IOM 530: Applied Modern Statistical Learning Methods
Final Group Project
Fall 2013

The premise of this project is very simple. You are to 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. The project will be completed in groups of 5 people each.

Deliverables

This project includes two deliverables:

1. A proposal for the project- one page long (**Due Thursday Oct 31, 2013**)
    a. Members' names
    b. Description of the problem
    c. Description of the dataset (dimensions, names of variables with their description)
    d. Supervised or Unsupervised?
    e. Regression or classification?
    f. Comments and/ or concerns?

2. A poster for the presentations would be (**Due Tuesday Nov 26, 2013**)
    a. Description of the data and the question/s that you are interested in answering.
    b. Review of some of the approaches that you tried or thought about trying.
    c. Summary of the final approach you used and why you chose that approach.
    d. Summary of the results.
    e. Conclusions.

    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.

3. A report to be handed in. (**Due Tuesday Dec 3, 2013**)
    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 R code (maximum 10 pages).

IOM 530: Applied Modern Statistical Learning Methods
Final Group Project
Fall 2013

Data Repositories

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

Suggestions

1. I highly recommend starting to work on your project R code as soon as you get my approval on your dataset. If your proposal is ready before the deadline, please feel free to send it to me for approval.
2. Since you will be applying the methods we learned in this class on your datasets, your assignments' R code should be very helpful!
3. Please don't wait until the last week, because you will be missing THANKSGIVING.
4. Thank you!