**Helping Students Succeed**

**An analysis of behaviors that best predict improvement in secondary school**

In Portugal, the United States, or any country in the world it is vital to provide a solid education to the youth of that nation to ensure continued growth. In a very real way the success of the student mirrors the success of the country. As such there is a great interest in finding the best ways to improve education, even though one must overcome several difficulties in doing so.

One such difficulty is agreeing on how to define success. Attendance is easy to measure, but may be too simple to be useful. Grades are more descriptive but can be highly subjective from school to school. Even standardized test scores can vary wildly from student to student based on a large number of environmental, educational, and personal factors. Our goal in this analysis is to investigate these factors and identify which among them best identify success for the student, which we will define as how their test scores have improved from their first year of secondary school to their third year of secondary school.

**Problem Statement:**

Using test scores and other related data observed from Portuguese secondary school students and donated to the UCI Machine Learning Library we will subtract each student’s first year test scores from their third year test scores to find the increase (or decrease) in score for each student. Using this change as a response variable we will then apply an ordinal logistic regression technique to determine which groupings of explanatory variables best describe the change in score for each student. We will also utilize a stepwise selection technique to reduce the number of variables, and a decision tree analysis to help support or disprove the findings of the logistic regression. This should allow us to accurately predict which behaviors will provide the best improvement for each student regardless of what level they were at when first enrolling.

**Data Set Used:**

We will be analyzing two separate but similar datasets, one that includes the scores on a standardized Math test and another that scores a Portuguese language test. Each dataset makes use of the same 30 explanatory variables and 3 response variables. We will be modifying the dataset such that instead of having 3 response variables, one each for first year, second year, and third year grades, we will only use 1 response variable which will be equal to the third year test score minus the first year test score.

Below we will include the description of the 30 explanatory variables as they were provided by the researchers for the donated data set:  
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", "at\_home" or "other")

10. Fjob - father's job (nominal: "teacher", "health" care related, civil "services", "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)

As can be seen from the variable descriptions above, the researchers gathered a wide range of data on the personal lives of these students as well as their scholastic history. It is also notable that the dataset contains no continuous explanatory variables. Only binary, nominal, or ordinal (numeric) variables are present.

**Scope:**

This was an observational study, and as such any inference made as a result of this study can only be one of association and not causation. The subjects are each students of two specific schools, and while we can assume they were selected randomly within the school’s population there is nothing to indicate that the two schools themselves were randomly selected. Therefore any inference of association will be limited to the student population of the two schools that were chosen for the study.

**Exploratory Data Analysis:**

The two datasets contain 649 and 395 observations respectively, so there will be no issues with sample size for any of the analysis methods we will use. Furthermore, neither dataset contains any missing values so no manipulation is needed in that regard:

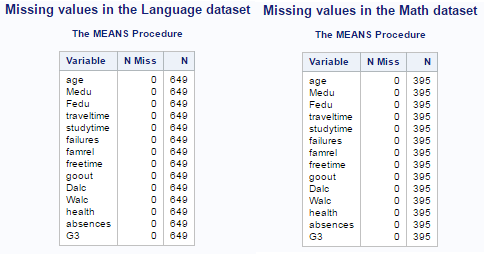


Figure 1: Checking for missing values

As mentioned previously there are no continuous variables used in either data set. Because of this standard multiple linear regression tools are inappropriate based upon the unmet assumptions of linear relationships and a normal distribution of Y for a given X. Neither logistic regression nor decision tree analysis requires these assumptions and so they will be more appropriate for the data.

