Predicting Student Performance in a Portuguese Secondary Institution

Leonid Shpaner, Juliet Sieland-Harris, and Dan Choi

Leonid Shpaner, Juliet Sieland-Harris, and Dan Choi

2021-04-15

R Programming Language - Code (Appendix)

Abstract

Qualitative and quantitative factors alike affect student grades. We observed 1,044 students collectively from three terms of math and language arts classes of a Portuguese secondary institution to determine which of these factors is directly influenced by performance. Student grades were tallied over the three terms, from which performance was bisected by the median and binarized into two classes of 0 and 1 ("bad", "good", respectively). The dataset was further subjected to an 80:20 train-test split ratio to evaluate the model performance of data outside the training set visa vie implementation of six algorithms. The C5.0 and CART models produced accuracy scores of approximately 63%; whereas logistic regression and random forest models performed approximately 1% lower in terms of accuracy. Implementation of Naïve Bayes classification in conjunction with the neural network model, yielded more accurate results of 65% and 69%, respectively. We discuss other metrics like error rate and precision and note that each model, when cross-validated, has its own limitations that may inhibit or facilitate the prediction of student performance holistically.

Keywords: student performance, machine learning, ensemble methods, data mining

Predicting Student Performance in a Portuguese Secondary Institution

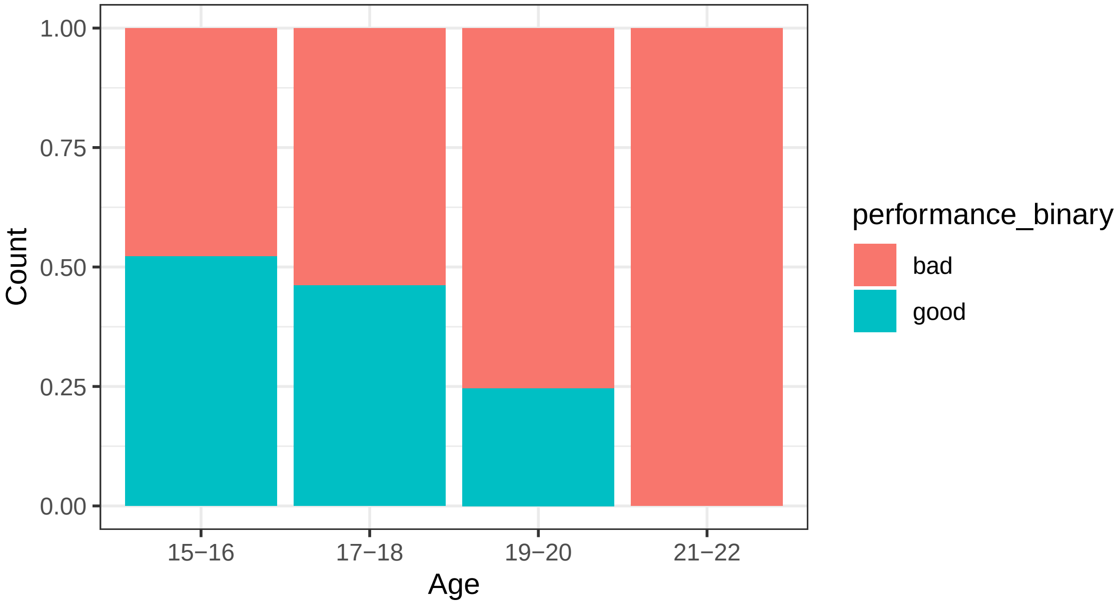
Predicting student performance closes the gap between socio-economic status and other external factors in one secondary educational institution, setting a precedent via proxy model for others to follow suit. The 2018 Program for International Student Assessment (PISA) found that "socio-economic status was a strong predictor of performance in reading, mathematics and science in Portugal. In Portugal, advantaged students outperformed disadvantaged students in reading by 95 score points in PISA 2018" (The Organisation for Economic Co-operation and Development [OECD], 2018). We aim to set a precedent for repeatability, allowing for subsequent iterations of our modeling techniques.

Methodology

Our goal was to measure the impact of socio-economic factors on student performance from which we would select a viable predictive model. We sourced our publicly available dataset from the UCI Machine Learning Repository (Dua, 2019) as two separate .csv files representing Math and Portuguese courses from schools of secondary education. We merged the two groups into one master "student_file.csv" file in R (R Core Team, 2020), ensuring that the dataset had no missing values; this provided us a total of 1,044 rows and 33 columns of observations. Rather than sampling the data, we chose the entire population for our ensuing exploratory data analysis, thereby allowing us to pre-emptively eliminate selection bias. Our preliminary review consisted of evaluating absences from school by age, relying on attendance to provide an initial read on overall student performance. Though we identified 480 outliers in absences, we persisted in building an inclusive model of performance; decidedly, school attendance was not a reliable target for performance. On the other hand, the following descriptive statistics provided important information about the overall distribution. Whereas the minimum student age was 15, the mean, and median were both approximated at 17, suggesting a normally distributed dataset. Performance itself was categorized into "good" if the students' grades were above the median, and "bad" if they were below. From here, we were able to determine performance by age and gender. Figure 1 shows the normalized age group by performance.

