Hwk #4: Regression and Classification
Biostat 202: Opportunities and Challenges of "Big Data"

For this homework we will use the "student-.csv" dataset available on the CLE

This data examines student achievement in mathematics at two secondary education Portuguese schools [1]. The data attributes include student grades, demographic, social and school related features. The data was collected using school reports and questionnaires.

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 (ordinal: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8 Fedu - father's education (ordinal: 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 (ordinal: 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 (ordinal: n if 1 <= n < 3, else 4, i.e., if n >= 4 failures = 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes taken in math (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 - in a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (ordinal: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (ordinal: from 1 - very low to 5 - very high)
26 goout - going out with friends (ordinal: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (ordinal: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (ordinal: from 1 - very low to 5 - very high)
29 health - current health status (ordinal: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

Target variables:
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)

We will focus on predicting student's final grades G3:

Data Preparation

1. (1 points). Read the dataset in from the course website. It is called student-mat.csv. Were there any issues that you encountered when reading in the data?
2. (1 point). Add a Derive Node after the Type Node to create a new binary (Flag) variable Pass3 that is equal to "pass" when G3 >= 10 and equal to "fail" when G3 < 10 – this derived variable will be used in the classification questions where the goal is to predict which children will have low scores so that they can receive interventional help (about the bottom 3rd).
3. (1 point). Use the Type node to appropriately set the Measurement type and Roles of all variables (based on above). Initially set G3 as the outcome ("Target") variable and set G1 and G2 to have the Role "None".
4. (1 point). Partition the data so that you have 50% training, 25% test, and 25% validation

Regression

5. (1 point) Add a Linear Node and build a model using "Best subsets" for Model Selection, and Best Subsets Selection with Criteria for entry/removal of "Information Criterion (AICC)". Otherwise use default settings.
6. (4 points) Describe the quality of the model fit based on the output of the golden nugget. What are the 3 most important predictors and what effect do they have on the outcome G3? (Note that this model represents the ideal goal of being able to predict student grades using only demographic variables.) Attach an Analysis Node and determine the Mean Absolute Error in what IBM Modeler calls the "Validation set" (this would be the test set in the conventional training-validation-test definition).
7. (3 points) Add in G2 as a predictor. Compare this model with the previous model when G2 was not included. Comment on the differences between the models (including the Model Summary and change in Mean Absolute Error). Why do you think the difference is so substantial?

Classification

8. (1 points). Go back into the Type node so that Pass3 can be set as the outcome flag variable and each of G3, G2, and G1 are set to have Role "None"
9. (2 points). Fit a support vector machine to the partitioned data. Use the default settings except that in the Analyze tab check "Calculate predictor importance". Do any predictors stand out?
10. (3 points). Add an Analysis Node. What is the prediction accuracy in the validation set? (Note that the validation set in Modeler is the final "test" set in conventional machine learning terminology.)
11. (3 points). Add G2 as a predictor and re-run the SVM. Look at the predictor importances. How far down the list does G2 come? What is the difference in prediction accuracy between this model and the one without G2 in the model? Are you surprised by this result?
12. (3 points). Repeat 9, 10, and 11 for the C5.0 classifier. Does C5.0 do better or worse than SVM in terms of validation set classification accuracy (with and without G2 as a predictor)?

Cross-validation

13. (5 points). Consider the following algorithm:
    a. Screen the predictors: perform separate analysis for each predictor to determine whether it is correlated with the class label. Select a subset of "good" predictors that show fairly strong correlation with the class labels (e.g. correlation > 0.8)
    b. Using just this subset of predictors, build a classifier.
    c. Use cross-validation to estimate the unknown tuning parameter of the classifier.
    d. Use cross-validation again to estimate the prediction error of the final model.

    Do you see a problem with this algorithm? If so, what would you need to do to fix it?

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.