Overview of Statistical Learning and Modeling – 36-600

Final Poster Project
First Draft Due Wednesday, December 10, 6:00pm
Final Draft Due Friday, December 12, 11:59pm

What are the expectations for the poster?

First, from a structural perspective, you'll want the following elements:

- An introduction section that specifies the research question. This can be short.
- A data section that describes the data and provides, as room allows, high-level EDA. Determining what to put here might be the hardest part.
- An analysis section that lists the models you used and the choices your made (like data splitting choices). This section would also contain tables (e.g., MSE for the models, or AUC/MCR for models) and any applicable figures or confusion matrices.
- A conclusion section that answers the research question. This can also be short.
- See the accompanying poster guide for more. . . er. . . guidance.

Note that your team should be working on one dataset! That sounds obvious, but what we mean is, don't have different team members split the data with different random number seeds and pursue different analyses, because the results cannot be merged together: any MSEs, for instance, are not directly comparable.

If your team is pursuing a regression project:

- Assume your goal is prediction
  - You do not need to show a correlation plot, nor compute VIF values
  - You do not need to explore PCA
  - You *should*, however, do best-subset selection (if possible, or do forward/backward stepwise selection), comment on the selected variables, and compare the BSS MSE with the MSEs for other models
  - You should show a table of considered models along with test-set MSEs for each, and specifically identify a best model
  - You do need to, among other things, show a diagnostic plot for the best model, i.e., a plot of predicted test-set response values (y-axis) versus observed test-set response values (x-axis)

If your team is pursuing a classification project:

- Assume your goal is prediction
  - You do not need to show a correlation plot, nor compute VIF values
  - You do not need to explore PCA
  - You *should*, however, do best-subset selection (if possible, or do forward/backward stepwise selection), comment on the selected variables, and compare the BSS AUC with the AUCs for other models

- You should compute AUCs for all models (show a table!) and pick the model with the best one, then use Youden's J statistic to determine the best threshold and ultimately determine a final confusion matrix and MCR value
- Bonus: it will look cool if you can overlay all ROC curves on one plot

As a final note: all team members are expected to contribute equally to the construction of posters. If any of your team members become "ghosts," let me know; any "ghosting" is considered not giving a good-faith effort and any "ghosts" will have their scores reduced by an appropriate amount.