Modeling to Predict and Interpret Secondary School Student Performance

Sude Almus, Michael Li, Martin Lim, Carrie Wang, Cheyenne Kim

Secondary School Student Data

Introduction

A student's performance in school is oftentimes a result of numerous exterior factors that are beyond the individual's control. This can include their demographics, home life, personal life, extracurricular activities, etc. Performance is a combination of nurture and nature and understanding the impact of these external factors can give important insight into the current education system and potential areas of improvement to help support students today. These reasons are what drove us to investigate a dataset of Portugese secondary students: to observe and interpret which attributes of secondary school students attributed to higher or lower final grades in their math class. This would give a good proxy of our objective question, but also is limited by it being data from just two schools in Portugal.

Our data had 3 key achievement variables that we could look at: G1, G2. and G3. These were all the grades of the students for each grading period. We chose to focus on G3 for our modelling as it was the student's final grade in the class. G3 was measured on a scale of 0-20. In order to test a variety of modelling techniques, we followed the Portugese grading scale to convert G3 to a categorical ordered variable *cat_g3* while followed this conversion: 0-3.4 “Poor,” 3.5-9.4 “Weak,” 9.5-13.4 “Sufficient,” 13.5-15.4 “Good,” 15.5-17.4 “Very Good,” and 17.5-20 “Excellent”. Following these same guidelines, we also created a binary categorical variable *pf* where a "poor" or "weak" grade is considered a "low" grade. A "sufficient" "good" "very good" or "excellent" grade is considered a “high” grade.

Data/Wrangling

We did analysis on a dataset[1] of 328 secondary school students from two Portugese schools in their math class found on Kaggle[2]. The data attributes include student grades, demographic, social and school-related features, and it was collected by using school reports and questionnaires.

The main goal of our analysis was to build a model that would explain how variables impacted student achievement. The dataset originally had 33 distinct variables, with 16 numerical variables, 8 binary variables, and 9 character variables. A key part of the data collection was that multiple variables, such as famsup (family support) or freetime, were self-reported ratings on scales of 1-5. These were quantitative variables, but due to the

[1] Refer to the appendix for a comprehensive data dictionary

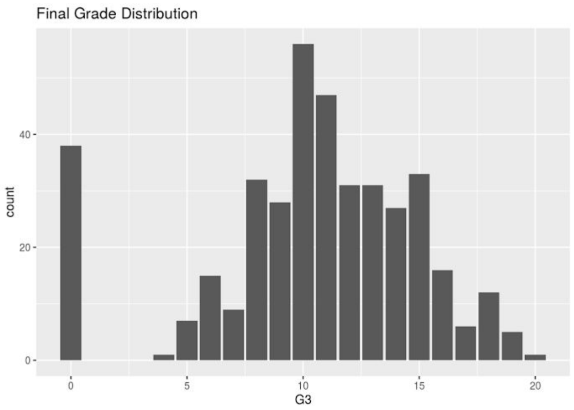
subjectivity of the self-report, we saw that it would be best to convert these types of variables into categorical variables for the sake of interpretability. Thus, scores of 1-3 would be factored as "low" and scores of 4-5 would be "high."

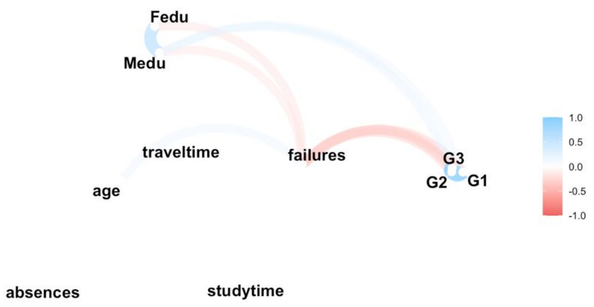
Additionally, we saw room to create informative variables based upon the given data that would provide more insight into our research question. We created the following variables:

1. *first_gen_college*: the given variables *Medu* and *Fedu* gave information about the student's parents education history. Using this, we created *first_gen_college* that indicates if the student would be a first generation college student if they decided to pursue higher education. This will give more tangible and clear insight to how parental education impacts student's performance.
2. *stable_learning_env*: if *famsup* is "yes" and *internet* is "yes", then *stable_learning_env* is "yes", otherwise "no"
3. *high_freq_absent*: if *absences* >= 10 for a student, we considered them a highly frequent student
4. *failed*: if *failures* > 0 then “yes” and "no" otherwise

