***BUS 347.01 Introduction to Business Analytics***

***Homework 7 (60 Points)***

**Homework Description:**

The attached file “HW7 Data.csv” is used in this homework assignment. “HW7 Data.csv” contains the following variables regarding whether a bank customer has subscribed a term deposit or not (the target variable y in the dataset).

In this homework assignment, you need to build a predictive model to forecast whether a client has subscribed a term deposit or not (target variable y in the dataset).

* bank client data:
  + age (numeric)
  + job: type of job (categorical)
  + marital: marital status (categorical)
  + education (categorical)
  + default: has credit in default? (categorical: “no”, “yes”, “unknown”)
  + housing: has housing loan? (categorical: “no”, “yes”, “unknown”)
  + loan: has personal loan? (categorical: “no”, “yes”, “unknown”)
* related with the last contact of the current campaign:
  + contact: contact communication type (categorical: “cellular”, “telephone”)
  + month: last contact month of year (categorical: “jan”, “feb”, “mar”, …, “nov”, “dec”)
  + day\_of\_week: last contact day of the week (categorical: “mon”,“tue”,“wed”,“thu”,“fri”)
* other attributes:
  + campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
  + pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
  + previous: number of contacts performed before this campaign and for this client (numeric)
  + poutcome: outcome of the previous marketing campaign (categorical: “failure”, “nonexistent”, “success”)
* social and economic context attributes
  + emp.var.rate: employment variation rate - quarterly indicator (numeric)
  + cons.price.idx: consumer price index - monthly indicator (numeric)
  + cons.conf.idx: consumer confidence index - monthly indicator (numeric)
  + euribor3m: euribor 3 month rate - daily indicator (numeric)
  + nr.employed: number of employees - quarterly indicator (numeric)

**Output variable (target variable):** y: has the client subscribed a term deposit? (binary: “yes”,“no”)

You need to apply two tree-based models to forecast the probability that a customer subscribe to the term deposit and then compare those two models. Answer each homework question to develop the model in your analytics report. Your submission should be an analytics report in the word format knitted from RMarkdown code.

In your analytics report, you need clearly label the following items:

* Question Number
* R Code
* R Output
* Conclusion, if applicable

You need to submit the knitted analytics report on Blackboard. Please carefully check your work before the submission, as you can only submit your work once. Late submissions will not be accepted.

**Homework Questions (Each Question has 10 point)**

**Q1.** Use a simple random sampling method to split the whole sample into training (30000 observations) and validation (the remaining 11188 observations). Use a random seed 12345 to generate the random row index. Show the proportion of y in both training and validation dataset in your analytics report.

**Q2.** Based on the training sample, create a default decision tree model by using all explanatory variables in the dataset. Plot the tree structure, display the variable importance table and plot the CP plot.

**Q3.** Based on the training sample, create a more complicated tree by setting the cp value equals 0.0001. Generate the CP plot of this complicated tree and discuss whether there is any need for pruning. If the tree needs prune, prune this complicated tree by choosing the cp value that minimizes the cross-validation error in the complicated tree’ cp table. Plot the pruned tree structure.

**Q4.** Based on the training sample, create a random forest model by using all explanatory variables and setting ntree=500 and mtry=5. (put the random seed = 67890 before running the randomForest() function). Plot the variable importance for this random forest model.

**Q5.** Based on the training sample, use the ROC curve and AUC to determine whether the pruned decision tree or the random forest model produces a better in-sample model fit. Draw your conclusion explicitly by showing the AUC. (Hint: y is “no”/”yes” in this dataset, not “0”/”1”; you need to adjust the class demonstration code accordingly to fit this question scenario).

**Q6.** Based on the validation sample, use the ROC curve and AUC to determine whether the pruned decision tree or the random forest model produces a better out-of-sample model fit. Draw your conclusion explicitly by showing the AUC.