THE UNIVERSITY OF CHICAGO

Policy Recommendation for Improving Secondary School Student Performance

Hyemin Jung, Jayoung Kang, Jeong Lim Kim

BUS 41201 - Big Data
Spring 2020

We pledge our honor that we have not violated the Honor Code during our assignment.

1. Executive Summary

In this project we analyzed the data of student academic performance in secondary education of two Portuguese schools in order to garner insight into what factors affect student performance. Based on our analysis outcomes we attempt to develop targeted policy recommendations based on the nature of the variables affecting student performance. Our original hypothesis was based on the assumption that the factors that would have the strongest effect on grades can be grouped by either family income related factors or by school related factors which will ultimately provide implications on where the primary policy focus should be for budget and resource allocation.

Our key questions for the analysis are the following:

1. What variables affect student performance, and can we predict student performance based on the variables?
2. Is there a causal relationship between demographic or school features and student academic performance?
3. Are there any underlying latent factors that allow us to group variables that will indicate the underlying factor that policy should address?

The key findings for each question are the following:

1. Not surprisingly, student performances are affected most by the number of past failures for the classes. Then, aspiration for higher education was important for predicting good school performance.
2. There is strong evidence for causal inference in that willingness to pursue higher education causes the outcome for pass/fail.
3. We were also able to identify a meaningful latent variable that grouped parents' level of education, which could be indicative of family income level.

Our policy recommendation therefore is focused on utilizing scholarship opportunities to encourage students to pursue higher education.

2. Data Introduction & Manipulation

The data for the analysis was collected from both student academic reports and questionnaires. The datasets were originally divided into two parts based on the performance of mathematics and Portuguese. There were students who either took just one of the two subjects or both. Our analysis focused on those who took both in order to analyze whether the influencing factors for student performance varied by the type of subject, especially on whether it differed by STEM or non-STEM subject. There are three grades for each subject representing the grades for each academic year. Based on our initial analysis of the grades, we could see strong correlation between the three grades and therefore took an average of the three scores to analyze as the outcome variable.

The 30 responses from the survey we used for analysis can be grouped as the following:

	Variable Name	Description Name	Data Type
1	**math**	Student's average Math grade	Numeric: from 0 to 20
2	**port**	Student's average Portuguese grade	Numeric: from 0 to 20
3	**sex**	Student's sex	Binary: 0 = female 1 = male
4	**age**	Student's age	Numeric: from 15 to 22
5	**Medu**	Mother's education	Numeric: 0 = none, 1 = primary education (4th grade), 2 = 5th to 9th grade, 3 = secondary education, 4 = higher education
6	**Fedu**	Father's education	Numeric: 0 = none, 1 = primary education (4th grade), 2 = 5th to 9th grade, 3 = secondary education, 4 = higher education
7	**traveltime**	Home to school commute time	Numeric: 1 = less than 15 min. 2 = 2 to 5 hours

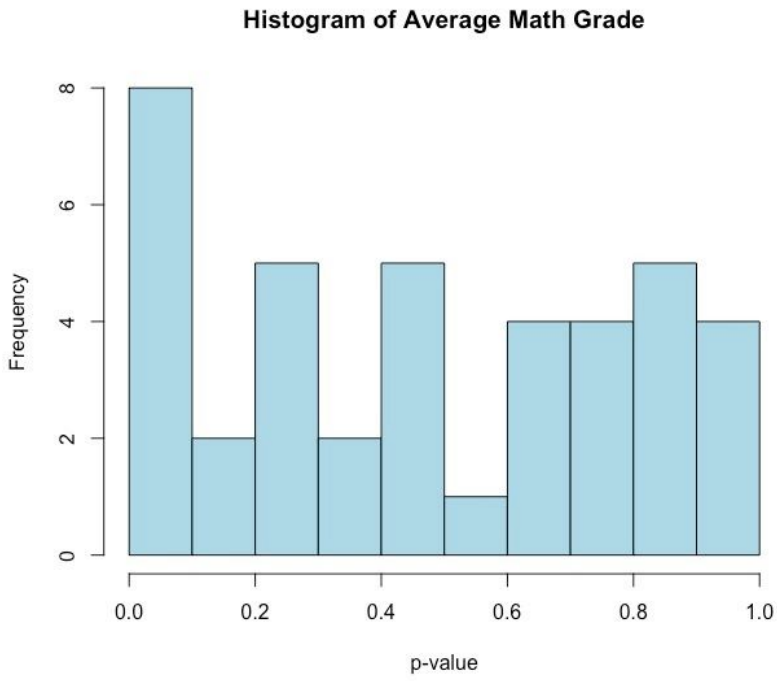
			3 = 5 to 10 hours 4 = greater than 10 hours
8	**studytime**	Weekly study time	Numeric: 1 = less than 2 hours 2 = 2 to 5 hours 3 = 5 to 10 hours 4 = greater than 10 hours
9	**famrel**	Quality of family relationships	Numeric: from 1 = very bad to 5 = excellent
10	**freetime**	Free time after school	Numeric: from 1 = very low to 5 = very high
11	**goout**	Going out with friends	Numeric: from 1 = very low to 5 = very high
12	**Dalc**	Workday alcohol consumption	Numeric: from 1 = very low to 5 = very high
13	**Walc**	Weekend alcohol consumption	Numeric: from 1 = very low to 5 = very high
14	**health**	Current health state	Numeric: from 1 = very bad to 5 = very good
15	**failures.x**	Number of past Math class failures	Numeric: n if 0<=n<3, else 3
16	**failures.y**	Number of past Portuguese class failures	Numeric: n if 0<=n<3, else 3
17	**absences.x**	Number of Math class absences	Numeric: from 0 to 93
18	**absences.y**	Number of Portuguese class failures	Numeric: from 0 to 93
19	**school**	Student's school	Binary: 0 = *Mousinho da Silveira* 1 = *Gabriel Pereira*
20	**address**	Student's home address	Binary: 0 = urban 1 = rural
21	**famsize**	Family size	Binary: 0 = less or equal to 3 1 = greater than 3
22	**Pstatus**	Parent's cohabitation status	Binary: 0 = living together 1 = apart
23	**Mjob**	Mother's job	Nominal: 'teacher' = teacher 'health' = health-related

