

Final Project

The goal of this final project is to apply the techniques covered in this class to some real data. Feel free to use any appropriate data set. You can find your data set from your own work or other sources such as one of the following online data libraries:

- UCI Machine Learning Repository
<http://www.ics.uci.edu/~mllearn/MLRepository.html>
- UCI KDD Archive <http://kdd.ics.uci.edu/>
- Datasets for Data Mining at KDnuggets: <http://www.kdnuggets.com/datasets/>
- KDD Cup Datasets from 1997 to 2008: <http://www.sigkdd.org/kddcup/index.php>
- StatLib - <http://lib.stat.cmu.edu/>
- Dr. John P. Klein's website – Survival Data (you might want to ignore the failure times and just focus on status in order to apply logistic regression.)
<http://www.biostat.mcw.edu/homepgs/klein/menu.html>
- The Royal Statistical Society
<http://www.blackwellpublishing.com/rss/Volumes/A162p1.htm>

You are encouraged to collaborate on ideas but work on the project independently - feel free to discuss your ideas with me and your classmates. You may find appropriate references. The deliverables should be your own though, e.g., if several of you work on the same data set, you should each produce something separate, even if the original ideas were developed jointly. Remember to provide credit where credit is due.

Deliverables

1. In-Class Presentation (Starting from 11/23/2009, Monday)

You will have 15 minutes to present your findings during the last couple of weeks or so of class. In the presentation, you might want to introduce the data source, background of the data, clearly specify the scientific question, explain your analytic goal and statistical methods employed, and then summarize the results and interesting findings.

See the table below for your presentation date. ***Every student is required to attend others' presentations and evaluate them on a 0-10 scale.*** The score that you receive would be the average of the evaluation scores from the rest of the class, after excluding 0's and 10's.

11/23, Mon	Ahmed	Chen	Dasser	Deshpande	Dong
11/25, Wed	Ekanem	Geng	Gu	Guo	Hassan
11/30, Mon	He	Hosie	Lochrane	Pariona Gil Pasquel	Salihelamin
12/02, Wed	Siddiqui	Soyuer	Straney	Walker	Wang
12/07, Mon	Wei	Wu	Yorkos		

2. Project Report (due on 12/09/2009, Wednesday)

The project report should contain the following:

The executive summary

The executive summary should briefly describe the *source* and background of your data and capture the questions you addressed and your key results (e.g. explained in the context of the specific application). Also mention any caveats to use of the data. Summarize with your main suggestions for how to act on these results. This should be roughly 10% of the length of the full report.

Main Report

The main part of your final report should be concise and contain only 4-10 pages. It should cover the following points:

1. What problems you specifically addressed, including details in technical as well as business terms
2. The techniques that you used and the process you followed
3. Interesting results. Include the scientific meaning of the result, as well as the specific result
4. Conclusions & References if any.

Appendix: Process and Detailed results

Give an appendix that contains information that would allow someone else to repeat your analyses (assuming a reasonable knowledge of the tools used, e.g., someone else in the class) and actual printouts of the results, annotated so that if someone did rerun your analysis, they would know how to get from the raw results to conclusions.

Submission: Hard copy preferred. Electronic submission is acceptable in certain circumstances.

Scoring

Scoring will be based on (in order of importance):

1. The Process You Followed: Is it correct (given the techniques you used), did you describe it well? This includes things such as data selection, preprocessing, etc.
2. Techniques Used: Given the scientific questions you chose to address, did you approach it in appropriate ways? And justify why.
3. Interpretation of Results: Did you correctly understand and interpret the raw results you obtained?
4. Quality of Presentation and Writeup: Did you present what you did well, in an understandable and usable manner?

The above questions are key issues and answer your knowledge of data mining. The following questions will be of interest, but will have a lower impact on your final score:

- Subject or scientific questions addressed:
- Quality of results: Since this is largely an artifact of the data and your initial selection of problems, lack of interesting results won't have that big an effect on your score.