The logistic regression method does, however, have an assumption of independence that will be violated if we use the first and second term scores, G1 and G2, as explanatory variables for a response variable of G3 because we cannot say that a student’s scores from one test taken earlier are independent from a similar test taken later. Indeed, a scatterplot of G1, G2, and G3 is suggestive of a positive linear relationship between how well you did on tests 1 and 2 with how well you did on test 3 as seen below:

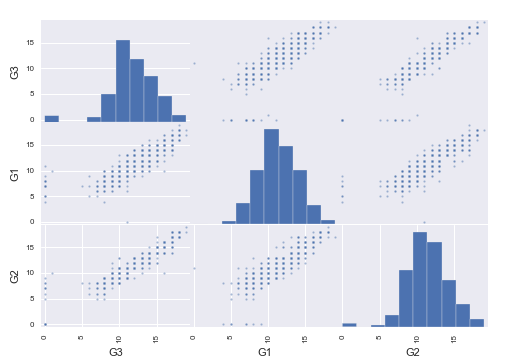


Figure 2: Scatterplot matrix of G1, G2, and G3 for the Portuguese data set

That being said, by not using G1 or G2 as explanatory variables and instead creating a new “GChange” response variable that represents the change from the first test to the third test (G3 – G1 = GChange) we will end up with a model that makes no violations against the assumption of independence.

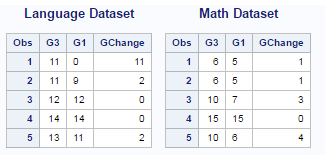


Figure 3: Creating the GChange response variable

Please note that additional detailed exploratory analysis is recorded in Appendix B

**Variable Selection:**

Both logistic regression analysis and especially decision tree analysis is more effective and more interpretable when we have a small but meaningful number of explanatory variables. In other words, a properly tuned model will include all of the available variables that are significant and none of the variables that are not significant. To accomplish this reduction we performed a stepwise selection technique which reduced our starting number of 30 explanatory variables down to only 4 for the language test, and only 2 for the math test as seen in the display below:

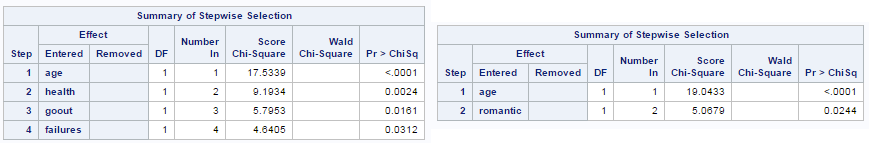
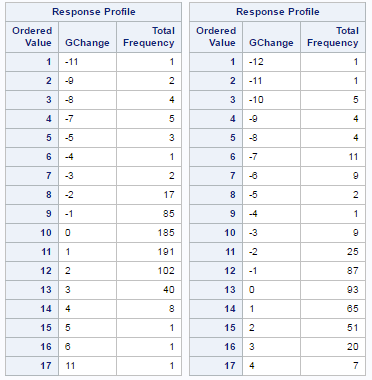


Figure 4: Logistic Regression Stepwise Selection Results, Language on the left and Math on the right.

The variables that were found to be statistically significant predictors for an increase in language scores are the age of the student, the student’s health, the number of times they failed the test in the past, and how often the student goes out with friends. The variables that were found to be statistically significant predictors for an increase in math scores are just the age of the student and whether they were in a romantic relationship at the time.

**Investigation of Outliers:**

While performing this stepwise selection we discovered that there were unusually high and low values of GChange reported in both datasets, as seen in the figure below:

The test scores range from 1 to 20, and as seen in the figure on the right we have the vast majority of the score increases falling between a decrease of 2 and an increase of 3. In fact, the distribution of these common changes appears to be approximately normal.

Some of the outlying values seem suspicious as it is unlikely that a normal student would see an increase or decrease as high as 12 points when the test scale is 1-20.

Figure 5: Response Profiles for Language (left) and Math (right)

Investigating these outliers further we find that there are student records in the data where either the first term test or the third term test (G1 or G3) has a reported value of 0. It is unknown whether these reported values of 0 are representations of missing data by the researchers, or if they are ‘legitimate’ scores in the sense that the student sat for the exam but did not actually make an attempt to complete any questions on the exam. Either way such scores are not valuable in providing insight into predicting increases in test scores.