Figure 1

Age Group by Performance: ("Good" or "Bad") - Normalized



Note. This normalized histogram assuages the comparison of performance across age groups, attributing the highest grades (248 of them) to 15–16-year-olds.

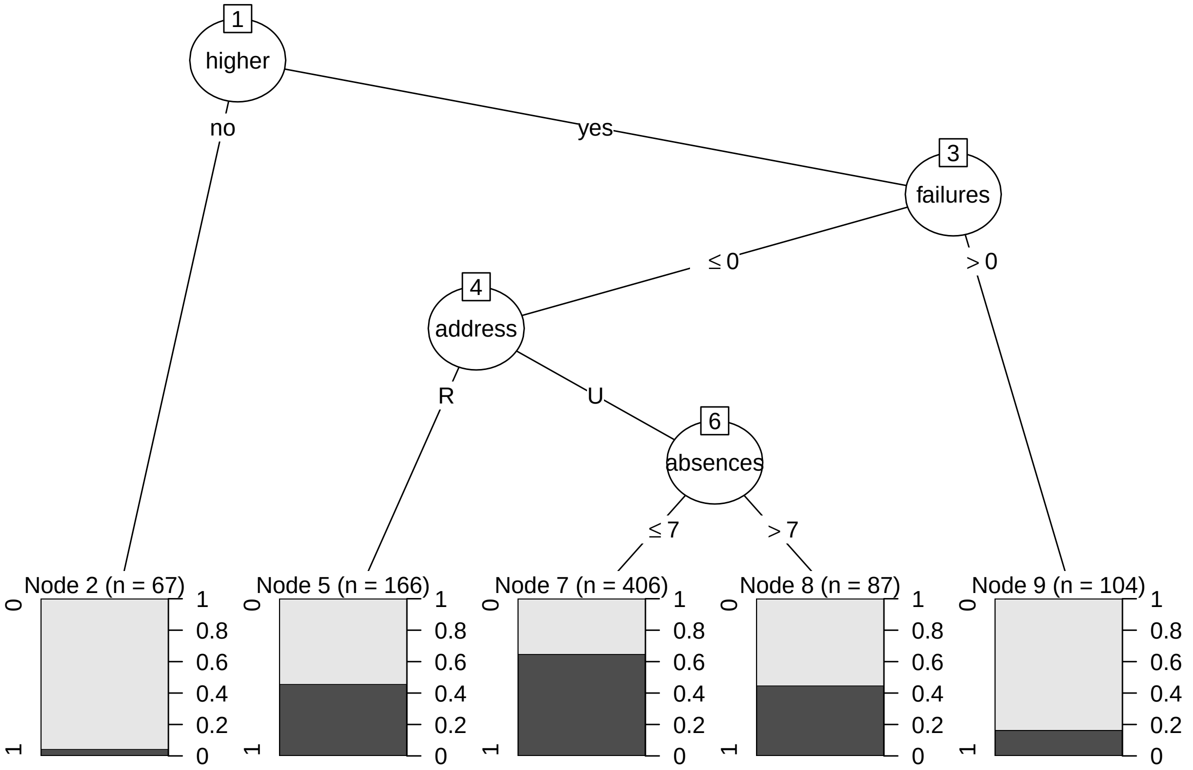
Six of the ensuing models were created with these predictor variables: address, family support, and study time. Nursery school, higher education, failures, and absences were presented as well for a total of seven. Logistic regression retained only the predictors of study time and absences.

C5.0

The C5.0 model only utilizes four of the variables for the four decision nodes. Figure 2 illustrates that the root node splits for whether the student is interested in higher education. Students who are not interested in higher education immediately terminate to the leaf node two which contains 67 records and has a low likelihood of performing well. The students interested in higher education branch to the next decision node which splits by students who have failed a class previously and students who have had no previous failures. Students who had at least one previous failure branch to leaf node nine, which represents 104 records with a low chance of performing well. The students who had not failed a class branch to the address decision node. If the student has a rural address, the branch terminates to leaf node five, which is comprised of 166 records and has a moderate likelihood of performing well. Students with urban addresses branch to the final decision node for absences. However, students with more than seven absences terminate at the leaf node (consisting of 87 records) where there is a moderate chance of performing well. Students with seven or less absences terminate at leaf node seven (406 records), the highest likelihood of performing well.

