis.

Also make sure to describe how you partitioned the data for subsequent analysis into training, validation, and test samples.

Analysis using Techniques B and C

Give a thorough, but concise description of the analyses you conducted to address the substantive questions outlined in the introduction. Make sure to include a description of how you used the training, validation, and test samples, as well as details of any variable selection (if relevant), model comparison criteria, decisions regarding selected models, etc.

Make sure that you provide sufficient details that anyone reading your report could replicate what you have done.

Include details of any relevant calculations, useful XLMiner output, charts, etc.

Figure 2. The upper left graph shows a lift chart for model such and such, etc.

Figure 2 contains a lift chart for model such and such – this shows

You can select from any of the data mining techniques covered in class to analyze your data, and you should use XLMiner to conduct the analysis.

If, as recommended, you employ more than one technique to analyze your data (if this is possible), make sure you compare and contrast the relative strengths and weaknesses of each technique in relation to your data.

This “analysis” section should be the longest section of your report, but don’t feel the need to provide all the “gory details” of everything you did. Model building is a very creative process, invariably involving many false starts and back-tracking. By all means briefly mention some of the less successful or inconclusive things you tried, but devote more of the limited space you have telling me about the things you did that worked and that proved to be important and useful for understanding the questions you were trying to answer.

Results

Summarize the most important and relevant results from your analyses in a separate section. Take care to present your findings in a compelling way, using tables and/or graphs if appropriate.

Figure 3. The upper left graph shows

Figure 3 contains

Table 2. ...

Table 2 shows

Also, round reported final figures in a meaningful way, e.g., if decisions based on your report are likely to involve millions of dollars, there is little point providing a final profit estimate of \$1,234,567.89, particularly if the “standard error” (statistical uncertainty) for this figure is on the order of \$10,000 – better to say \$1.23m or \$1.235m.

You might consider doing additional calculations based on your results that show the value of your analysis – this might require you to make some assumptions about costs, or customers, or something else so that you can provide some “potential scenario outcomes.” For example, you might be able to come up with some projected net profit figures assuming costs are such and such – if costs are twice this, net profit would be reduced by \$X.

Conclusions

Draw appropriate conclusions from your results and relate them back to the questions from the introduction.

Present your conclusions clearly, but don't overstate what you've found, e.g., many of the techniques rely on particular assumptions being satisfied, and you may well have found significant relationships between certain variables, but these are more than likely evidence of probable associations, not necessarily evidence of cause and effect.

You could also mention possible follow-up studies that would be useful based on your results and conclusions – remember the “virtuous cycle of data mining.”

Appendix

The page limit for the main part of the report is 8 pages (11 or 12 point font, 1” margins), but if you need to include supplementary material (e.g. background reading, XLMiner output, etc.) you can include this in an appendix. I can’t guarantee that I’ll read the appendix however, so make sure that everything important goes in the main part of the report.