Homework 5

MSCS 6520 Business Analytics
Spring 2018
Assigned: April 2, 2018
Due: April 9, 2017 (by beginning of class)

Exercises

For this assignment, you're going to repeat Homework 3 using Logistic Regression (LR) instead of k Nearest Neighbors (kNN) and then compare and contrast the two algorithms.

1. Review your Homework 3 submission to identify the classification dataset (Animal Scat Data, Cell Body Segmentation Data, German Credit, DHFR Inhabitation) from Caret and associated features that you used for Homework 3
2. Perform a round of forward feature selection to build a Logistic Regression model using the 10 features. Which are the accuracies for models built with each feature? Which feature is best?
3. Perform a second round of forward feature selection using the 5 best features from the previous step. What were the accuracies? Were you able to improve the accuracy using two features vs a single feature? Which two features performed best?
4. Keep performing the forward feature selection until either you see no improvements or all 5 of the features are included in the model.
5. Repeat steps 2 – 6 for another dataset from the Caret package (only if working in pairs)
6. Reflection: Do you notice any differences between building a model with LR vs kNN? For example, was one model more accurate than the other? Did the accuracy of both models go down for "bad" features?

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.