PSTAT 131/231 COURSE PROJECT

The course project is an opportunity for students to investigate a data mining problem that interests them. The course project should apply data mining techniques to real-world problems. Data and software for the project can be obtained from various Internet sites, or developed by students.

You are encouraged to collaborate on this project with the rest of the class - feel free to discuss your ideas with other students. The "deliverables", however, should be unique to you. Your project will be graded on its merits; plagiarism will result in a score of 0.

Deliverables and deadlines There will be two "deliverables" on this project.

Project report is due **June 8^{th}, 2015**. Hard copy is preferable, but electronic submission of the project by e-mail to feldman@pstat.ucsb.edu is acceptable as well. Please submit in one file only.

Project defense is required in addition to a written report. It should take place during finals week or before; **scheduling of the defense is your responsibility.** Please prepare a 5-minutes summary of your findings.

Project Report. The project report should contain the following:

1. Abstract or Executive summary should be one–two short paragraphs and summarize briefly the questions you addressed, your data mining techniques, key results, and conclusions.

2. The main body of the report should cover the following:

Introduction. Restate your problem, including details. Describe the data set and explain why this data set is interesting or important. Provide a clear description of the problem you plan to address using this dataset and techniques you use. Describe results (positive and negative) and briefly state your conclusions. Please acknowledge the source of your data and software used. should end with a summary of following sections.

Sections.

- Describe your dataset, preprocessing steps, visualization techniques. Provide analysis of your specific observations and conclusions. Describe what actions were taken (e.g., records with missing attributes removed from the dataset, etc.)
- Include all necessary plots either in the body of the report or in appendix. When including R outputs, analyze the outputs and explain what you plan to do next, why and how.
- Describe data mining techniques you used and their applications to your dataset. Explain each step.
- Compare at least two models to chose final model. Describe model validation techniques you used to choose your "best" model.
- Apply your model to do prediction when appropriate and provide analysis of your results.

3. Conclusion Section. Reiterate your conclusions referring to the goals of your project. Were these goals achieved? State directions for further work. Give acknowledgment to all individuals who helped you with this project.

4. References.

5. Appendix. Include your code with comments.

Report should not be long; please do not add words for the sake of being wordy. The report must be self-contained, i.e. if you use formulas, write them. Data is available from several libraries; urls are listed in a separate file posted at Gaucho Space. Please acknowledge the source of your data in the project.

GOOD LUCK !!!