To complete this homework, submit a report by May 4th, in PDF format, with an embedded URL in which you link to a repository (e.g., Box, Dropbox, Google Drive) that contains the data and code necessary to replicate your results.

1. Find a replication archive (e.g., via a journal's Dataverse site) for a recently published article in which the authors estimate a logistic regression model. Confirm that you can reproduce the results from the article, estimating the same coefficient values for the corresponding variables in the model (standard error estimates vary based on the software used and settings selected, but you should be able to reproduce the coefficients). Write up a brief introduction to the article, variables, and model, and present the replicated results in your report.
2. Split the data into 70% training and 30% test sets. Use the training set alone for this part of the exercise. Evaluating performance via cross-validation, find a change to the original model specification (e.g., dropping variables, adding interactions, adding polynomial functions of variables) that improves the predictive performance. Comment on whether you think the improvement in performance reflects you moving the model closer to the true data generating model; or if the improvement in model fit reflects the contradiction highlighted by Shmueli, that mis-specified models can fit better in predictive terms.
3. Compare the original model and the new specification you discovered, estimated on the full training data, in predicting the test data. Be sure not to estimate either model using the test data. Comment on which model fits better on the test data, and interpret your results in light of your response to (3).
4. Report the new specification, estimated on the full dataset, and comment on whether or not you think you discovered anything new about the underlying process being modeled, via this exercise.