We made the decision to limit our analysis to observed increases in test scores for students who made an attempt for both the first term test and the third term test. We then found and removed all rows for which there was a 0 value for either G1 or G3. This process removed 16 of the 649 observations on the language test, leaving 633 remaining. This process also removed 38 of the 395 observations on the math test, leaving 357 remaining. As seen in the figure below, this had a dramatic improvement in the response profiles for both tests, and only slightly changed the results of the stepwise selection:

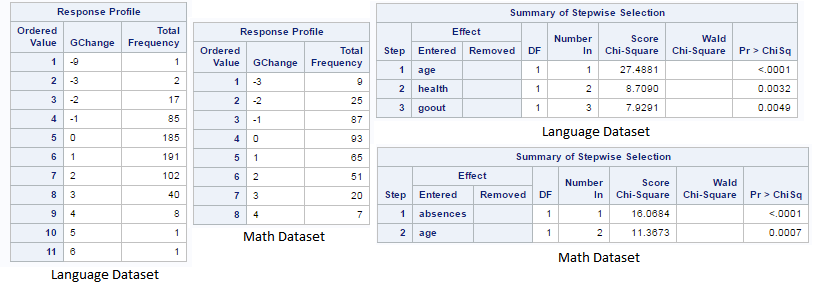


Figure 6: Updated Response Profiles and Stepwise Selection Results after removing 0 score tests

Specifically, after removing those outliers we see that the number of test failures is no longer included in the selected variables for the Language dataset, and rather than math scores being predicted by a student’s romantic status and age they are predicted by the number of their absences and age.

This is a significant reduction from 30 explanatory variables down to 2 or 3, but not entirely unexpected. Most of the variables the researchers measured might conceivably estimate how well the student scores on tests, but are either not expected to change between the first test taken and the last test taken or that change was not recorded. For example, the reason why a student chose to attend a particular school is not likely to change, and if the cohabitation status of the student’s parents changed between test G1 and test G3 it was not recorded.

Because of this we will accept the selected variables as they are now, and combine them together into a single model that will be run against both datasets for clarity of interpretation. This produces the following predictive model:

GChange = β0 + β1Age + β2Health + β3GoOut + β4Absences

**Logistic Regression Analysis:**

Using the model selected above we perform our logistic regression analysis on both datasets which results in the odds ratios displayed below:

