Student Performance

Group B: Akanksha Tandon, Ceren Ungan, Madeleine Beck, Andrew Yohanan

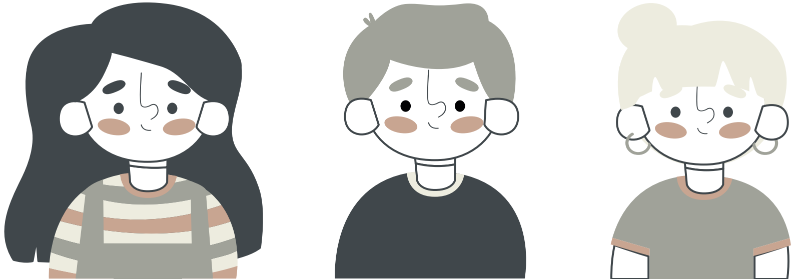
Agenda

Data Exploration

Modeling

Important Predictors

Overall Findings and Recommendations



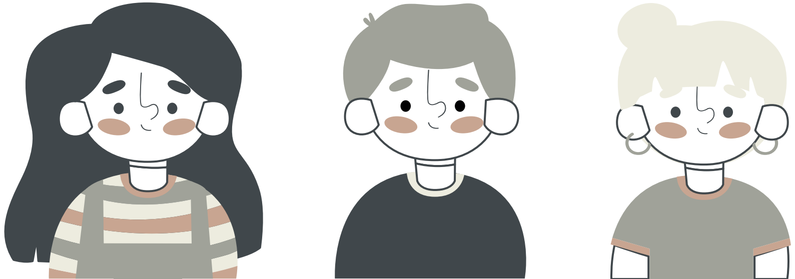
Agenda

Data Exploration

Modeling

Important Predictors

Overall Findings and Recommendations



Problem Statement

How is student achievement in Math affected by different life situations?

- What are the characteristics of a **stellar** student?
- What are the circumstances of an **underperforming** student?
- What are the most meaningful **areas of intervention** that could help students?

Analysis entails two Portuguese secondary schools with a total of 395 observations.

Variables

School

Sex

Age

Address

Family Size

Parent's cohabitation status

Mother's education

Father's education

Mother's job

Father's job

Reason to choose school

Guardian

Travel Time

Study Time

Failures

School Support

Family Educational Support

Paid

Activities

Nursery

Higher- wants to take higher education

Internet

Romantic Relationship

Family Relationship

Free Time

Go out

Dalc- workday alcohol consumption

Walc- Weekend alcohol consumption

Health

Absences

<u>Created Variables</u>

Grade: Composite for all three exams (G1+G2+G3)/3

Pass: If Grade => 13 Pass, else Fail

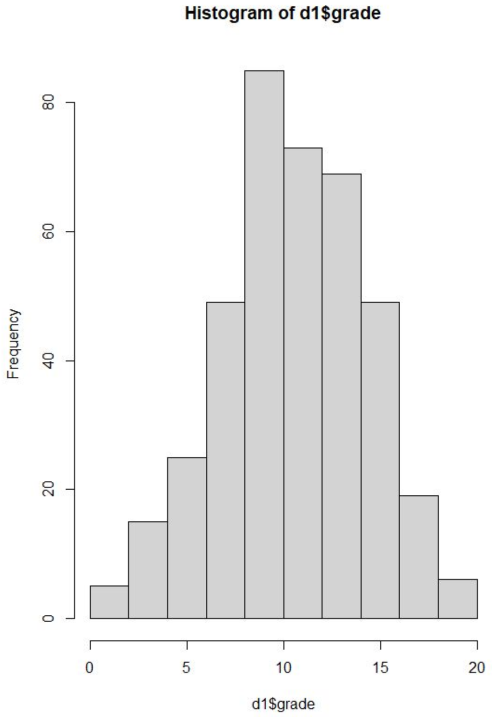
Drop: If G3 is 0, student dropped the class

Correlations between Variables

Upon compiling a correlation matrix of all numerical variables, the following are highly correlated:

