Third Assignment — Stat 474/974 — Spring, 2017

A high school guidance counsellor wants to be able to advise students about a college into which they are likely to be admitted. Will they be admitted or not? What factors are driving the admissions prediction?

As a "demonstration of concept," the guidance counsellor wants to try out a forecasting procedure for admission to a particular "elite" university. To this end, she is able to get data from that university about the previous year's applicants. The dataset ("admissions974.rdata") has 8700 observations and 9 variables. The response is the variable "admit." The other variables are predictors chosen because they are thought to be related to admission. The following variables are in the data.

1. admit — Coded "1" for admit and "0" for reject.
2. anglo — Coded "1" for anglo and "0" otherwise.
3. asian — Coded "1" for asian and "0" otherwise.
4. black — Coded "1" for black and "0" otherwise.
5. gpa.weighted — high school GPA weighted for AP courses.
6. sati-verb — SAT verbal score.
7. sati.math — SAT math score.
8. income — household income capped at $100,000.
9. sex — Coded "1" for male and "0" for female.

Suppose you are the guidance counsellor. Your job is to undertake a forecasting analysis and write it up to show to your principal. Use random forests to arrive at your algorithmic forecasts. A key will be to employ a sensible and well-justified cost ratio for false positives and false negatives with the primary performance measures taken from confusion tables. But you need to explain as well what is driving the the forecasts. Use importance plots and partial dependence plots to do that. Think hard about this because you need to be able to explain to the principal what your results mean.

As before, start with a problem statement followed by a description of the data. Then proceed from univariate statistics, to bivariate statistics to random forests (as a classifier): confusion table, predictor importance, and partial plots. Make sure to finish with a summary of your findings and overall conclusions. How well do you forecast? What predictors seem to matter for accuracy? How do they matter? Be sure to address whether you are doing a level I or level II analysis and with what justification. If level II, what are you estimating? And make sure you clean your data where needed.

Write with precision and clarity. Five pages of single-spaced text should be plenty. Number your pages, figures and tables. Include your R-code in an appendix in case we cannot figure out what you did.