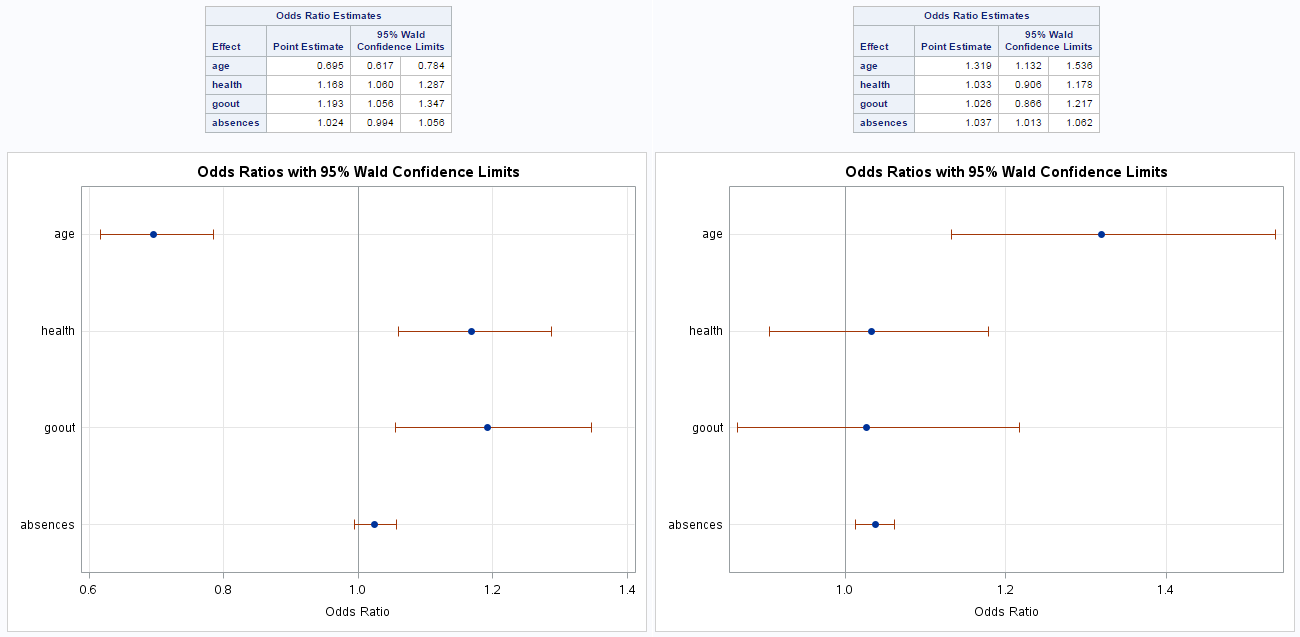


Figure 7: Odds Ratio Estimates and related confidence limits for Language (left) and Math (right)

Interestingly enough the two types of test paint very different pictures of what students are expected to improve their scores for it. It appears that a student who is younger and goes out often with their friends for healthy activity is most likely to increase their language score during their time at school. Meanwhile, a student who is older and dutifully attends all of their classes is more likely to increase their math score while in school. We will now check these findings against a decision tree analysis before coming to a formal conclusion.

**Decision Tree Analysis:**

When we check the results of the ordinal logistic regression compared to a decision tree analysis, we get very similar results with a few notable exceptions. Below are the visualized decision trees, with the language dataset on the left and the math dataset on the right:

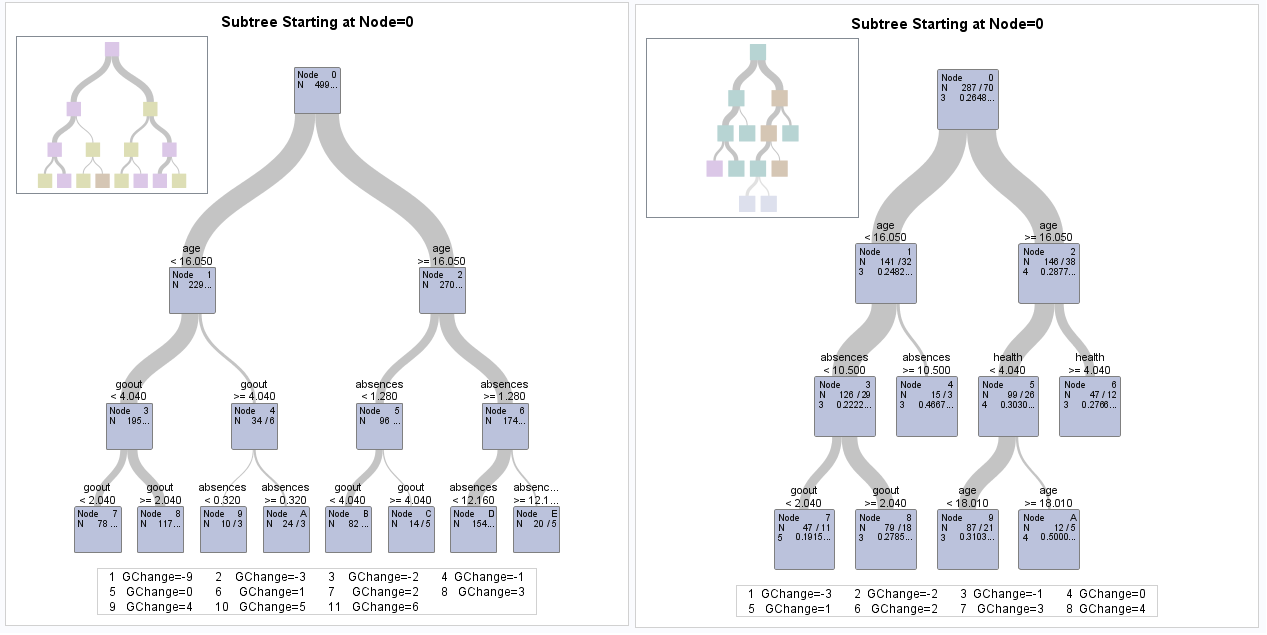


Figure 8: Decision trees for language (left) and Math (right)