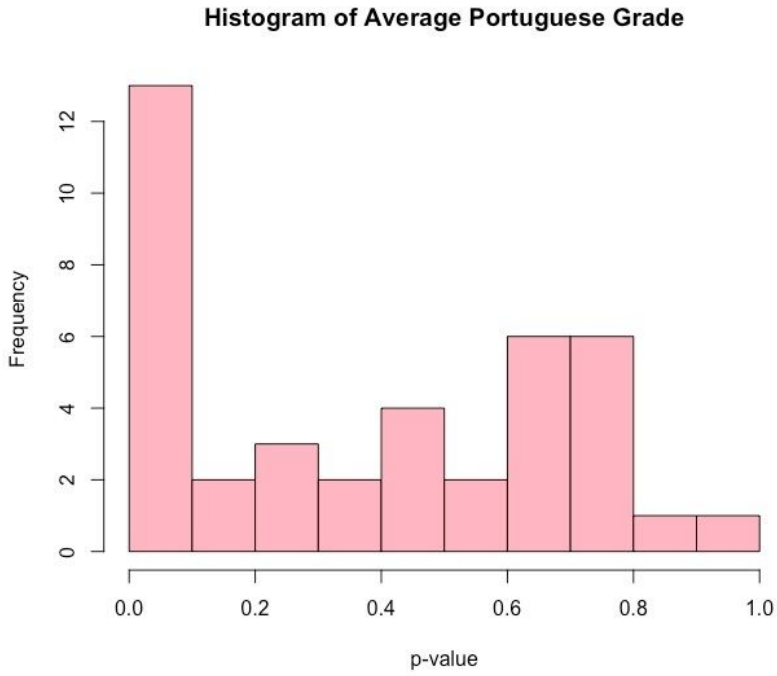
			‘services = civil services ‘other’ = at home or other
24	**Fjob**	Father’s job	Nominal: ‘teacher’ = teacher ‘health’ = health-related ‘services = civil services ‘other’ = at home or other
25	**guardian**	Student’s guardian	Nominal: ‘father’ = father ‘mother’ = mother ‘other’ = other
26	**activities**	Extra-curricular activities	Binary: 0 = yes 1 = no
27	**nursery**	Attended nursery school	Binary: 0 = yes 1 = no
28	**higher**	Wants to take higher education	Binary: 0 = yes 1 = no
29	**internet**	Internet access at home	Binary: 0 = yes 1 = no
30	**romantic**	With a romantic relationship	Binary: 0 = yes 1 = no
31	**paid.x**	Extra paid classes for Math	Binary: 0 = yes 1 = no
32	**paid.y**	Extra paid classes for Portuguese	Binary: 0 = yes 1 = no

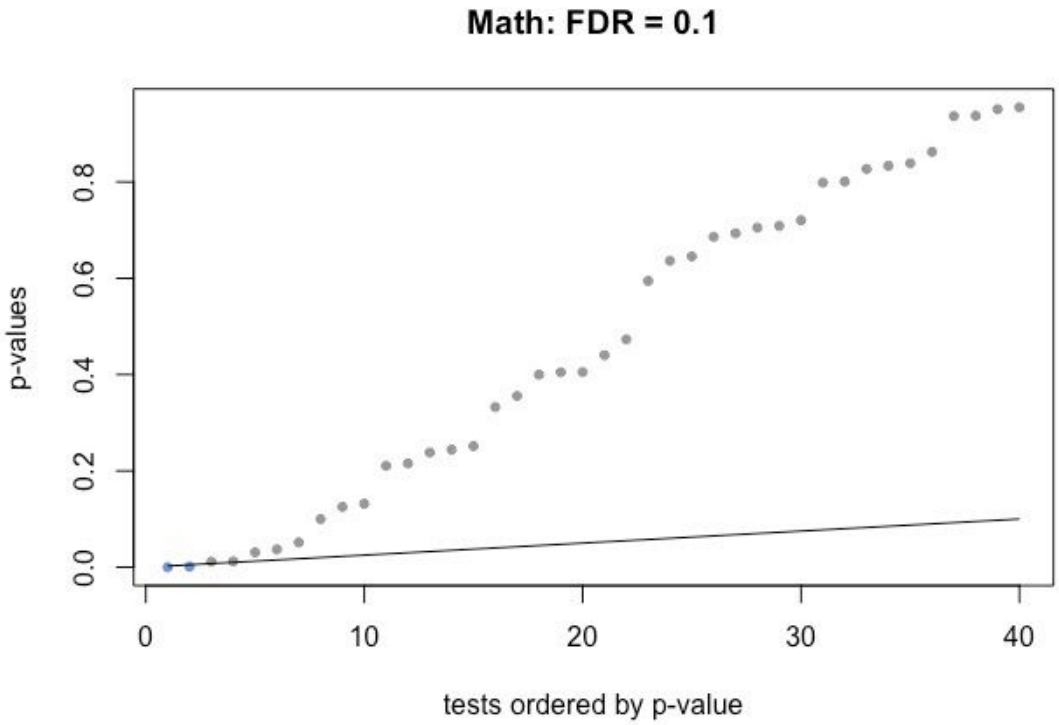
Nominal data such as parent jobs, and reason for school attendance were converted into numerical factor variables in order to conduct statistical analyses such as FDR and LASSO.

