15.077 Project Proposal

Shi Wang

1. Source of dataset:
student performance dataset
https://archive.ics.uci.edu/ml/datasets/Student+Performance

2. Brief description of dataset
The tentative task of this dataset is to predict the student performance in secondary education. There are totally 649 observations and 33 explanatory variables. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. The dependent variable is G3 (final year grade, which is a continuous, integer variable ranging from 0 to 20.).

3. Work plan
(1) Descriptive analysis

To study the influence of these explanatory variables on the response variable. This part may include histograms, ANOVA test, correlation analysis, etc. This part aims to get good understanding of this particular dataset, and provide guidance for further modeling variable selection.

(2) Since we are not in a data-rich situation, and the goal of this study is not to infer the predictive capability of the developed models, so for assessing the predictive performance, a number of *K*-fold cross-validations would be conducted to estimate prediction error.

(I) Parametric regression

In this section, various ways of regression (LS, Ridge, PCR, etc) are going to be applied to this dataset, and the prediction performance would be further studied. Various tests related to regression diagnostics would be conducted to ensure the validity of the chosen regression method. Due to the larger number of variables, I think one main challenging task would be effective subset selection.

(II) Non-parametric regression

In this section, *k* nearest neighbor (*k*NN) would be used as to predict the value of response variable. Although it is mainly used for classification, it can also be adapted for regression purpose. Specifically, the main focus of this section would be the selection of influencing factors, the definition of distance between observations, the best value of neighbor size *k*. A weighted distance is expected to be developed on this dataset to improve prediction capability.

(III) Classification

From the literature where the original data is collected and analyzed, we find the dependent variable can be converted to discrete variables (five levels from *I*- very good to *V*- insufficient) or (pass, fail) by grade converting system . So a groups of classification techniques (e.g., Decision tree, Neutral networks, SVM, *k*NN) can also be applied to obtain fitted model.

(4) The final prediction results would be compared and studied. The predictive performance for each modeling approach would be evaluated and discussed.

Appendix: Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:

31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)