Modern Statistical Computing

Final Project, Second Trimester 2023

Pass or Fail - Predicting Nerds and Losers

Submitted by:
Martin Blasi and Jonas Wallstein

Supervisor:
David Rossell

Universitat Pompeu Fabra
Barcelona, March 23, 2023

1 Introduction

What makes a successful student? This question baffles students, teachers and policy-makers alike. Students as well as teachers want to know what can make them improve. As researchers we want to find out how much room for improvement there is or if pre-determined or exogenous factors explain most of students' test results. Insights regarding factors that determine good education is of high interest to policy-makers. On the one hand, education is a major influence of economic growth. On the other hand, obtaining a good education can change peoples lives and support social cohesion. To get behind the factors determining educational outcomes, we study a rich data set of student performance that has been obtained by other researchers. We consider different models to fit the data in the most adequate way. Our linear regression model obtains the most consistent estimates for the most important factors explaining student performance. Furthermore, we also build different prediction models for students passing/failing a course and are able to predict with up to 93% accuracy which student is going to fail the course. The remainder of this report is organized as follows: First we explain the source and structure of our data. Then, we explain our different models and the results we obtain. Finally, we conclude our report.

2 Data

We use student performance data from the secondary education level in Portugal. The datasets were compiled for Cortez and Silva (2008) and publically available at the UCL machine learning repository.

2.1 Data Collection

The secondary grade in Portugal is made up of the final three years of school, after 9 years of basic education, before students can go on to obtain higher education. Cortez and Silva gather data on students and their performance in the school year of 2005/06 in two steps. First, they use school reports to obtain students' grades and the number of absences. Second, they run a survey over the observed students to obtain several demographic (e.g. mother's education, family income), social/emotional (e.g. alcohol consumption) and school related (e.g. number of past class failures) variables that they felt were expected to affect student performance. Socioeconomic status, parental involvement, student motivation, student health and peer group influence are indeed generally believed to be important predictors of student success. It should be noted that quality of teaching could not be measured, but usually is considered an important factor for student success as well.

Cortez and Silva end up with two datasets that both feature 33 variables, one containing test scores in Mathematics (395 observations) and the other test scores in Portuguese (649 observations). As the data was already processed and tidy, no further reprocessing was necessary. For the full list of variables please refer to our Appendix. Otherwise, we provide summary statistics of the most important covariates in the following section.

2.2 Grading System

The grading system in Portugal's secondary education features grades that range from 0 to 20. A student fails the course if the final grade (G3) is a 9 or lower. Students can get a final grade of 0, however, this only happens in extreme cases (e.g. the student skips multiple classes without an excuse). The school year in Portugal is split in three parts. Consequently, next to the final grade, we also observe grades at the end of the first two terms (G1 and G2) that represent the student's status up until the given point. One should note that all the grades that we observe are given as integer values. Whether this is due to rounding or just the nature in which the grades are stated could not be found out by us.

3 Descriptive analysis

The main thing we want to achieve in our report is to get a better understanding of what predicts student outcomes the best. For that reason we consider first, how the most important variables (according to significance in our regression analysis) are distributed. Furthermore, we consider interesting patterns in the grades from which we draw insights about certain sub-samples of students.

3.1 Sample Frequencies

The following graphic shows histograms for the most important variables in our regression analysis. Only the sample distributions for Mathematics are reported as the distributions for the Portuguese grades dataset are very similar.

Regarding the student's family background one can see that most students have fairly good family relations. Ranking at 1 or 2 on the family relations scale is rather an exception. At this point, one should consider that this is not neutrally observed data, but obtained from a survey. Consequently, what students said about their family relations could be tilted towards being more positive than in reality.

As far as individual characteristics towards study commitment go, the descriptive analysis delivers a diverse picture. The study time of students is typically low and studying a lot is rather an exception. Most students have passed their previous classes. As far as party behaviour goes, one can see that going out and drinking alcohol is fairly common.

What sticks out is the picture for absences. While most of the students are never skipping class, there are extreme outliers with up to 93 missed classes.

With most students in the middle and low extreme values, the first period grade seems almost normally distributed. The extremes get stronger as the course goes on however. In the second period grade, the portion of students with a zero grade get larger. This proportion becomes a lot larger in the final grade. It is worth to note that, while the same picture can be seen in the Portuguese data, the proportion of zero grades is not as large.

