Week 6 Assignment

Note: Late homework assignments are not accepted because the solutions are discussed at the start of every class on Wednesday. Your solution to this assignment must be uploaded on eCampus as a PDF file. A common approach is to put the solution into a word document and then save that into a PDF file. Please only submit PDF files.

Assignment: You can complete one of the two parts to this assignment, or both. If you do both parts, you can obtain extra points for this assignment. You are expected to complete at least one part successfully.

Data File: OilProduction.sas7bdat

Part 1: Create a SAS EM project named "Week 6 Homework". In that project read this data file for this assignment. Import the data, ensuring that all attributes have the proper metadata described in the data dictionary. In this case, the target is 'Log_Cum_Production', an interval target.

These data do no contain missing values or outliers. There is no need to "clean" these data. Use 'gini' as the split criterion.

Fit these data to a Random Forest fit using the SAS default settings, and compare this to a decision tree. For the decision tree select the best depth from 3-15 using 4-fold cross-validation.

Part 1: SAS EM Solution Upload (all screen shots must be readable)

1. A screen shot of your project diagram
2. A screen shot or listing of ALL SAS code used in your diagram.
3. A table of the metrics for each of the decision tree cross-validation folds.
4. Describe which decision tree you selected and why.
5. Compare the best decision tree to the random forest solution.

Part 2: Using the data file OilProduction.xlsx, compete the same assignment as Part 1 using Python.

Create a random forest solution and compare it to the best decision tree selected using 4-fold cross-validation. Review depts from 3-15. Use the default settings for the random forest.

Use one-hot encoding for the nominal attributes in these data.
Prepare a report containing:

1. A listing of your python code.
2. A table of the metrics (recall, accuracy, precision and F1) calculated for each of your 4 cross-validation folds.
3. Describe which model you selected from Cross-Validation, and why.
4. Compare the best decision tree solution to the random forest solution.