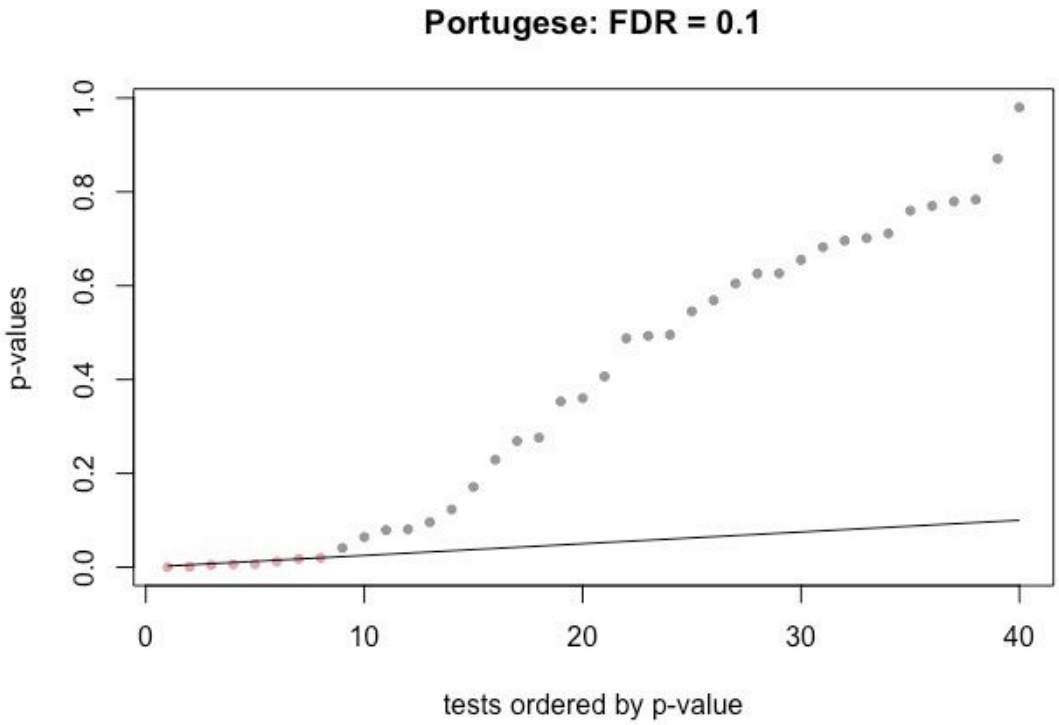
3. Identifying Significant Variables for Predicting Student Performance

FDR Analysis






Math: FDR = 0.1
p-values
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
tests ordered by p-value


Portugese: FDR = 0.1
p-values
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
tests ordered by p-value

A spike near zero indicates hopeful signals. When the p-value distribution deviates from a uniform distribution, the histogram demonstrates the spike. Compared to the average math grade, the histogram of average Portuguese grade demonstrates clearer peak around zero.

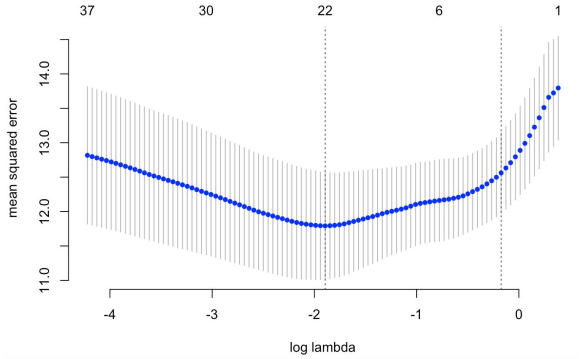
To reduce the risk of false discovery, we ran a 10% FDR cut and found the following:

- Average math grade has **two** variables significant enough to fall below the threshold of **0.0009858427**:
    - the number of past failure math classes (failures.x)
    - the receipt of extra educational support (schoolsup_yes)
- Average portugese grade has **eight** variables significant enough to fall below the threshold of **0.01979525**:
    - weekly study time (studytime)
    - current health status (health)
    - the number of past failure Portuguese class (failures.y)
    - school Mousiho de Silveira (school_MS)
    - male student (sex_M)
    - urban address (address_U)
    - the receipt of extra educational support (schoolsup_yes)
    - desire for higher education (higher_yes).

The FDR analysis is only valid when p-values are independent. In our multiregression, multicollinearity may cause inaccurate interpretation. Thus, we need to further examine through LASSO techniques.

LASSO Analysis

We ran LASSO regression of **average math grade** on 42 potential inputs, to select the coefficient. This analysis found 3 coefficients to be significant under CV 1se selection. The number of past failure math classes(failures.x), the receipt of extra educational support (schoolsup_yes) were most significant coefficients excluding intercept, and the other significant coefficients are presented in the table below.



CV min deviance chooses 22 with log(lam) -2.13
CV 1se chooses 3 coefficient with log(lam) -0.453

	Variable	Estimate
1	**intercept**	**13.07**
2	**failures.x**	**-1.517**
3	**schoolsup_yes**	**-1.328**
4	**higher_yes**	**1.259**
5	**Mjob_health**	**0.983**
6	**Fjob_teacher**	**0.957**
7	**Mjob_services**	**0.848**
8	**sex_M**	**0.781**
9	**famsup_yes**	**-0.533**
10	**romantic_yes**	**-0.464**
11	**famsize_LE3**	**0.456**
12	**reason_reputation**	**0.312**

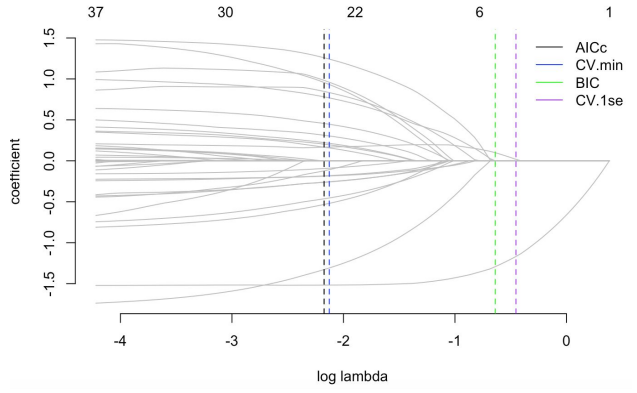
13	Fjob_other	-0.263
14	goout	-0.261
15	studytime	0.222
16	address_U	0.203
17	age	-0.187
18	traveltime	-0.184
19	Medu	0.167
20	internet_yes	0.167
21	guardian_other	-0.122
22	health	-0.094

Top 2 significant LASSO coefficient interpretation (excluding intercept)

Average math grades decrease by 1.517 with an additional number of failures in math class while keeping everything else fixed.

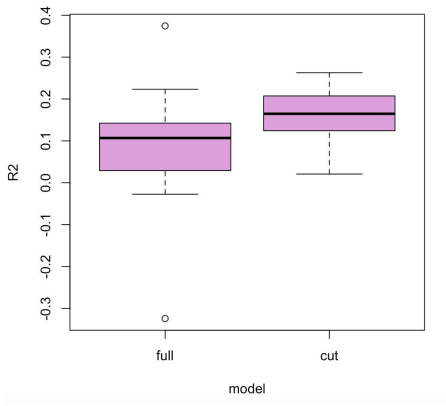
Average math grades are 1.328 lower for students who had extra math education from school than the students who didn't, while keeping everything else fixed.

