CSC570E Machine Learning
Final Project

College Ranking

(due by the end of the day on Thursday, July 25th)

Download the dataset colleges.csv. It contains 17 variables for 777 different colleges in the US. The variables are:

Private : Public/private indicator
Apps : Number of applications received
Accept : Number of applicants accepted
Enroll : Number of new students enrolled
Top10perc : New students from top 10% of high school class
Top25perc : New students from top 25% of high school class
F.Undergrad : Number of full-time undergraduates
P.Undergrad : Number of part-time undergraduates
Outstate : Out-of-state tuition
Room.Board : Room and board costs
Books : Estimated book costs
Personal : Estimated personal spending
PhD : Percent of faculty with Ph.D.'s
Terminal : Percent of faculty with terminal degree
S.F.Ratio : Student/faculty ratio
perc.alumni : Percent of alumni who donate
Expend : Instructional expenditure per student
Grad.Rate : Graduation rate

Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide colleges into two groups based on whether the proportion of students coming from the top 10% of their high school classes exceeds 50%, i.e., college$Top10perc >50.

The task is to predict whether a college is elite or not.

1. Apply two machine learning algorithms to the dataset. One of the algorithms should have not been used in previous homeworks. In other words, one of the algorithms must be different from kNN, Naïve Bayes, linear regression and C5.0. Use confusion matrices and different performance measure to compare the algorithms.

2. Perform automated parameter tuning for both models (if they allow it) using the caret package.

3. Try to improve the performance of each algorithm by using ensemble learning (one method of ensemble learning of your choice) and the caret package. Compare the algorithms.

Write a report (not to exceed 3 pages) describing what you have done and found, the algorithm comparisons and summarizing the results.

Your submission must consist of two files:
- a report as a txt, docx, or a pdf file
- a script with history of your session. Save the session into script using savehistory("script.Rhistory"). You can manually edit the script in order to remove any unnecessary commands, such as trials and errors. The order of the commands must follow the order stated above.