CREATIVE ASSIGNMENT 2
CORE METHODS IN EDUCATIONAL DATA MINING
PROFESSOR RYAN BAKER
FEATURE ENGINEERING
YOUR ASSIGNMENT IS DUE OCTOBER 19, 7PM USA EASTERN
YOUR RESPONSE POSTS ARE DUE OCTOBER 24, 7PM USA EASTERN

The goal of this assignment is to build a better behavior detector (classifier), using the data from Creative Assignment 1, as well as new data provided for this assignment.

These two datasets represent the same data set, but at two different grain-sizes. Specifically, the new data represents individual student actions within educational software, while the previous data set 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 original data set, using the data in the new data set. You must create at least 10 new features that cannot be created using just the original data set, and add the new features to the original 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 (or just the new feature set). Ideally, the model with new features will have both better AUC and Kappa than the model with old features – and also will have better AUC and Kappa than your hand-in for Creative Assignment 1. However, it is OK to just use Kappa, if you are using RapidMiner and cannot get the A' code to work.

As with Creative Assignment 1, 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.

You must build the detector using an automated algorithm. You cannot simulate the algorithm in Excel. You can use any data mining package you want.

Please post to the forum, in a new thread within the CA2 folder:

- 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 AUC and Kappa (averaged together) under appropriate cross-validation, gets the bonus.

New features:
Row = arbitrary order
Lesson = lesson in online learning software
Prod = cognitive skill involved in current action
Cell = interface widget involved in current action/ problem step
Pknow-1 = estimate of student skill after current action (or -1 if estimate did not change)
Pknow-2 = estimate of student skill after current action
Pchange = did probability that student knows skill change? (i.e. is it the first attempt for this student on the current problem step)

PART TWO: YOUR RESPONSE POSTS

After completing your own assignment, you are expected to also provide substantive comments on at least four other students' submissions, as a response within that student's assignment thread. For these posts, there is no length requirement, but the posts must offer a critical and meaningful perspective on how that student did the assignment. (i.e. "Great job! You did really awesome!" and "Terrible! You totally messed up!" are insufficient)

This is not just for the benefit of the student whose solution you are commenting on. Seeing how other students did this assignment will be informative to you as well.

Although there is no requirement to do this, you are encouraged to give feedback to students who have received fewer feedback responses so far – i.e. I would like to avoid having one student get feedback from every classmate, and another student get feedback from no one. Thanks.