Lambda and variable selection



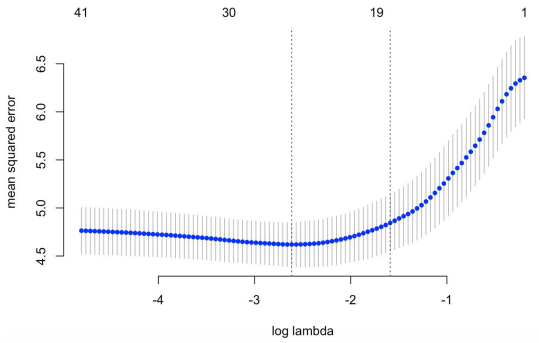
Number of selected variables follows: AICc > CV.min> BIC > CV.1se

Goodness of fit



Goodness of fit comparison follows:
In-sample R^2 (0.307) > Cut model mean for 10 folds OOS R^2 (0.1586) > full model R^2 (0.0837)

Then we ran LASSO regression of **average portugues grade** on 42 potential inputs, to select the coefficients(CV 1se). This analysis found 19 coefficients to be significant; desire for higher education (higher_yes), the receipt of extra educational support (schoolsup_yes), the number of past failure Portuguese class (failures.y), urban address (address_U), school Mousiho de Silveira (school_MS), father's job teacher(Fjob_teacher), extra paid classes for Portuguese(paid.y_yes), studytime(studytime), male(sex_M), workday alcohol consumption(Dalc), mother's job health related(Mjob_health), less than 3 family size(famsize_LE3), health extra-curricular activities(health, activities_yes), mother's job other(Mjob_other), mother's education level(Medu), frequency of going out with friends(goout), traveltime(traveltime), father's education level(Fedu), excluding intercept. Other significant coefficients are presented below.



CV min deviance chooses 27 with log(lam) -2.61
CV 1se chooses 19 coefficient with log(lam) -1.59

	Variable	Estimate
1	intercept	10.655
2	higher_yes	1.95
3	schoolsup_yes	-1.217
4	failures.y	-0.649
5	address_U	0.609
6	school_MS	-0.578
7	Fjob_teacher	0.427
8	paid.y_yes	-0.422
9	studytime	0.376
10	sex_M	-0.371
11	Dalc	-0.316
12	Mjob_health	0.232
13	famsize_LE3	0.18
14	health	-0.177
15	activities_yes	0.153
16	Mjob_other	-0.15
17	Medu	0.144
18	goout	-0.098
19	traveltime	-0.06
20	Fedu	0.047
21	absences.y	-0.026
22	age	0
23	famrel	0
24	freetime	0
25	Walc	0

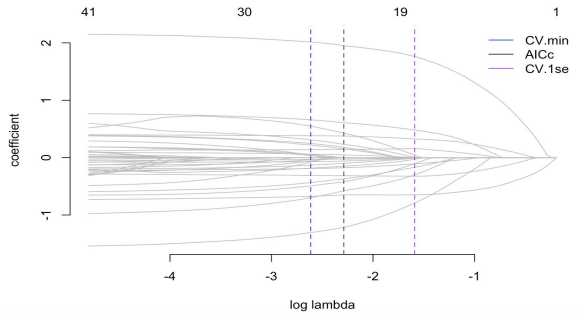
26	Pstatus_T	0
27	Mjob_services	0

Top 2 significant LASSO coefficient interpretation (excluding intercept)

Average portuguese grades are 1.94 higher for the students who wish to pursue higher education than the students who didn't,while keeping everything else fixed.

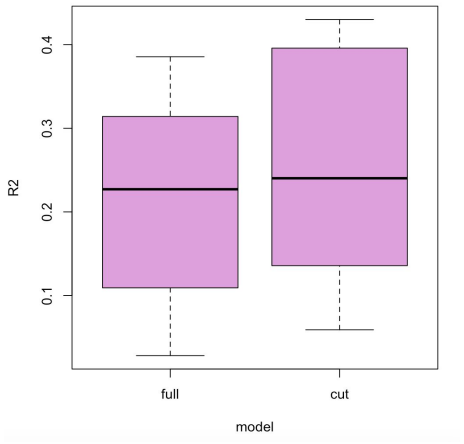
Average portuguese grades are 1.217 lower for students who had extra portuguese education from the school than the students who didn't, while keeping everything else fixed.

Lambda and variable selection



Number of selected variables follows: CV.min> AICc > CV.1se

Goodness of fit

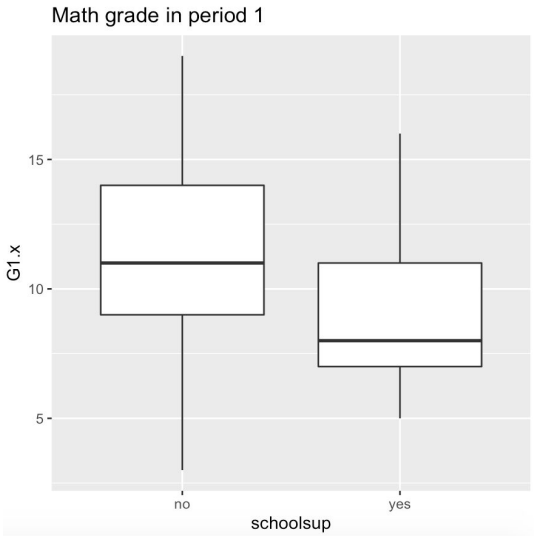


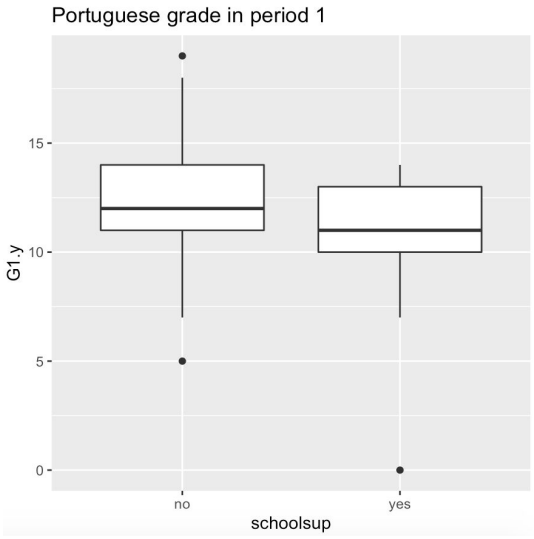
Goodness of fit comparison follows:
In-sample R^2 (0.359) > Cut model mean for 10 folds OOS R^2 (0.248) > full model R^2 (0.216)

Caveat of LASSO analysis result interpretation

Lasso analysis suggests that school support has high prediction power for both math and portuguese classes. However, school support has negative correlation with grades, which one would not normally expect. This suggests either school support is worsening students' performance or the selection bias into school support treatment.

