STAT606 Assignment Instructions

Find a data set to perform binary classification. Follow the tasks below and prepare a presentation of the results. The presentation should be no more than 10 minutes.

1. Give a brief introduction to the data and problem.
2. Familiarize yourself with the data. Fix any inconsistencies or missing values that may exist (you can use any software for this). How these were fixed should be *briefly* mentioned and motivated in your presentation.
3. Give some basic characteristics of the sample and target variable as well as state what attributes/features were included.
4. You may use existing attributes to create new ones to be used in the models if you think they would be more suitable/informative.
5. Using an 80:20 train vs test set split on the data, apply ***four*** machine learning techniques to classify the response based on the set of attributes in the data, one of which must be an ***ensemble*** method.
6. Comment on the results in terms of any overfitting, underfitting, high variance etc. that may be present. Determine the best performing model according to a measure of your choice. **Substantiate the measure you use**.
7. Further improve the performance of your model by finding the optimal cut-off point. State which optimal cut-off was obtained and explain what it means with respect to classifying your target variable (i.e. how is the cut-off point used to make a classification).
8. Provide the top 5 most important features based on your best performing model.

Presentations will take place between 12 and 20 May. Both group members are expected to contribute to the presentation.

Your presentation must be submitted in PDF form by 09h00 on Tuesday 20 May.