Data Mining Semester Project Guidelines

Summary

The purpose of the data mining project is to bring the variety of classification techniques learned in this course to bear on a dataset of interest. This project will provide the opportunity for you to consider a variety of methods that you find applicable rather than being directed to use a certain method by me or the author.

Methodologies

Please limit yourself to those supervised learning techniques introduced in Part IV of the Schmueli et al. textbook (Chapters 6 – 12). Thus, you will need to look for a dataset with a response variable (that directs or "supervises" the learning process). The project provides a chance to compare these methodologies. For a fair comparison, please try at least three (3) of the six methodologies. (I am counting regression – Chapters 6 & 10 – as one.) Note that some techniques such as nearest neighbor and CaRT also require validation data to "tune" the model, optimizing the sizes of the neighborhood *k* and pruned tree, respectively. So it might be best to partition the data into 3 datasets: training, validation, and test/evaluation.

Software

You are free to use whatever software you wish, be it Minitab, SPSS, SAS, R, Analytic Solver Platform (with XLMiner) from solver.com, etc. or any combination thereof.

Datasets

There are a variety of datasets (downloaded from the Schmueli et al. textbook site) on Canvas that students have used in the past. (Only half of those were sufficiently large; the others were small classroom examples.) This year, I would like you to find your own dataset off the internet or from other source materials. I would like each group to look at a different dataset. Assigned permission will be given on a first-come first-serve basis, so let me know what dataset you would like to examine as soon as possible.

Group Size

I have assigned student project groups on Canvas. These teams were composed for diversity of backgrounds and abilities, mixing Math/Stat, ASOR, MSA graduate students and BA&I undergraduate students for relative balance. There are 10 teams, each with one MSAnalytics student, one female student, at least one domestic student, and at most one undergraduate student. All but one group has 4 members, so the 4th member could be in charge of the evaluation process and the other 3 members tasked with analyzing the data with one of the appropriate data mining techniques.

Report Requirements

A written report (5-10 or so pages) should include a **1-page Executive Summary** that summarizes the rest of the report which details these features:

- ✓ **Introduction** – describe the problem setting and response variable of interest
- ✓ **Data Modification** – describe any data preparation, modification/ transformation that you performed.
- ✓ **Data Exploration** – describe any data exploration that you performed, including any pertinent visualizations of the data.
- ✓ **Model** – explain your modeling of the problem, including assumptions made, choice of predictor variables, and data mining techniques employed. Also discuss you modeling "hurdles," i.e., things that didn't work and how you creatively overcame these difficulties.
- ✓ **Results** – report output of the models, briefly describing the trained model parameters and their subsequent performance on a set of validation data. This section should include a comparison of the methods, with a report of errors, lift ability, statistical comparison, etc.
- ✓ **Conclusions** – summary of what you consider the best model(s) and what, if anything, you can interpret from the model.
- ✓ **Appendices** – attach computer file(s) to Canvas assignments area; due to limitations on file size, you may wish to zip your work first. I would suggest you store XLMiner results in separate workbooks, one workbook per technique as those files can be quite large.

Oral Report

A brief (~15 minutes) oral report of your conclusions will be made during the last week of classes.

Grading

The project constitutes 10% of the course grade. The project grade will depend on:

- ✓ modeling accuracy & validity
- ✓ correctness & thoroughness of results & conclusions
- ✓ report readability & style
- ✓ oral presentation of results

Due Date

Friday, Apr 24: project written report