Statistical Models: Final Project

Pick a real data set for which you believe there are interesting questions to answer. You will then try out all the different statistical learning approaches that we have covered in this course to try to find the best way to answer these questions.

Deliverables

A proposal for the project - 1 page long - Due February 1st

- Name
- Description of the problem
- Description of the dataset (dimensions, names of variables with their description)
- A discussion of the possible techniques to use
  - Supervised or Unsupervised?
  - Regression or classification?
- Comments and/ or concerns?

A poster for the presentations - Due April 17

- Description of the data and the question/s that you are interested in answering.
- Review of some of the approaches that you tried or thought about trying.
- Summary of the final approach you used and why you chose that approach.
- Summary of the results.
- Conclusions.
- A set of slides for presenting in class on April 18. Aim for a maximum time of 15 minutes.

In preparing the presentation you should be aiming it at a smart audience with statistical training to the level of multiple linear regression but not beyond. Hence, you should not just say "We did KNN" but also explain the basic idea of how it works, why it might be better than linear regression etc. Among other things, points will be allocated for clear articulations of the question of interest, the approach you used to solve it, the reason you chose that approach, and the conclusions you were able to draw.

A report to be handed in. First draft due March 20. Final version due April 17.

The report will contain a summary of the material covered in the presentation (maximum 3 pages). The first page should be an executive summary. The report must also include the slides from the presentation and a technical appendix, which should include your code (maximum 10 pages).

Some Data Repositories

- Open Gov. Data: http://www.data.gov , http://www.data.gov.uk, http://www.data.gov.fr, http://opengovernmentdata.org/data/catalogues/
- Kaggel: http://www.kaggle.com
- KDD Nugets: http://www.kdnuggets.com/datasets/
- UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
- StatLib: http://lib.stat.cmu.edu
- TwitteR: http://cran.r-project.org/web/packages/twitteR/index.html
- rfigshare: http://figshare.com, http://cran.r-project.org/web/packages/rfigshare/index.html

You are not restricted to this repositories.