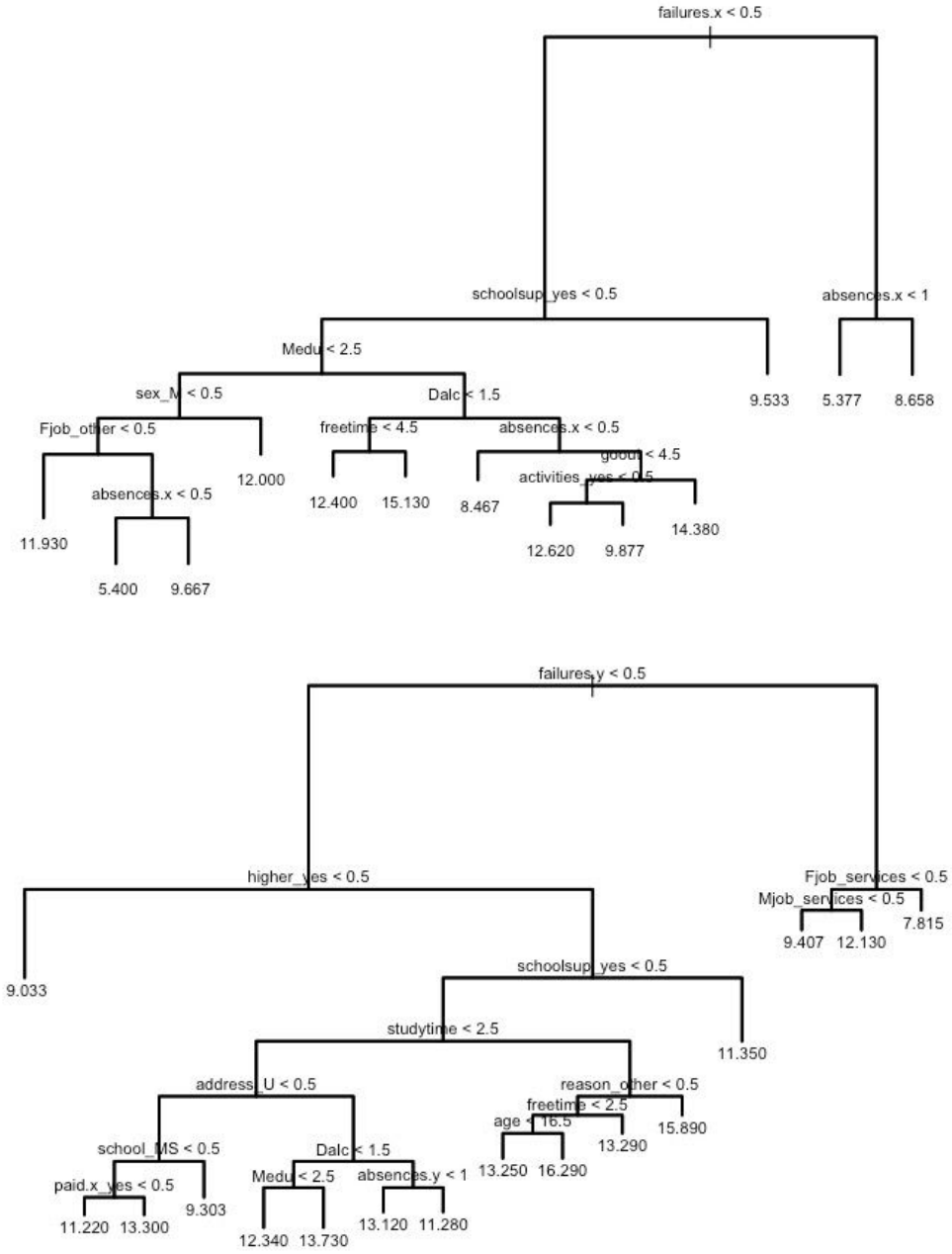
Comparing the first math and portuguese score by school support treated to untreated, we realized that the students who received school support from school had approximately 5 points lower in grade(math), and 2 points lower in grade(portuguese) to begin with. This has created downward bias for the school support effect.





