Capstone Project Proposal

What is the problem you want to solve?

Identifying students who may need additional help is an issue for all educational institutions. Test results during a course or program provide insight into this issue, but this measure is delayed; educators may not be aware an issue exists until it is too late to intervene. Being able to identify at risk students early on would allow educators to provide assistance to at risk students before problems arise and improve student outcomes.

Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

The outcome of this analysis could be used by any number of educational institutions, from primary and secondary schools, to post secondary institutions. Early identification of at risk students would allow for the development of early intervention planning, with the goal of improving student outcomes.

What data are you going to use for this? How will you acquire this data?

The Student Alcohol Consumption data set from the UCI Machine Learning repository will be used for this analysis. This data set contains 32 attributes, ranging from social factors such as age, gender, and family size, collected via a student survey, as well as numerical attributes including number of absences and grades. Two data sets are included, one for students in a Math class and the other for students in a Portuguese class.

A description of the data, and the data sets can be found at
http://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION

In brief, outline your approach to solving this problem (knowing that this might change later).

The data sets will need to be combined, and then split into training and test sets. Next, an examination to identify variables that may be useful, or not useful in predicting class outcomes (grades) will be conducted. After removing the non-predictive variables, further examination will be needed to examine interrelationships, as well as the possibility for feature engineering. A classification model will be developed in order to identify the characteristics that are predictive for each group. This model will then be tested on test set to confirm accuracy and validity.

What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The deliverables will include a GitHub repository which will include the raw data, R code that was used to perform the analysis, and a written report and slide deck outlining the problem, steps taken in the analysis, and the results.