ASSIGNMENT 5
SPECIAL TOPICS IN EDUCATIONAL DATA MINING
PROFESSOR RYAN S.J.d. BAKER
REGRESSION
DUE NOON, MONDAY MARCH 11

The goal of this assignment is to build a regression model, using a regression algorithm of your choice, using the data in Asgn5-dataset.csv

This data set involves a set of student actions from an intelligent tutoring system. Each action is labeled with a probability of slip, P(SLIP|TRIO). These values are calculated using the method in (Baker, Corbett, & Aleven, 2008), and for the purpose of this assignment can be treated as ground truth.

You must build a regression model to predict this label. For the purposes of this assignment, you can ignore student-level independence issues during cross-validation. For those aware of this issue, you can also ignore truncation issues for extreme values of pknow. You should still use cross-validation. You can use any feature selection, or feature engineering approaches you want, and any appropriate algorithm.

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 set you input into the data mining package, if different than the original data set
- The model built on the full data set
- Evidence of model goodness, when the model is applied to new students (see the Diagnostic Metrics lecture)
- 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 correlation (r) under either k-fold or leave-out-one cross-validation, gets the bonus.