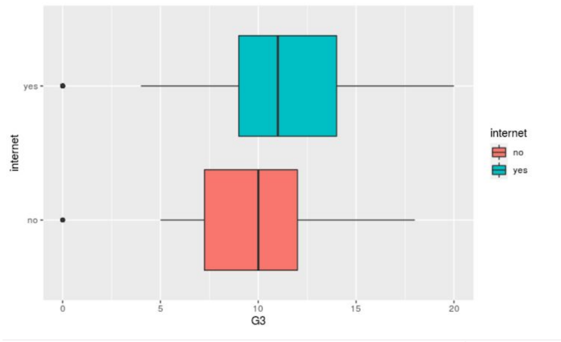
Exploratory Data Analysis

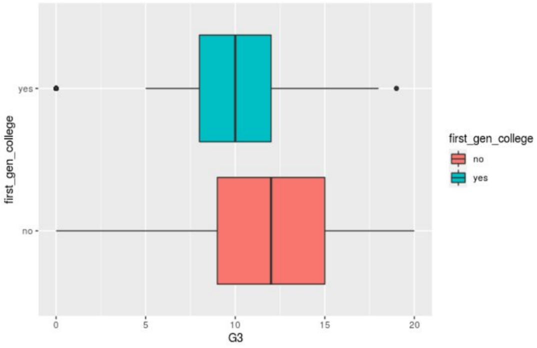
First, we looked at the distribution of our response variable, G3 (final grade). There are a large number of observations that are 0, but the rest of the observations seem pretty normally distributed, alleviating concerns that transformations would be necessary for regression. Next, we fit a correlation network of the entire feature space. As the plot shows, the vast majority of potential predictors are uncorrelated and thus do not even meet the threshold for inclusion in the plot. While modelling, we ensured that we examined potential multicollinearity issues between weakly correlated variables.

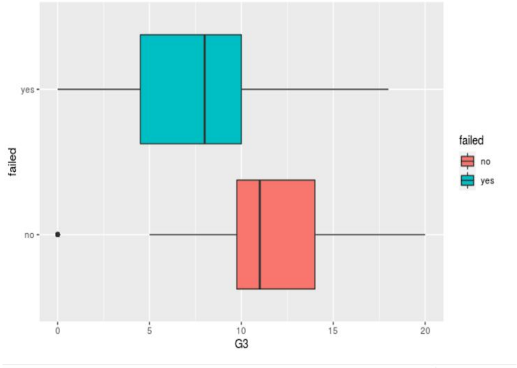


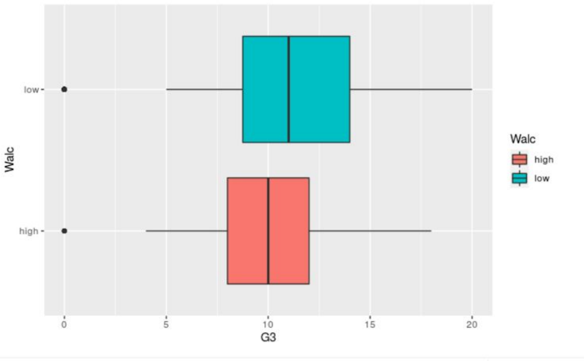


We also fit a few bivariate plots of key variables we thought would have impact on G3, the response variable. From our EDA, we found a few key possible trends. Students who had at least one of the following traits: failed a class previously, without internet, were frequent drinkers on the weekend, and were first generation students, on average had lower final grades than their counterparts.









Modeling

In our modeling, we first approached final grades as a continuous, quantitative variable. However, this posed several problems with data irregularities, poor model fit with a variety of techniques, and poor predictive performance. As discussed above, this motivated our creation of categorical variables based on Portuguese education system's own classifications. Our approach was to fit a variety of models for each of these three responses. We fit the models on a randomly selected training set consisting of 80% of the data, and tested the models on the remaining 20% to evaluate predictive performance.

Regression

Model 1: Linear Regression

The first model that we tried to fit was a linear regression model with all base variables with a continuous final grade variable as the response. We examined multicollinearity, performed stepwise selection, and added significant interaction terms to determine the predictors. We fit an expanded model based on the model chosen by two-way stepwise selection and interaction effects. From that, we fit a final model with the active base variables and significant interactions that allowed us to reach the best fit and lowest test error.