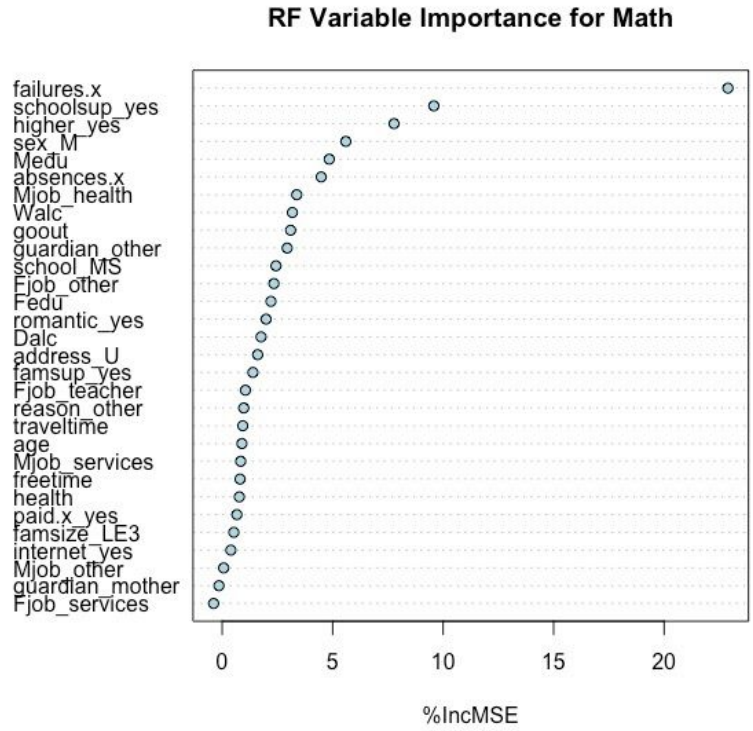
CART and Random Forest

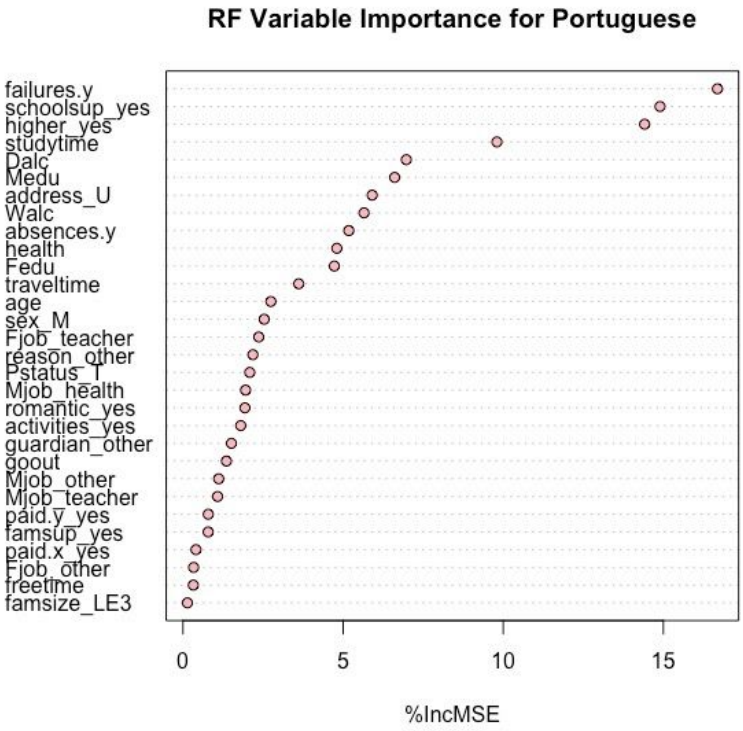
Furthermore, we used the CART algorithm and random forest technique to predict student performance. The tree for Math and Portuguese grade is shown below, respectively:



The CART model shows for both Math and Portuguese grades, the number of past failure classes on its respective subject places the highest importance. For both subjects, the model also shows

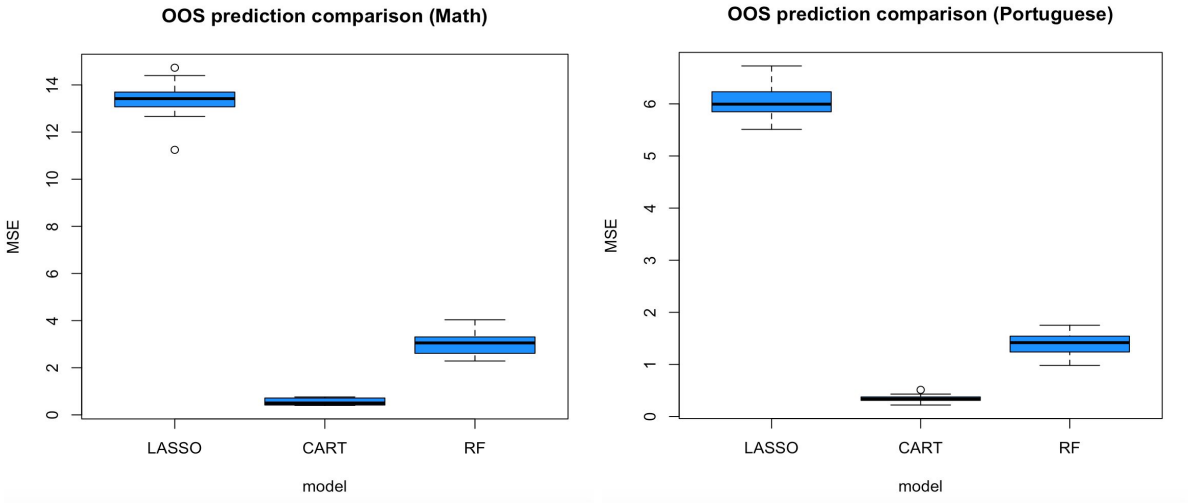
its importance on extra school support. Similar results show when we run random forest technique:





Although the order of importance is slightly different, the random forest model finds many of the factors significant as previous models we tested. Based on the random forest graphs above, both subjects emphasize the importance of the number of past subject failures, extra school support on the subject, and the desire for higher education.

OOS prediction over 10 random folds for the student performance regressions indicates that the trees outperformed LASSO and CART is better than RF.



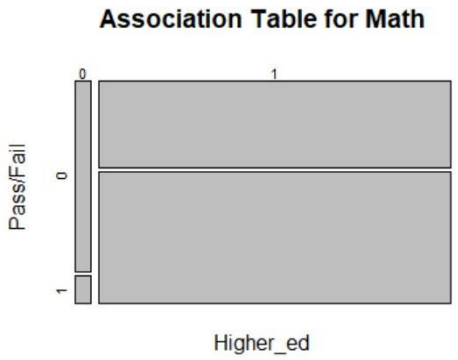
It is also worth noting that there are many variables significant in determining the average grade for Portuguese compared to the average grade for Mathematics in our FDR, LASSO, CART, and random forest analysis.

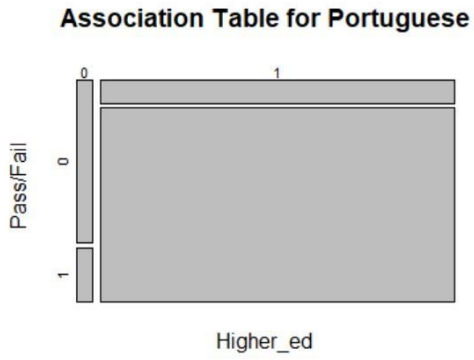
4. Causal Inference of Significant Variables

In order to identify variables that could provide insight into causal inference, we conducted the double LASSO and bootstrap analysis. Of the candidate variables our choice was to utilize the outcome from the previous analyses to determine the treatment variable. In investigation of the variables that were highly significant for both courses, the applicable variables are student willingness to pursue higher education and availability of school support. In the analysis of causal inference, our primary focus was to see if there is a causal relationship between the treatment variable and whether the student could on average receive a passing score for the course. Therefore, we converted the average grade into a pass/fail grade by assigning those with above 50% as pass and those below as fail.

