DS502/MA543: Statistical Methods for Data Science
Final Group Project

The idea of this project is that you are to pick a real data set for which you believe there are interesting questions to answer. You will then try out at least **two statistical learning approaches that we have covered in this course (though you are certainly allowed to do more)** to try to find the best way to answer these questions (you are, of course, welcome to choose *additional* approaches either from class or your own reading). An important aspect of your results will be **appropriate error assessments** for the results you provide. The project will be completed in groups of 3-5 people each.

Deliverables

This project includes **three deliverables**:

1. A proposal for the project- 1-2 pages long (**Due Tuesday October 30**)
   a. Members' names
   b. Description of the problem
   c. Description of the data set (dimensions, names of variables with their description)
   d. Regression or classification?
   e. The methods you plan to try.
   f. The error metrics you plan to use and the algorithms for assessing them.
   g. Comments and/or concerns?
2. Slides for an approximately 15-minute presentation (**Due November 27, with presentations November 27 and December 4, if necessary**).
   a. Description of the data and the questions that you attempted to answer.
   b. Review of the approaches that you tried or thought about trying. It is interesting and useful to discuss both successes and failures!
   c. Summary of the final approach you thought worked best and why you chose that approach.
   d. Summary of the results.
   e. Conclusions.

   Points will be allocated for the explanation of the question of interest, the descriptions of approach you used, the reasons you chose your final approach, and the conclusions you were able to draw, both positive and negative.

3. A report to be handed in. (**Due November 27)**

   The report will contain a summary of the material covered in the presentation (between 10-15 pages).

Note: As a matter of fairness, the final version of all materials (report and presentation) are due on November 27, even if your presentation is not on that day. In particular, the slides you submit on November 27 are the slides you must present, even if your presentation is December 4.

All materials to be submit by email to Prof. Paffenroth (rcpaffenroth@wpi.edu).

Data Repositories that you might consider

1. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
2. Statlib datasets: http://lib.stat.cmu.edu/
3. Kaggel: www.kaggle.com
4. Open Gov. Data: www.data.gov, www.data.gov.uk, www.data.gov.fr, http://opengovernmentdata.org/data/catalogues/

Project description inspired by: http://www.alsharif.info/#!iom530/c21o7