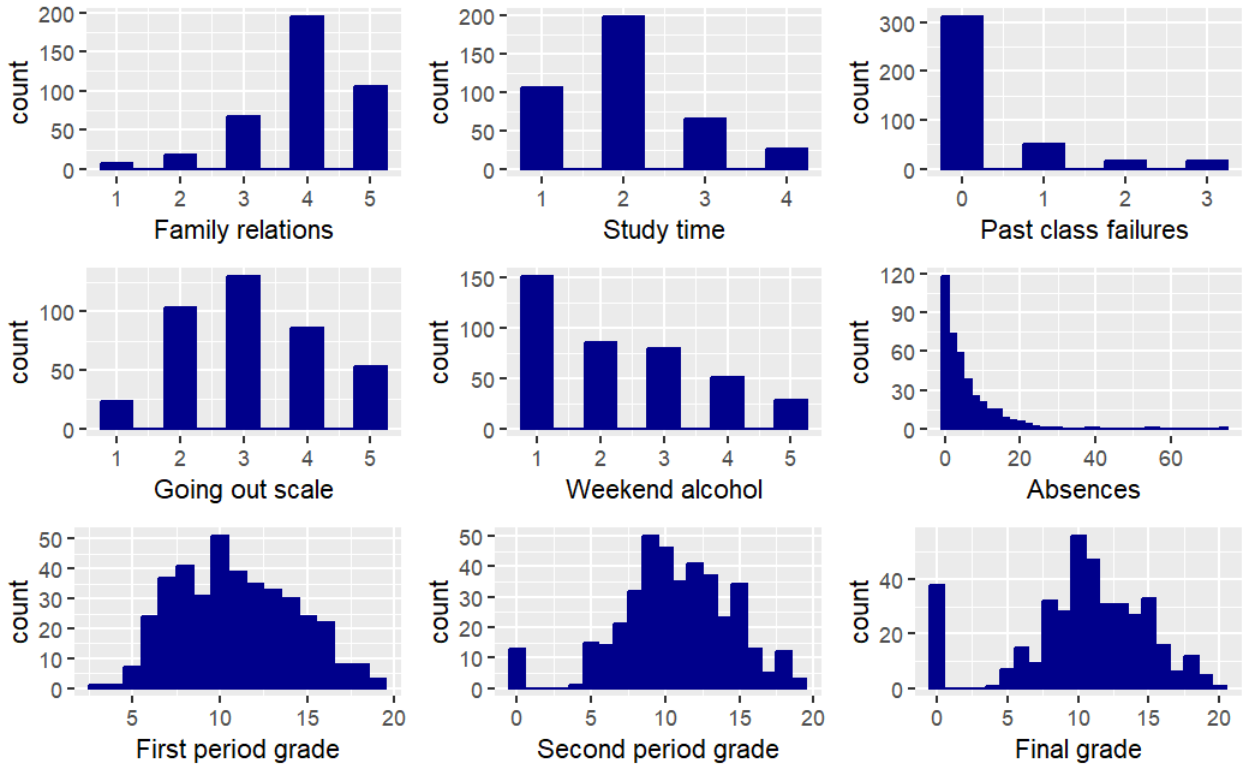

Figure 1: Histograms for the most important variables in Mathematics

3.2 Grade Differences

Analysing the relations of the grades throughout one school year yields various interesting insights. The first as well as the second period grade are the strongest factors in explaining the final grade. The regression analysis later will show the importance of past performance for future test results.

From looking at the two plots below one can see that there are different groups of students. In particular, there are two insights to draw from this: First, one can tell groups of students apart depending on when they start to fail the course with a zero grade. The graph on the left shows that there are students who start out with low grades (5-10) in the first period, but then manage to end up with a zero for the final grade. (Note that all the points on the bottom of the graph are zeros for the final grade. We

added a bit of noise to the plot to avoid overlapping points.) We suppose that this group of students is different from other students as there are others who start with the same initial grade but manage to pass in the end (there are also students who still fail, but at least with a positive grade). The same dynamic can be found in the plot on the right. There is a group of students that have a zero in the second period and stay there until the final. There are, however, also students who have a positive grade in the second period but still end up with a zero for the final grade. To start with a positive grade but end up with a zero, is a phenomenon that we cannot exactly explain. Plotting the final grade against the absences shows that the people with a zero final grade actually have zero absences. To further explore potential heterogeneities across different groups of students we perform a regression analysis on the sub-samples of students who fail and who pass. The summary of the results of this can be found in the Appendix.