Effect of 'Willingness to Pursue Higher Education'

The association between student performance in mathematics and willingness for higher education can be seen as below. Since most students are willing to pursue higher education there is a skewed proportion in the data distribution. However, based on the association graph it is possible to observe that there is a positive correlation between passing and willingness for higher education. A regression analysis of the two variables also supports this idea in that the interpretation of the regression reads that willing to pursue higher education raises the odds of passing the mathematics course by 10.6 times. A similar output can be seen for Portuguese as the regression indicates that the odds increases by 25 times.

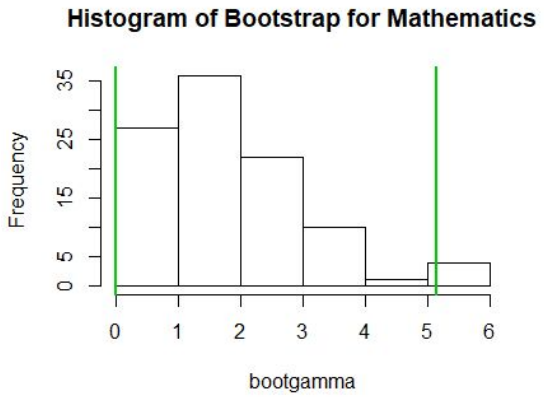


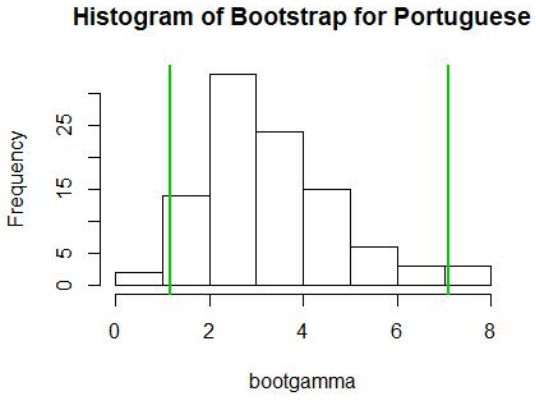


To determine whether this is causal, we need to isolate the 'pure' effect of willingness to pursue higher education that is not affected by other covariates. The regression of higher education willingness on all other variables for mathematics shows that approximately 22% of the variability is explained by the other variables and 78% is independent of the other variables. For Portuguese related variables, 81% is independent of other variables.

The double lasso coefficient of the treatment variable for mathematics is 1.130638, indicating the possibility of a signal that there is an effect of willingness to pursue higher education on student grades when keeping all other variables constant. The result differs from the naïve LASSO which is also indicative of a possible signal. For Portuguese, the double lasso coefficient is 1.893189, telling a similar story to that of mathematics.

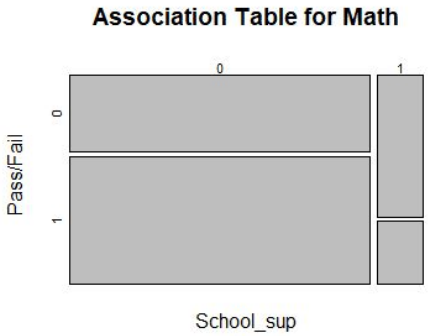
To investigate the how variable the treatment effect is, and to consider the validity of the outcome we conducted a bootstrap analysis to draw out the distribution of the treatment effect and created a confidence interval and identify the standard error. The 95% confidence interval of the distribution can be seen as below. Assuming that we have reasonably controlled for other variables, since for both mathematics and Portuguese the distribution starts above zero, we can infer that there is strong evidence for causal inference in that willingness to pursue higher education causes the outcome for pass/fail.

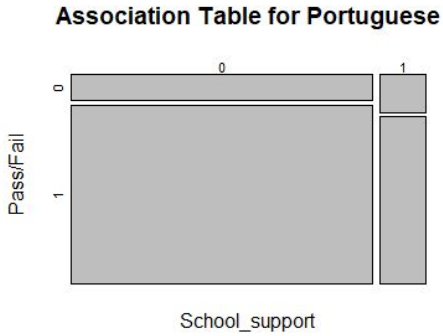




Effect of 'Availability of School Support'

In contrast to what we hypothesized, the availability of school support seems to have a negative correlation with student performance for both mathematics and Portuguese. The association between student performance in mathematics and availability of school support can be seen as below. A regression analysis of the two variables shows that students who had school support had the odds of passing the mathematics course become 0.27 times and for Portuguese the odds changes by 0.65 times.



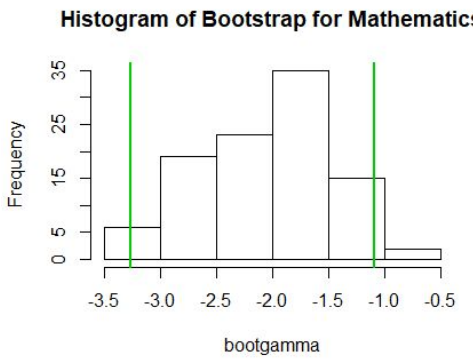


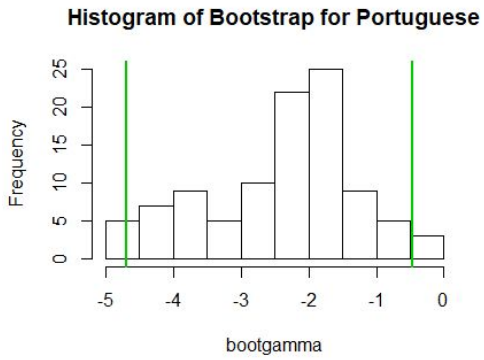
The regression of taking school support on all other variables for mathematics shows that approximately 11% of the variability is explained by the other variables and 89% is independent of the other variables. For Portuguese related variables, 91% is independent of other variables.

The double lasso coefficient of the treatment variable for mathematics is -1.546, indicating the possibility of a signal that there is a negative effect on student grades when keeping all other variables constant. The result differs from the naïve lasso which is also indicative of a possible signal. For Portuguese, the double lasso coefficient is -0.869, telling a similar story to that of mathematics.

To investigate the how variable the treatment effect is, and to consider the validity of the outcome we conducted a bootstrap analysis to draw out the distribution of the treatment effect and created a confidence interval and identify the standard error. The 95% confidence interval of the distribution can be seen as below. Although the analysis from the bootstrap and the double

lasso seem to yield a counterintuitive outcome for causal relationship, we can infer that this is a result from selection bias. Students that are constantly failing are likely to take extra school support. However, this could also be indicative of the fact that the school support is not effective enough to counteract the effect of the selection bias by sufficiently improving student scores.

