ASSIGNMENT 4
ADVANCED METHODS AND ANALYSIS FOR THE LEARNING AND SOCIAL SCIENCES
PROFESSOR RYAN S.J.d. BAKER
FEATURE ENGINEERING
DUE NOON, WEDNESDAY FEBRUARY 15

PLEASE TURN IN THIS ASSIGNMENT BOTH TO RYAN AND TO SUJITH GOWDA, SUJITHMG@WPI.EDU

The goal of this assignment is to build a better behavior detector (classifier), using the data in Asgn3-dataset.csv and the data in Asgn4-dataset.csv.

These two datasets represent the same data set, but at two different grain-sizes. Specifically, Asgn4 represents individual student actions within educational software, while Asgn3 is at the grain size of all the actions that occurred during 20 second field observations by trained coders. Note that the individual student actions are labeled with the same UniqueID labels as the observations are (each UniqueID corresponds to a single field observation).

In this assignment, you must conduct feature engineering to improve the features in the Asgn3 data set, using the data in the Asgn4 data set. You must create at least 10 new features that cannot be created using just the Asgn3 data set, and add the new features to the Asgn3 data set. You can create new features in Excel, or in any automated fashion you like.

Then you must build a detector of the behavior OffTask (e.g. a detector that can predict if the column OffTask is Y or N), using both the old and new feature sets. The model with new features must have both better A' and Kappa than the model with old features – and also must habe better A' and Kappa than your hand-in for Assignment 3. If you did not hand in Assignment 3, then you must do better than the average solution to Assignment 3 turned in by your classmates. I'll announce this standard in class.

As with Assignment 3, you should make sure that your detector is not over-fit, paying particular attention to making sure that your detector does not use features that could not be used when applying the model to new data or new students. This can be done both by restricting the features used during model fitting, and setting up cross-validation in an appropriate fashion. (Hint: Try Batch Cross-Validation).

You must build the detector using an automated algorithm. You cannot simulate the algorithm in Excel. You can use any data mining package (e.g. SAS, R, Weka, KEEL) you want, but I strongly recommend using RapidMiner 4.6.

Please turn in:

- The data sets you input into the data mining package, both for the old and new features.
- The model built on the new data set

- Evidence of model goodness, when the model is applied to new students, for both the old and new feature sets
- All data mining code you used to generate the outputs
- A document explaining how you completed the assignment

You will be graded on completeness and comprehensibility of your hand-in, whether you correctly and validly apply the method you choose to this data, and whether the methods you chose fit the requirements of this assignment.

BONUS: The student who succeeds in producing the detector with the best A' and Kappa (averaged together) under appropriate cross-validation, gets the bonus.