Figure 2

C5.0 Decision Tree Predicting Student Performance

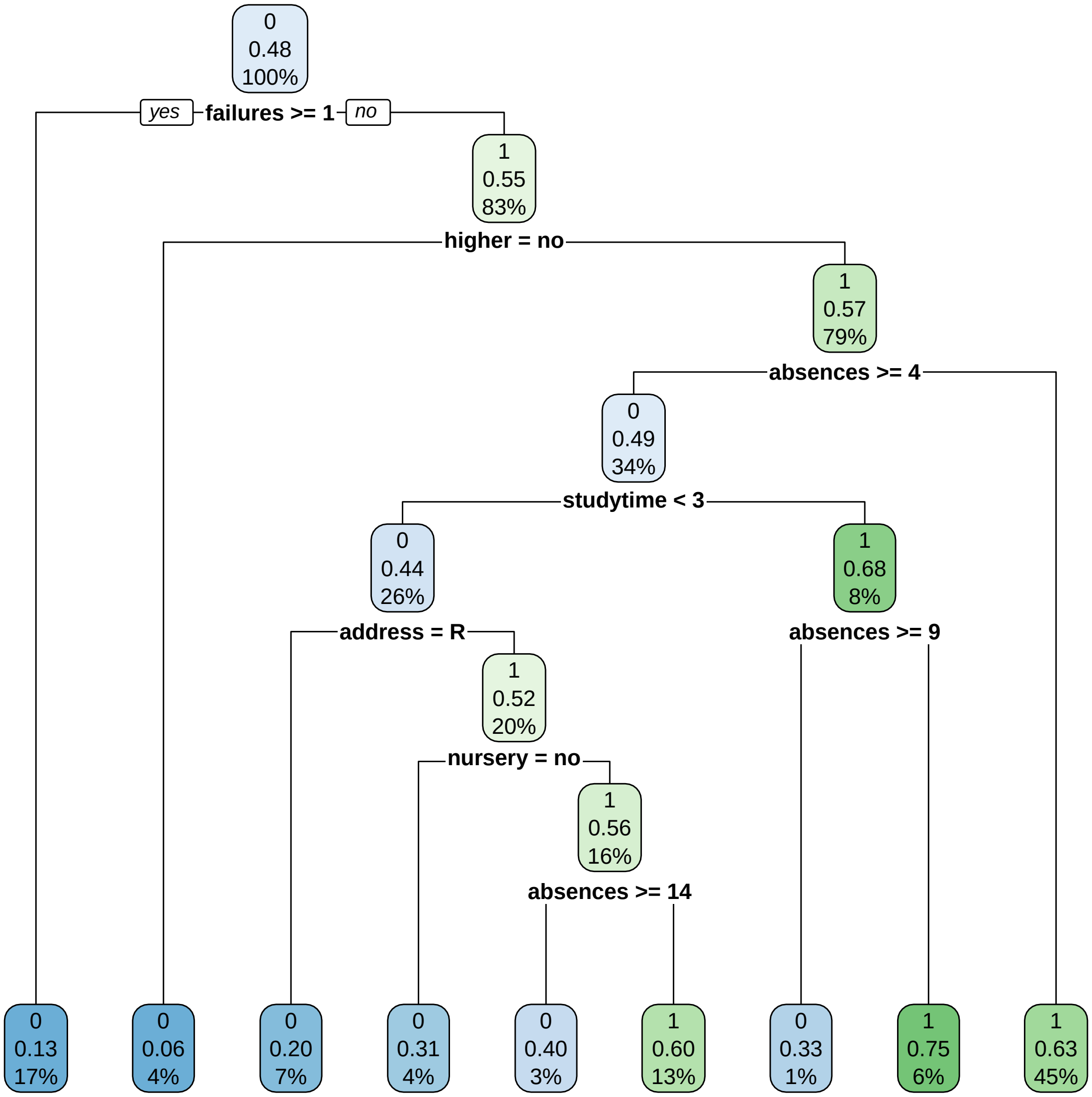


CART

The resulting decision tree (Figure 3) has one root node, seven decision nodes, and nine leaf nodes. The root node begins with the "failures" variable, immediately relegating students with one or more failures to the first leaf node. These students make up 17% of the data set; 13% have a high-performance value, strongly likening a student's previous class failures to continued poor performance. The next decision node splits based on student interest in higher education, with those not interested in higher education terminating to the second leaf node. They make up 4% of the data set; 6% of these students have a high-performance value, demonstrating that lack of interest in higher education is a strong indicator for likelihood of low performance. Another decision node splits students by address, with rural addresses terminating at the third leaf node. These students make up 7% of the data set and 20% of them have a high performance, making address another strong indicator of student success.

Figure 3

CART Decision Tree Predicting Student Performance



Note. After evaluating this model with the test data set, the accuracy is determined to be 63.55%.

Logistic Regression

Logistic Regression was selected as the regression model of choice for this study due to performance being a binary response. The logistic regression equation takes the parametric form of the logistic regression model:

$$p(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)} + \varepsilon$$

Study time and absences are used to predict performance using this descriptive form:

$$\hat{p}(performance) = \frac{\exp(b_0 + b_1(study\ time) + b_2(absences))}{1 + \exp(b_0 + b_1(study\ time) + b_2(absences))}$$

Plugging in the coefficients from our training data we have:

$$\hat{p}(performance) = \frac{\exp(-0.589 + 0.365(study\ time) - 0.052(absences))}{1 + \exp(-0.589 + 0.365(study\ time) - 0.052(absences))}$$