STAT330/430 - Statistical Learning

Assignment 3
Due date - May 9th

Note: You must submit two files for this assignment:

- a pdf document of your solutions, and
- a complete and concisely annotated *R Script* file of all **R** analysis that was undertaken to produce your results.

In addition, to receive marks for Question 1 you are required to engage in the Topic 3 moodle discussion forum.

Please refer to the Assignment assessment criteria document for additional guidance.

Question 1 [5 marks]

In each assignment you will be able to earn up to 5 marks based on your engagement on the moodle Discussion forums from the topics associated with the given assignment. See the Assignment assessment criteria document for more details.

Question 2 [25 marks]

The activities of the instant chat function on a company website were reviewed to identify key features that allowed a given query to be successfully resolved. The data-set *Chatter* contains the following variables:

Variable	Description
Worker	10 point self-assessment score on session given by the employee
Max_msg	Maximum number of characters used in any message
Min_msg	Minimum number of characters used in any message
Exchanges	The total number of messages exchanged during the session
Total_time	Total time the customer was active for
Time_length	Average time (in secs) customer waited for a response from the employee
Age_client	Age of the customer
Resolved	Whether the customer considered the issue resolved ("No" or "Yes")

(a) Using only complete records - use a combination of figures and text to conduct an exploratory data analysis and summary of the *Chatter* dataset.

(b) Using a 80:20 train:test split create a decision tree for the *Chatter* data-set. Explore whether the tree should be pruned. Provide an overview of your results, final decision tree, and a succinct summary of all relevant summary values.

(c) Fit a bagging *OR* random forest model using 10-fold cross-validation to the *Chatter* data-set. Use at least 5 different values for each of the relevant tuning parameters. Provide carefully labeled and well-formatted tables and/or figures - as well as an informative summary of your model and results.

Question 3 [15 marks]

(a) Using the same train and test data-sets as Question 2 fit a SVM to the *Chatter* data-set. Use a range of values to tune the models appropriately. Provide carefully labeled and well-formatted tables and/or figures - as well as an informative summary of your model and results.

(b) Compare and contrast the results and modeling approach of all analyses conducted in both Questions 2 and 3.