Second, the two graphs reveal interesting insights about student improvement. The 45° line indicates the point where students stay exactly on the same grade. The two fitted regression lines (using a non-linear method) almost lie perfectly on this line, meaning that on average students stay on their initial level. Apart from that, the students in the left graph who lie above the black line improve compared to the first period. The ones below get worse. To study the potential heterogeneous dynamics behind the divergent learning outcomes of students we perform another sub-sample regression analysis later. Again, the results of this are described in the Appendix.

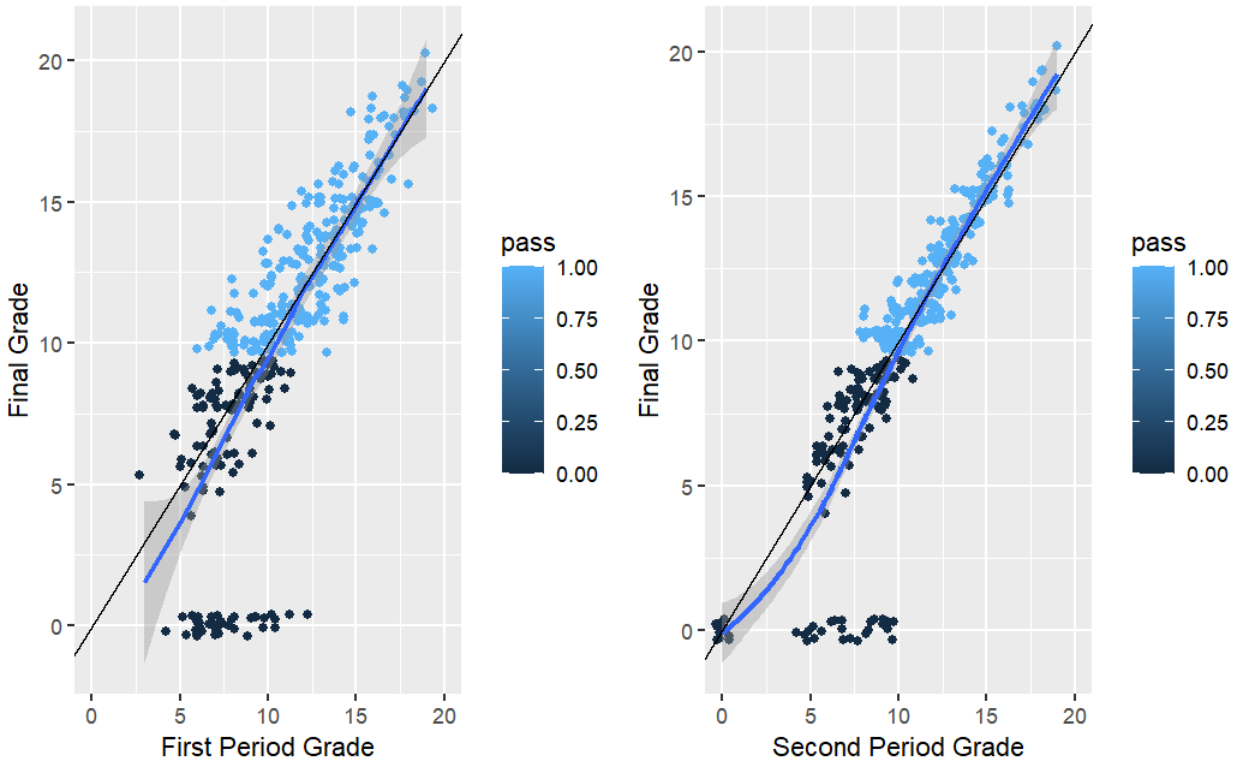

Figure 2: Final Grade plotted versus Second and First Period Grade

4 Methods and Results

Our goal is to create valuable insights on the factors that explain student performance. Knowing what improves or worsens students' learning outcomes is of high importance for schools as well as policy makers. To come to trustworthy conclusions, we first perform various checks to make sure we have the correct model to fit the data in the most adequate way. Furthermore, we account for non-linear effects as well as various heterogeneities.