G3 ~ first_gen_college + stable_learning_env + failures* absences + first_gen_college * failures + absences * stable_learning_env + schoolsup * absences + schoolsup * first_gen_college + sex * first_gen_college + sex * failures + studytime * schoolsup

Multiple R-squared: 0.3049, Adjusted R-squared: 0.2701, AIC: 1792.389, Test MSE: 16.71672

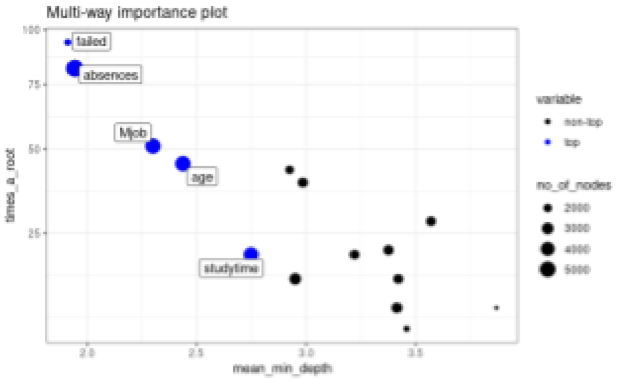
According to this model, the most significant predictors were first_gen_college, stable_learning_env, and the interactions between sex*first_gen_college, studytime*schoolsup, failures*absences, failures*first_gen_college. However, the low adjusted R-squared value of 0.27 indicates a very poor fit to the data. With a test MSE of 16.72, it is clear that the predictive performance was also poor. The linear regression model did allow us to observe some interesting interactions between some variables. For example, the interaction between first_gen_college and schoolsup shows that being a first generation college student and receiving extra educational support has a +5.25 effect on final grades. Another significant interaction is between first_gen_college and failures, which shows us that being a first generation college student and having previous class failures has a +2.34 effect on final grades for each failure. While these relationships may be interesting to try and understand, the poor predictive power and the inability to control the variables in these interactions make these insights less practical.

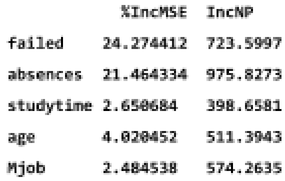
Model 2: Random Regression Forests

The linear model did not seem a good fit for the data and exhibited poor predictive performance. We chose to create a random forest model with a continuous G3 as the response variable, as it is robust to outliers, fits well on non-linear data, and generally displays higher accuracy. Our initial model included all base variables with some removed to address multicollinearity and variables of interest (for example, removing Medu and Fedu, which were re-coded as first_gen_college, as we are more interested in the latter variable). We fit successive pared-down models, as our primary concern is inference, by removing the variables with lowest importance. All random forest models were trained using 500 trees and 3 predictors as candidates at each split. We were able to achieve the best fit and predictive performance with the following model:

G3 ~ failed + absences + schoolsup + first_gen_college + age + studytime + Pstatus + famsize + guardian + freetime + Mjob + romantic + paid + sex + goout

This model explained 30.49% of the variation in the data and had a mean of squared residuals of 15.28, which indicates a poor fit. The mean squared error for the testing set was 14.39569, which indicates an improvement in predictive performance as compared to the linear model. This model examines variable importance with two primary measures: 1. The percent increase in mean squared error after the variable is permuted (%IncMSE) 2. mean increase in node purity by splits on the variable (IncNP). We have displayed the 5 strongest predictors according to these measures in the following figures.





```
           %IncMSE   IncNP
failed     24.274412 723.5997
absences   21.464334 975.8273
studytime  2.650684  398.6581
age        4.020452  511.3943
Mjob       2.484538  574.2635
```

The insights about relationships between variables derived from such a model may not be very useful. However, we would like to note that *failed* and *absences* were significantly more important than any other variable according to both %IncMSE and IncNP values. This makes sense; a student who has previously failed a class likely displays a pattern of subpar academic performance, and is likely to perform worse than his or her peers. A high number of absences would greatly impact the student's comprehension of the class material, likely resulting in lower final grades.

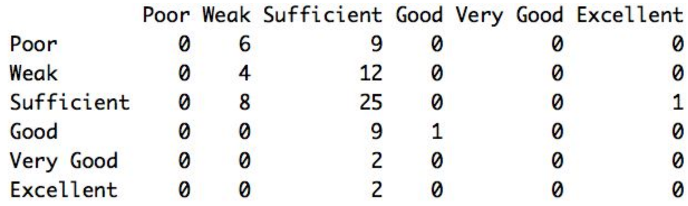
Six-Level Classification

Model 3: Ordinal Logistic

As mentioned previously, we created an ordinal categorical variable *cat_g3* that partitioned the continuous dependent variable G3 into six categories, turning this into a classification problem. Given that the values of *cat_g3* are fully specified by the values of *G3*, we opted to fit the ordinal logistic regression model using the same predictors and interactions as the simple linear regression (Model 1). From there, insignificant variables and interactions were removed according to their p-values and t-values; if the confidence interval for a given coefficient contained zero, we omitted the variable. The final ordinal logistic regression model (AIC: 892.6798) had the following form:

polr(cat_g3 ~ failed + absences + goout + sex + guardian + romantic + schoolsup + first_gen_college + schoolsup * studytime + schoolsup * first_gen_college)

