Homework 3

MSCS 6520 Business Analytics
Spring 2018
Assigned: February 12, 2018
Due: February 19, 2017 (by beginning of class)

Readings

Chapters 5 and 7 of *Data Mining for Business Analytics*

Chapters 1, 3 – 5, 17 in Caret documentation

Exercises

1. Choose one of the classification datasets in Chapter 23 of the Caret package documentation: Animal Scat Data, Cell Body Segmentation Data, German Credit, DHFR Inhabitation
2. Choose 10 of the features for study (use the `select()` function). If factors have been expanded into binary dummy variables, you will need to use all of the columns associated with the factor (e.g., CreditHistory, Purpose in the German Credit dataset). If using the German Credit dataset, you can't use the same 10 features I used in the slides.
3. Perform exploratory data analysis. Plot each feature versus the variable you are trying to predict. Which features do you think will make the best predictors?
4. Perform a round of forward feature selection. Which are the accuracies for models built with each feature? Which feature is best?
5. Perform a second round of forward feature selection. What were the accuracies? Were you able to improve the accuracy using two features vs a single feature? Which two features performed best?
6. Keep performing the forward feature selection until either you see no improvements (e.g., after four features there is no improvement) or all of the features are included in the model.
7. Repeat steps 2 – 6 for another dataset from the Caret package (only if working in pairs)

Prepare a document containing the answers to the above questions and plots. Submit the document as a PDF to D2L. You may work in pairs, in which case, you should only submit one PDF per group.