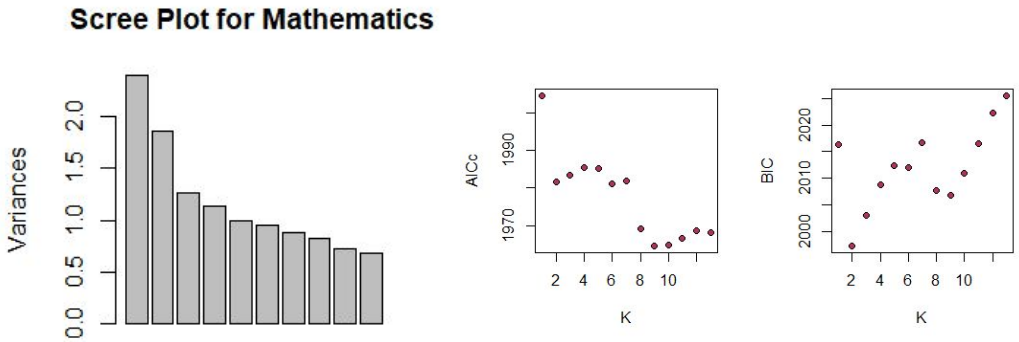




5. Grouping Variables through Factor Analysis

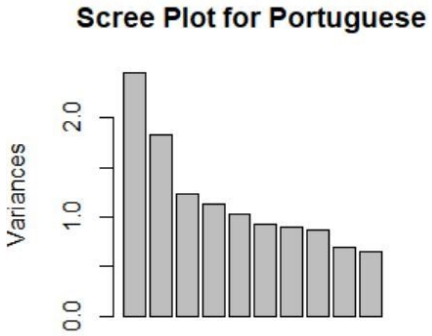
Given the dataset, our primary concern was that some key variables that could be grouping the given variables could be missing. Namely, common variables related to student performance such as family income and student innate ability. Our hypothesis is that these variables could appear as latent variables that appear through factor analysis.

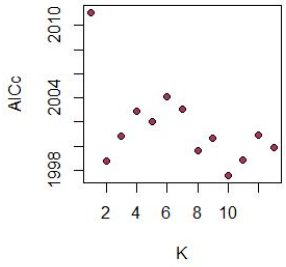
In analyzing the principal components for variables related to mathematics, there are a total of 13 principal components. The PC scree plot indicates that there is a large difference between the variance of the first two PCs and the rest of the PCs. An interpretation of the two PCs indicates that the first PC relates to student's alcohol consumption and time spent going out. The second PC groups parent's level of education. However, it is difficult to conclude that these variables are necessarily indicative of a more meaningful latent variable. The AICc and BIC select 9 and 2 Ks respectively, while lasso chooses 3. The first two PCs and the eighth PCs are selected under lasso.

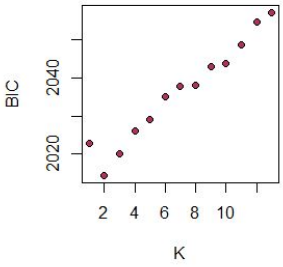


	PC1	PC2	PC8
age	-0.2	0.1	0.4
Medu	0.1	-0.6	0.2
Fedu	0.1	-0.6	0.3
traveltime	-0.2	0.2	0.1
studytime	0.3	0	0.1
famrel	0.1	0	-0.2
freetime	-0.2	-0.1	-0.2
goout	-0.4	-0.2	0.1
Dalc	-0.5	-0.2	-0.1
Walc	-0.5	-0.2	-0.1
health	-0.1	0	0.2
failures.y	-0.3	0.2	0.5
absences.y	-0.2	-0.2	-0.5

For Portuguese, the two PCs chosen under LASSO are the first two PCs which show similar groupings to that of mathematics. However, it is difficult to conclude that this PC is meaningful for analysis because daily and weekly alcohol consumption are directly related. The grouping for parent education can be indicative of a matching of skills in marriage and thus family income.







	PC1	PC2
age	-0.2	0.2
Medu	0.1	-0.6
Fedu	0.1	-0.6
traveltime	-0.2	0.2
studytime	0.3	0
famrel	0	0
freetime	-0.2	-0.2
goout	-0.3	-0.2
Dalc	-0.5	-0.2
Walc	-0.5	-0.2
health	-0.1	0
failures.y	-0.3	0.2
absences.y	-0.2	-0.1

6. Conclusion

From the three methods for attempting to analyze significant variables for the mathematics and Portuguese course, the outputs were relatively similar. Although we expected some differences between the FDR and lasso outputs, it is possible that they did not diverge significantly because the multicollinearity between variables that could have invalidated the FDR was not significant. Another finding was that the analyses yielded a different set of significant variables for the mathematics and Portuguese course, indicating that there are some inherent differences between the subjects. Based on the output from lasso, it seems that Portuguese is more affected by external factors such as study time and location of residence, while mathematics is not. This could be indicative of the fact that performance in non-STEM courses can be influenced by personal efforts and availability of resources, while for STEM courses inherent ability is more meaningful.

Based on the analysis of variables that were commonly significant for both mathematics and Portuguese, we were able to identify causal relationships for both cases on whether a student on average received a pass or fail score. We were also able to identify a meaningful latent variable that grouped parents' level of education. Although this provides some support to the hypothesis that family income could be a possible latent variable, more data that could be indicative of this should be grouped to support this.

These findings indicate some meaningful implications for policies that aim to improve student performance. To encourage students to pursue higher education attempts should be made to identify the reason behind those who do not want to enter higher education and develop policies accordingly. If the student faces financial hardships that deter them from pursuing higher education they will be underperforming and not reaching their full potential knowing that there is a limit to their education. Scholarship opportunities should be presented to such students. However, for students who are unwilling to pursue higher education because they wish to pursue other career interests, there is no need to alter their behavior through policy.

Also, It is likely that for non-STEM courses such as Portuguese is more affected by variables such as study time and location of residence because external resources and input into studying can change performance results while STEM courses are more affected by innate ability. In such cases it could be recommended that additional resources be provided for students living in rural settings to aid their coursework and to increase mandatory study sessions at school to improve academic outcome.