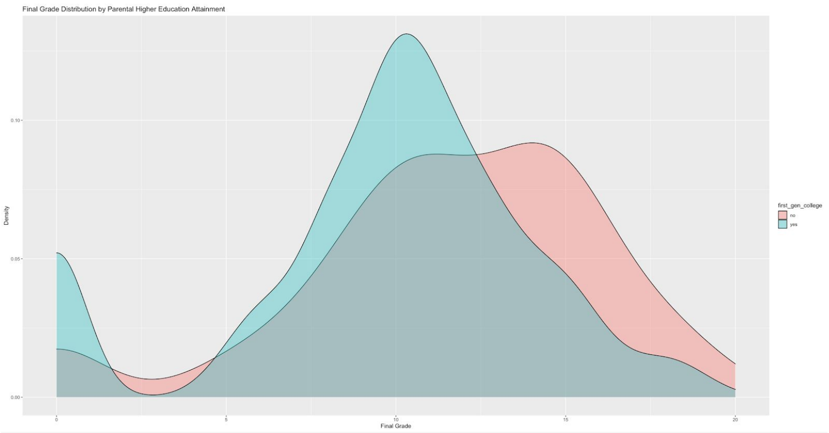
The fitted model had several undesirable features. After being trained on a portion of the complete dataset, the model achieved a classification accuracy of 37.97% on the test set, as can be seen in the confusion matrix on the right. More problematic than the classification accuracy itself was the tendency to disproportionately assign predictions to the Sufficient category due to the prevalence of these observations within the data. Consequently, we believed that variable selection approaches would prove inadequate, given a more fundamental problem with the use of the cumulative logit model in this form. We decided to build off the ordinal logistic model by relaxing a key assumption and allowing a greater flexibility in our final model.



```
          Poor Weak Sufficient Good Very Good Excellent
Poor         0    6          9    0         0         0
Weak         0    4         12    0         0         0
Sufficient   0    8         25    0         0         1
Good         0    0          9    1         0         0
Very Good    0    0          2    0         0         0
Excellent    0    0          2    0         0         0
```

Model 4: Partial Proportional Odds Model

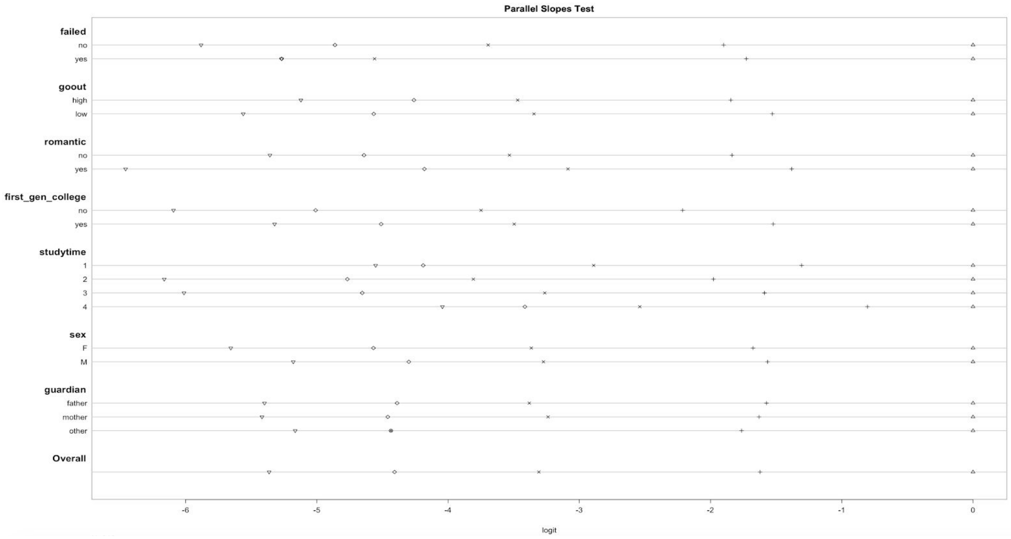
The primary concern with the ordinal logistic regression model was the proportional odds assumption for the entire feature space. Put simply, the ordinal logistic regression assumes that the coefficients that describe the relationship between the lowest ordinal category versus all higher categories are the same as those that describe the relationship between the next lowest category and all higher categories, and so on. This assumption is likely to be false for several important