As seen above, the major predictive decision points for the language test scores seem to come from how old the student is, followed by how often they attend class versus how often they go out with friends. This is slightly different from the logistic regression that suggested the health of the student was more predictive than their absenteeism. The decision tree for the math test, however, agrees much more with the logistic regression with age and absences being the most predictive factors followed by health and going out with friends as being less predictive.

We cross-validated the decision tree findings by an 80/20 split, with 80 percent of the observations in the training set and 20 percent of the observations in the validation set. As seen in the figure below, the cross validation found similar misclassification scores for both the training set and the validation set. This suggests no overfitting of the model to the training data.

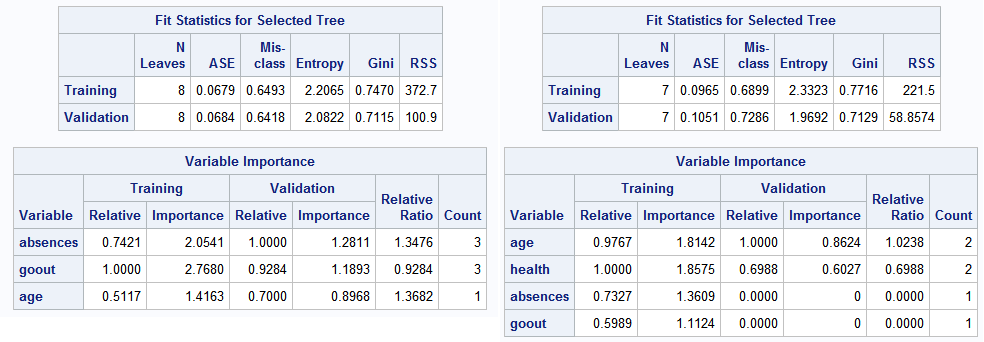


Figure 9: Fit Statistics for the Decision Tree analysis for Language (left) and Math (right)

That being said, although the misclassification rates are close between the training set and the validation set, neither score is very impressive with as many as 64.9 percent of language test scores being misclassified and as many as 72.9 percent of math test scores being misclassified.

In performing the decision tree analysis we found that it provided little evidence against the inferences suggested by the ordinal logistic regression analysis. We also found that the high misclassification rate and lack of additional interpretability made it an unappealing option to use in place of logistic regression. Therefore we will neither reject nor replace the logistic regression and proceed with drawing our final conclusions from the logistic regression analysis.

**Conclusion:**

When performing an ordinal logistic regression analysis we found that there is sufficient evidence at the alpha 0.05 level to suggest that age has a negative association while health and getting out of the house often both have positive associations with an increase in language test scores from term 1 to term 3. For the highest positive effect, going out, an odds ratio estimate is that getting out of the house increases the odds of an improved language score 1.193 times (with a 95% confidence interval between 1.056 times and 1.347 times).

The logistic regression analysis also found sufficient evidence at the alpha 0.05 level to suggest that both age and absences have positive associations with an increase in math test scores from term 1 to term 3. For the highest positive effect, age, an odds ratio estimate is that being an older student increases the odds of an improved math score 1.319 times (with a 95% confidence interval between 1.132 times and 1.536 times).

Based upon these data nations such as Portugal can best ensure their students see improved combined test scores while in secondary school by finding a balance between attending classes and getting out of the house for healthy social activities. Because this was an observational study we cannot say that attending class, being healthy, or being social causes an increase in scores, only that we observe an association. Furthermore, this association can only be inferred to apply to the students attending the two Portuguese secondary schools studied, and any inference made beyond that must be considered an extrapolation of the data.