Data Mining I
Homework 2

1) In this exercise, we will predict the number of applications received using the other variables in the College data set in the ISLR2 package.
*** be sure to look closely at this data, you may want to consider the multi-scale nature of the problem, and perhaps use a transformation on some of the variables.***
(a) Split the data set into a training set and a test set. Fit a linear model using least squares on the training set, and report the test error obtained.
(b) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.
(d) Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.
(g) Comment more generally on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

2) The insurance company benchmark data set gives information on customers. Specifically, it contains 86 variables on product-usage data and socio-demographic data derived from zip area codes. There are 5,822 customers in the training set and another 4,000 in the test set. The data were collected to answer the following questions: Can you predict who will be interested in buying a caravan insurance policy and give an explanation why? Compute the OLS estimates and compare them with those obtained from the following variable-selection algorithms: Forwards Selection, Backwards Selection, Lasso regression, and Ridge regression. What methods performed the best in the prediction?
(The data can be downloaded from https://kdd.ics.uci.edu/databases/tic/tic.html.)

3) ESL textbook exercise 2.8 modified: Compare the classification performance of linear regression and k-nearest neighbor classification on the zipcode data. In particular, consider only the 7's and 9's for this problem, and k = 1, 3, 5, 7, 9, 11, 13,15. Show the test error for each choice of k. Describe your results – are you surprised by the differences in performance?

The zipcode data is available https://hastie.su.domains/ElemStatLearn/ in the "Data" tab. The data can be obtained as test/training and/or by number.