Final Grade Distribution by Parental Higher Education Attainment predictors in the model. Considering the *first_gen_college* explanatory variable plotted above, it is evident that students who had at least one parent attend an institution of higher education tend to achieve higher final grades than students whose parents have not. Given this, we decided to use a partial proportional odds model of the form below. This model allows certain predictors to have coefficients that are the same for all response values *j*, while other predictors— represented here by X3— are able to have coefficients that vary depending on the value of the response. This should improve classification accuracy, as it affords the model greater flexibility.

$$P(Y_i > j) = \frac{\exp\left(\alpha_j + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_{3,j} X_{3,i}\right)}{1 + \exp\left(\alpha_j + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_{3,j} X_{3,i}\right)}, j = 1, 2, \dots, M-1$$

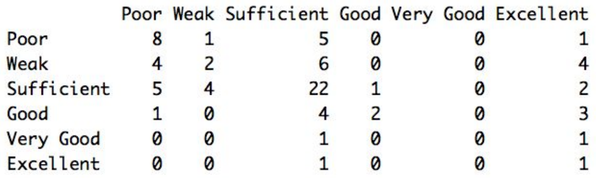
To determine which variables satisfy the proportional odds assumption, we created a function to calculate the log odds of *cat_g3* being greater than or equal to a given level. This effectively allowed us to evaluate the parallel slopes assumption by checking the equality of coefficients for a series of binary logistic regression across varying cutpoints on the dependent variable. We would expect the differences between each of the estimates for different levels of *cat_g3* to be similar and independent of the predictor's value.



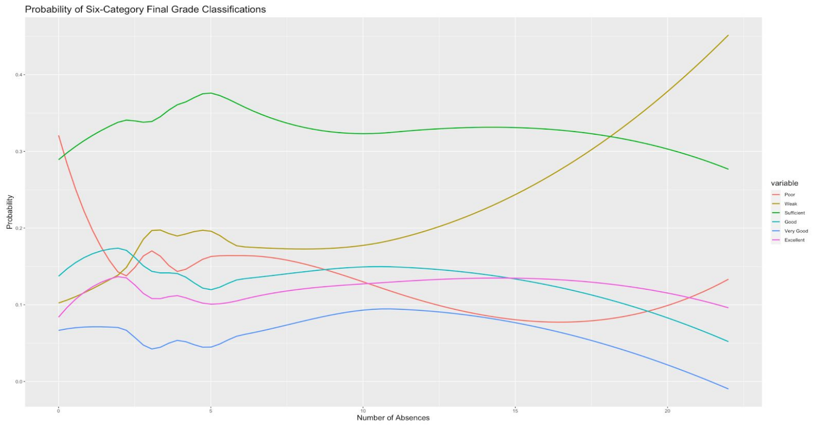
Analyzing the above plot which shows the differences in log odds estimates compared to the baseline case when *cat_g3 ≥ Weak (2)*, it is clear that the assumption is only reasonably upheld for *goout, sex,* and *guardian*. We thus fit the following vector generalized linear model, applying the proportional odds assumption to the aforementioned variables and allowing nonparallelism for the remaining variables:

> ***vglm(cat_g3 ~ failed + absences + goout + sex + guardian + romantic + schoolsup + first_gen_college + schoolsup * studytime + schoolsup * first_gen_college, family = cumulative(parallel=TRUE~goout + sex + guardian))***

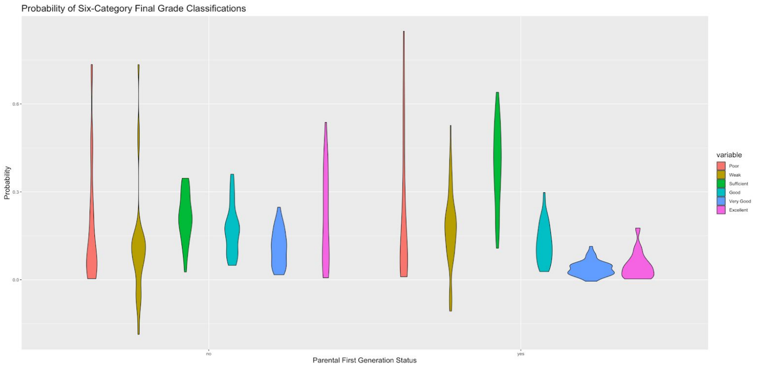
Fit on the same training set and evaluated on the same testing set as the ordinal logistic model, the partial proportional odds model achieved an improved test classification accuracy of 44.30%, which is relatively impressive given the noisiness of the data. Notably, even as the model does misclassify slightly more data points that carry the label Sufficient and Weak in the testing set, it is much improved at not over-predicting the proportion of data under the Sufficient label. This offers evidence that the greater flexibility afforded by relaxing the proportional odds assumption for certain variables resulted in a more robust model for identifying trends among the variables for inference purposes.



```
           Poor Weak Sufficient Good Very Good Excellent
Poor          8    1          5    0         0         1
Weak          4    2          6    0         0         4
Sufficient    5    4         22    1         0         2
Good          1    0          4    2         0         3
Very Good     0    0          1    0         0         1
Excellent     0    0          1    0         0         1
```



