To complete this homework, submit a report by April 6th, 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, identify a support vector machine, random forest, and neural network that you think perform well in predicting the dependent variable, using the same independent variables, or a subset thereof, used in the model replicated for Question 1.
3. Compare the original logit model and the other machine learning methods using the test data. Be sure not to estimate any model using the test data. Comment on which model fits better on the test data.
4. For each method---logistic regression, SVM, random forests, and neural networks, use the full data to evaluate the importance of each variable in terms of its contribution to the predictive performance of the model. Comment on whether the importance of variables varies significantly across the model types.