ISYE 6501, Week 7 HW

**Question 1**

Using the same crime data set as in Homework 5 Question 2, find the best model you can using (a) a regression tree model, and (b) a random forest model. In R, you can use the tree package or the rpart package, and the randomForest package. For each model, describe one or two qualitative takeaways you get from analyzing the results (i.e., don’t just stop when you have a good model, but interpret it too).

**Response –**

Using regression tree model, the best model came out to be

Crime = 819.6 + 9.49\*Ed + 11.395\*Pop – 3164.075\*Prob – 12.13\*Time

The R-squared value = 0.4485

The random forest model gave only a R-squared value of 0.419 which is less than the regression tree model.

With regression tree model, the parameters considered in the model are clearly known but not with the random forest model.

**Question 2**

Describe a situation or problem from your job, everyday life, current events, etc., for which a logistic regression model would be appropriate. List some (up to 5) predictors that you might use.

**Response –**

Logistic regression model can be used to predict if a political candidate can win the election or not.

Predictors –

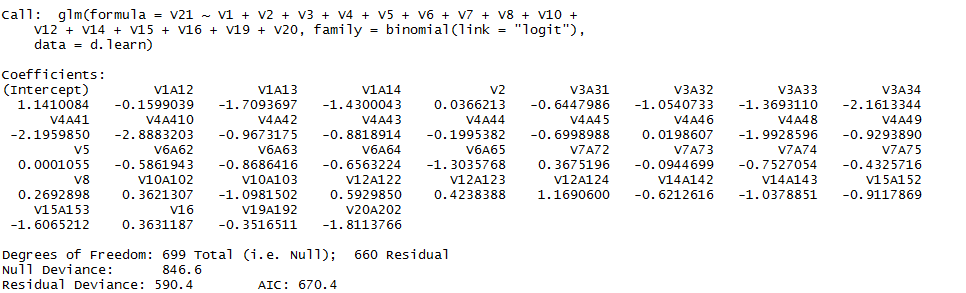
1. Time spent on campaigning
2. Money spent on the campaign
3. Incumbent president or not
4. Years in politics
5. Popularity (in percentage)

**Question 3**

Using the GermanCredit data set germancredit.txt from http://archive.ics.uci.edu/ml/machinelearning-databases/statlog/german / (description at http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29 ), use logistic regression to find a good predictive model for whether credit applicants are good credit risks or not. Show your model (factors used and their coefficients), the software output, and the quality of fit. You can use the glm function in R. To get a logistic regression (logit) model on data where the response is either zero or one, use family=binomial(link=”logit”) in your glm function call. 2. Because the model gives a result between 0 and 1, it requires setting a threshold probability to separate between “good” and “bad” answers. In this data set, they estimate that incorrectly identifying a bad customer as good, is 5 times worse than incorrectly classifying a good customer as bad. Determine a good threshold probability based on your model.

**Response –**

Software Output



Good threshold – 0.4