Plotting the various class probabilities as a function of our predictors reveals several key insights and two have been highlighted below. Looking first at the relationship between a student's number of absences and their final grade, there are some intuitive results and other more surprising ones. The probability of having a Weak performance, shown by the gold line, appears to increase in an exponential manner with absences, indicating that as a student misses more and more school, the probability that they will not perform well shoots up. The probability of having Very Good or Excellent scores generally decreases with an increasing number of absences. However, the Poor and Sufficient curves display interesting and perhaps unintuitive behavior. The probability of having a Poor final grade actually decreases rapidly with an initial increase in absences from zero and generally tends to decrease at a slower rate for large values of absences. The initial highly negative relationship between absences and probability of a Poor grade may indicate that higher aptitude students can afford to miss several days of school during the year. If a student has no or very few absences, this could possibly be a reflection of their inability to understand the material and a subsequent need to attend school consistently. In the case of the Sufficient category, it is evident that absences have a relatively indeterminate relationship with the probability of scoring in this category. This suggests that for middle-of-the-road-performing students, other factors may be much more important in explaining their performances than the number of absences.

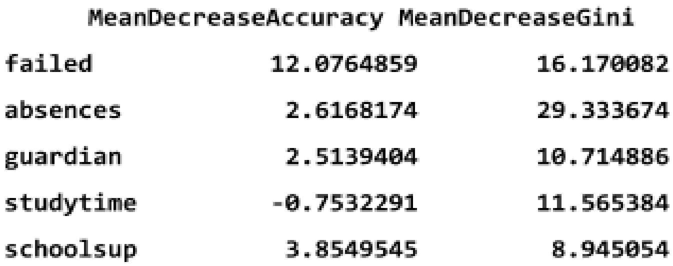

Probability of Six-Category Final Grade Classifications

The second crucial insight is the importance of the parental attainment of higher education. A value of ‘yes’ indicates that a student does not have a parent who attended university. The distributions of potentially first-generation college students who obtained final scores belonging to the higher three tiers of Good, Very Good, and Excellent are noticeably concentrated more closely around lower probabilities than those of their counterparts who have more educated parents. This seems reasonable: academic performance is likely to be emphasized more in highly educated households and the educational attainment of parents likely proxies for a measure of socioeconomic status, as wages tend to rise with years of education. Students with college-educated parents are likely to have greater support and access to resources promoting academic growth. Yet, having more educated parents is not a guarantee of student success. Potential first-generation college students are more likely to achieve a score that is “Sufficient”, as evidenced by the fact that the green distribution for this group is shifted upward substantially. Additionally, the distributions for Poor and Weak scores across the classes of this predictor are reasonably similar. The combination of these patterns may hint at another lurking relationship: the spoiling of children in educated households. While generally there appears to be a positive relationship between having a college-educated parent and a student’s own educational success, the non-trivial existence of students who come from educated households but do not have strong final scores lends credence to the idea that they may not have been pushed toward achieving academic excellence.