4.1 Linear Model vs. Poisson

Student performance data as measured on the given 0-20 scale with discrete values only is a special case of a dependent variable. In order to find the model that explains the data in the most adequate way, we considered different methods.

The final grades feature a significant proportion of zeros and are observed in a narrow range. That could be a case of count data in which a poisson model might be fitting. This model is checked on its own as well as compared to the baseline linear regression model.

To get a first picture of the significance of covariates we run full models. Using the Bayesian Information Criterion we obtain model specifications that deliver the highest log-likelihood considering the number of parameters used (bestBIC).

In the poisson model we obtain a severe case of overdisperion. To account for that, we run a quasi-poisson specification. We check this model by computing bootstrapped 95% confidence intervals and compare it to the confidence intervals of the quasipoisson model. It turns out that the confidence intervals are far different from each other. The permutation test we perform for the poisson model delivers also far different coefficients in sign and significance from the poisson model.

The linear model, in comparison, performs very well. The errors seem relatively well-behaved and the bootstrapped 95% confidence intervals as well as the coefficients from the permutation tests are all very similar to our bestBIC-specification.

This said specification can be seen in our table 1, presenting the coefficients for the estimation of math final grades in column 1 and Portuguese final grades in column 2. The linear model for Mathematics features age, family relations, absences as well as the grades in previous periods as the most important factors. Interestingly, the Portuguese specification only features the previous grades, G1 and G2, and reaches even a slightly higher R^2. This exemplifies the high importance of previous performance in explaining future performance in tests. While the outcomes for age and family relations are more intuitive to understand, a positive effect of the number absences on the final grade should be questioned. As it could be a non-linearity in absences biasing our estimation, we run a general additive model to account for that. The estimate for absences stayed positive and the other coefficients did not significantly change. We cannot state with certainty what might be behind this unintuitive result, however, it might be a certain case of selection.

People who know that they will be able to pass may select to skip class more often.

Table 1: Linear Regression Models Comparison

	Dependent variable:	
	G3	
	Mathematics	Portugese
	(1)	(2)
age	-0.202^{***}	
	(0.077)	
famrel	0.357^{***}	
	(0.106)	
absences	0.044^{***}	
	(0.012)	
G1	0.158^{***}	0.149^{***}
	(0.055)	(0.036)
G2	0.978^{***}	0.897^{***}
	(0.049)	(0.034)
Constant	-0.078	-0.171
	(1.376)	(0.215)
Observations	395	649
R^2	0.834	0.848
Adjusted R^2	0.831	0.847
Residual Std. Error	1.881 (df = 389)	1.262 (df = 646)
F Statistic	389.754^{***} (df = 5; 389)	$1{,}798.671^{***}$ (df = 2; 646)
Note:		*p<0.1; **p<0.05; ***p<0.01

4.2 Binary Discrete Choice Model

To assess what factors are important for students to pass their class we run a binomial regression on the binary outcome pass or fail. For Math, we find that the most important factors are the father's education, family relation , going out, weekend alcohol consumption, and G2, according to bestBIC (Table 2). Surprisingly, weekend alcohol consumption has a positive effect, which might be due to self-selection rather than causal. With Portuguese, the BIC suggests that the final grade is mostly determined by the previous grades. For both subjects, the binomial models based on BIC explain around 84%

of the variation.

Table 2: Binomial Regression Models Comparison

	Dependent variable:	
	pass	
	Mathematics	Portugese
	(1)	(2)
Fedu	-0.567^{**}	
	(0.241)	
famrel	0.974^{***}	
	(0.327)	
goout	-0.632^{***}	
	(0.239)	
Walc	0.637^{***}	
	(0.202)	
G1		0.586^{***}
		(0.135)
G2	2.464^{***}	1.525^{***}
	(0.354)	(0.218)
Constant	-24.750^{***}	-18.451^{***}
	(3.881)	(2.141)
Observations	395	649
Log Likelihood	-61.551	-94.394
Akaike Inf. Crit.	135.102	194.789
Note:	*p<0.1; **p<0.05; ***p<0.01	

4.3 Prediction

To predict grades we choose to use the binary outcome pass/fail as it is highly important in practice to be able to support students, especially those who are expected to fail. First, we predict a student's probability to fail with our binomial bestBIC-specification. To address the overfitting issue, we split the data first into a training and test sample and further use cross-validation. For the training/test split, we find that an 80/20 split produces the most accurate predictions. For the Math class, we misclassified 16.5% of the data

in the test sample using our binomial prediction model. For the Portuguese class, that figure is 9.2%. The lower rate is partly due to combining covariates from both bestBIC models, so that we include previous grades, Father's education, family relation, going out and weekend alcohol consumption as predictors. Combining covariates only improved accuracy for Portuguese but not for Math. Also, the Portuguese data has almost double the observations, which allows to fit the model better to the training data.