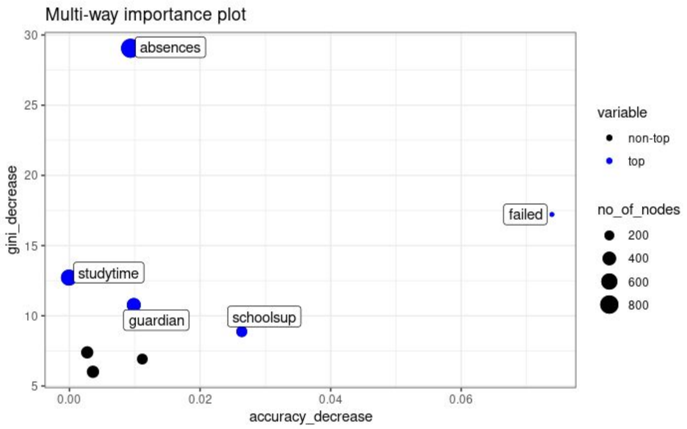
Binary Classification

Model 5: Random Forest

For “low”-”high” classification, our primary goal was improving predictive performance and getting a clear picture of important variables. This is why we focused on tree-based methods such as decision trees, bagging, and random forest. We chose to proceed with random forest modeling due to its greater interpretability, higher accuracy, and clear indication of variable importance. Our initial model included all base variables, and we performed selection by successively removing the least important predictors until we achieved a model with the greatest accuracy rates. All random forest models were trained using 300 trees and 3 candidate predictors at each split.



```
          MeanDecreaseAccuracy MeanDecreaseGini
failed              12.0764859        16.170082
absences             2.6168174        29.333674
guardian             2.5139404        10.714886
studytime           -0.7532291        11.565384
schoolsup            3.8549545         8.945054
```



The best-performing model is as follows:

pf ~ failed + absences + guardian + studytime + goout + schoolsup + first_gen_college + Walc + famsup

On the training data set, the model’s prediction accuracy was 90.19%, indicating a good fit to the data. Its predictive performance was also much higher than any previous model: it correctly classified 70.89% of the test data.

As mentioned, one of the key advantages of the random forest model is the ability to examine the relative importance of predictors. In the classification setting, this is done by examining two importance measures: 1. mean decrease in prediction accuracy after that variable is permuted (MeanDecreaseAccuracy) 2. Mean decrease in the Gini index, which measures node impurity by splits on that variable (MeanDecreaseGini). Higher values for each indicate greater importance. We have chosen to display the five most important variables in the model. We can derive a few key insights from these variables. *Absences* and *failed* seem to greatly affect students’ grade classification as “low” or “high”. This makes sense--a student who is more frequently absent misses a larger amount of class material, which should lead to lower performance. A student who has failed a class before likely displays a pattern of low academic performance. *Guardian, studytime,* and *schoolsup* have moderately high impact on classification. Students who spend more time studying are more likely to receive high final grades. If a student received educational support from the school, they likely display a pattern of performing poorly in previous classes; they would be more likely to receive a low grade again. It is intriguing to see that students' type of guardian (mother, father, or other) would be an important factor in their performance. Perhaps this indicates stability in the family environment, which could have an impact on performance. This certainly deserves to be explored further.

Results

Our final model selection is informed by inference in the context of the problem as well as model fit and accuracy. In terms of inference, a student's particular numerical grade is not as important; one's score can often increase or decrease by small margins without impacting the student's true understanding of the academic material. The 6-category student grade not only conforms to Portuguese national standards of academic performance, it is a more reliable indicator of student performance, allows us to understand general trends that affect student success, and would allow for extrapolation to a wider context. The two-category response variable that labeled grades as “low” or “high” allowed us to develop a model with the best fit and highest accuracy rate. However, this model only classifies a student as receiving a low or high grade; while its accuracy is high, it poses a danger of oversimplification. We are most interested in trends *across* categories. This is why we are much more concerned with understanding which factors affect a student's math grade classification as “Poor”, “Weak”, “Sufficient”, “Good”, “Very Good”, and “Excellent” and predicting this classification.

Based on this reasoning, our final model is the partial proportional odds model specified previously. The form of the model to allow coefficients to depend on the value of the response variable if the proportional odds assumption is not satisfied makes it the most justifiable and theoretically sound classification model. Additionally, it allows for improved predictive power in comparison to the original ordinal logistic model and inference on individual variables due to its generally additive nature.

Conclusion & Limitations

Our goal throughout this project was to identify key factors that impact students' academic performance, understand these particular impacts, and translate these into actionable steps on part of school policymakers. The particular model chosen is less significant than the relationships we were able to glean through the variety of modeling approaches and the three settings of students' final grades in mathematics.

1. Most of our strong models indicated a strong relationship between absences and final grades. Generally, increase in absences results in increased probability of lower performances. However, according to our final partial proportional odds model, it seems that for middle-of-the-road-performing students, other factors may be much more important in explaining their performances than the number of absences, and that higher aptitude students can afford to miss several days of school during the year. Absences can and should be addressed by school policy, but simply penalizing absences would not be productive. A more nuanced approach is necessary.
2. Parents' attainment of higher education seems to be an important factor in student performance. First-generation college students are less likely to score in the higher three tiers of Good, Very Good, and Excellent than their counterparts. However, the distributions for Poor and Weak scores across the classes of this predictor are similar, indicating that other factors may have a greater impact for students performing poorly.
3. Failing previous classes seems to have a very large impact on the final grade in all of our models. Previous failures are associated with increased likelihood of poor final grades. This makes sense; previous poor academic performance indicated future poor academic performance. Luckily, this factor lends itself well to policy: the school should consider instituting remedial programs or further educational support for students who have previously performed poorly in their classes.

We urge school administrators and policymakers to take steps to improve academic performance. Based on these key factors, we would first suggest addressing previous poor performance: for example, the school may institute remedial programs and mandatory tutoring for students who have failed classes previously. Administering harsh penalties on absences is not advisable, but the school may consider incentivizing attendance as well as addressing absences non-punitively: perhaps requiring meetings with counselors after a certain threshold of absences is met. Regarding the conclusion about parental education, essentially a correlate of socioeconomic status, schools and administrators may choose to offer resources or support to those with lower family income, such as free textbooks for those who cannot afford them, or free lunches so that students experience less stress about getting their next meal and can therefore focus on their education.

An obvious limitation of our analysis is extrapolation to a wider context. The dataset contains data from only two schools in Portugal; performance could also be impacted by school quality or social context. Our analysis is very useful for school policymakers who would like to improve student performance at their particular schools. However, we cannot generalize the trends to the Portuguese education system, and we certainly cannot generalize the trends internationally. Our response variables also only measure performance in mathematics. Due to its limited scope, it is not a comprehensive measure of academic success. This exploration, then, gives us an idea of what factors could impact academic success overall. However, to truly understand what impacts success, we would need much more comprehensive data.

Another limitation, due to the fact that we chose to prioritize the interpretability and applicability of our results over predictive power, is the relatively low test accuracy rate. However, due to the nature of the maximum likelihood classifier, this remains expected, as there are often categories that no observations are classified as; in our case, this happened with the "Good" classification. Though rather simplistic, our binary random forest model determining past-fail achieves higher predictive performance, with a 0.72 test set accuracy, though this is to be expected with the lower number of classes. The higher accuracy for the test set hints that if predictive power were to be more thoroughly explored for the pass-fail binary, there may be an even stronger model. Such a model could be used to individually target students who may fail and give them guidance ahead of time.

Appendix

1. Data Dictionary

Variable	Type	Description
school	character	School’s initials
sex	binary	Gender of student (‘M’ - male, ‘F’ - female)
age	integer	Student’s age
address	binary	Student’s home address type (urban or rural)
famsize	binary	Student’s family size (<=3 or >3)
Pstatus	binary	Parent’s cohabitation status (alone or together)
Medu	integer	Mother’s education level (0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)
Fedu	integer	Father’s education level (0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)
Mjob	character	Mother’s job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	character	Father’s job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	character	Reason for choosing school (close to 'home', school 'reputation', 'course' preference or 'other')
guardian	character	Student’s guardian (‘mother’, ‘father’ or ‘other’)
traveltime	integer	Student’s time to travel from home to school (1 - &lt;15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - &gt;1 hour)
studytime	integer	Weekly study time (1 - &lt;2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - &gt;10 hours)
failures	integer	Number of classes failed previously (0, 1, 2, 3, 4+)
schoolsup	binary	Receiving extra educational support (yes or no)

famsup	binary	Receiving family educational support (yes or no)
paid	binary	Taking extra paid classes in the subject (yes or no)
activities	binary	Involvement in extracurricular activities (yes or no)
nursery	binary	Attended nursery school (yes or no)
higher	binary	Wishes to pursue higher education (yes or no)
internet	binary	Has internet access (yes or no)
romantic	binary	In a romantic relationship (yes or no)
famrel	integer	Quality of family relationships (1 - very bad to 5 - excellent)
freetime	integer	Amount of freetime after school (1 -very low to 5 - very high)
goout	integer	Frequency of going out with friends (1 - very low to 5 - very high)
Dalc	integer	Workday alcohol consumption (1 - very low to 5 - very high
Walc	integer	Weekend alcohol consumption (1 - very low to 5 - very high)
health	integer	Current health status (1 - very bad to 5 - very good)
absences	integer	Number of school absences (0 to 93)
G1	integer	First period grade (1 to 20)
G2	integer	Second period grade (1 to 20)
G3	integer	Final grade (1 to 20)

2. Data source:

While obtained through Kaggle, the data is collected from the following study:

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.

https://www.kaggle.com/uciml/student-alcohol-consumption