To be able to make better predictions, we compare the binomial model to a random forest model, which performs best in this setting, according to Cortez and Silva. Random forest is an ensemble learning model that combines multiple decision trees to improve predictive performance and reduce overfitting. Decision trees are a type of model that recursively split the data based on the most informative features (like previous test scores) to create a tree-like structure that predicts the target variable (in our case passing/failing) based on a set of conditions. Each tree is trained on a random subset of the data and features, and the final prediction is based on the majority vote of the individual tree predictions. Unlike Cortez and Silva, we used the randomForest R package to implement the random forest model and then to find the optimal number of features that each decision tree considers. We found that four features per tree produce the highest accuracy. Using the random forest model enabled us to lower the share of misclassifications considerably from 16.5% to 12.7% for Math and from 9.2% to 6.9% for Portuguese.

	Binomial	Random Forest
Share of Misclassifications	0.165	0.127
Root Mean Squared Error	1.086	1.27

Figure 3: Mathematics: Binomial vs. Random Forest Prediction

	Binomial	Random Forest
Share of Misclassifications	0.092	0.069
Root Mean Squared Error	0.995	1.416

Figure 4: Portuguese: Binomial vs. Random Forest Prediction

The root mean squared error (RMSE) on the other hand is higher for the random forest model which implies lower accuracy. However, the percent of misclassification is arguably a more important measure in the case of a binary outcome, since it is not important how much the predicted probability differs from 0 or 1 but only if it is above or below 0.5. [1]

[1]Cross Validation could not be compared since the cross-validation random forest function does not produce a delta to compare to.

5 Discussion

In our analysis we study what factors contribute to student performances in the case of Mathematics and Portuguese classes in secondary education in Portugal. We find that previous test scores, absences and family relations are the most important factors across explanatory models. We believe that our models could be improved with additional variables that measure teacher quality and student intelligence or ability as these play a vital role in performance. Further, we predict whether students are likely to pass or fail based on these factors and achieve a predictive accuracy of 87% for Mathematics and 93% for Portuguese.

Such a prediction model could be used to make early interventions with students that are expected to fail. A further application could be a tool that lets students input a few key factors about themselves which are then used to give the teacher a passing prediction for that student. However, it is important to note that the amount of support from the family as well as extra support by the school or through paid courses did not seem to contribute to the student performance. Further research could analyze why these support schemes do not seem to affect performance and how they could be improved.

6 Appendix

The following figure shows the full set of variables and their respective scale.

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
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).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Figure 5: Full list of variables, from Costa and Silva (2008)

6.1 Heterogeneity in Passing and Failing

We, furthermore, try to understand if, compared to our main binary regression on passing or failing, there are different factors driving students to pass and fail respectively. We account for the potential heterogeneity across the groups by running full linear regression (with all covariates and the final grade as the regressand) on the sub-samples. What we get is that the coefficients for the two sub-samples in the Mathematics as well as the Portuguese dataset are mostly in line (in terms of which ones appear as significant) with the coefficients of the binary regression above. There are certain differences between the groups in certain variables becoming significant like health or type of guardian. That

probably reflects more sample peculiarities than actual heterogeneities.

6.2 Heterogeneity in Improvement

We suspect a second source of heterogeneity in the effects for students that improve their grades from the first to the final grade versus those that do not improve. For the group of improvers (i.e. the points that lie above the 45° line in figure 2) the age, failures, daytime alcohol consumptions, and absences determine their math grade difference. For non-improvers, the school, education of both parents, failures, romantic relationship and absences determine the difference. With Portuguese, only age has a significant effect for improvers, while for non-improvers only absences affect the grade difference significantly. Interestingly, the amount of support from the family as well as extra support by the school or through paid courses did not seem to contribute to the grade difference.

References

Cortez and Silva (2008). *Using data mining to predict secondary school student performance: Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, EUROSIS.*