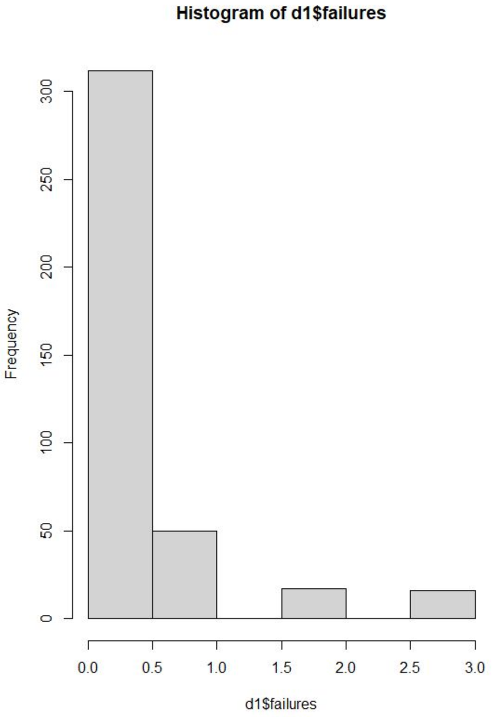
- First, Second, and Third Period Grade
- Workday and Weekend Alcohol Consumption
- Father's and Mother's Education

School Distribution

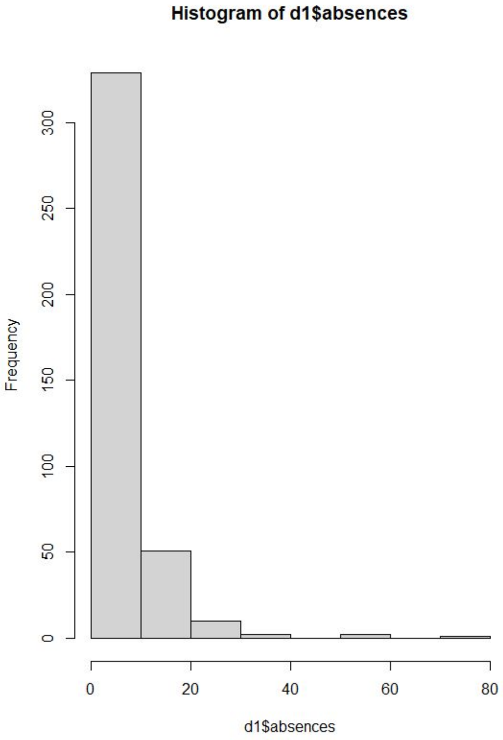


Histogram of d1$grade

Dependent Variable

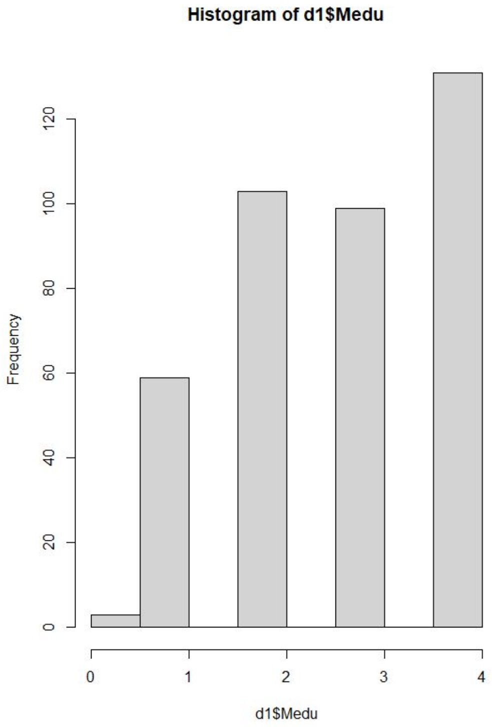


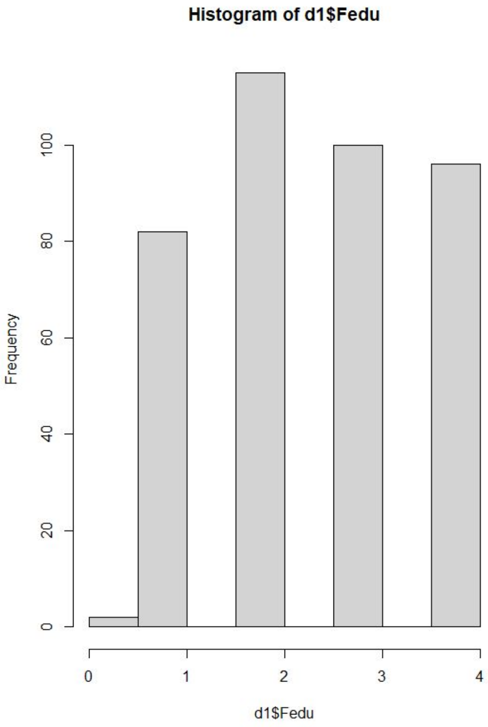
Histogram of d1$failures

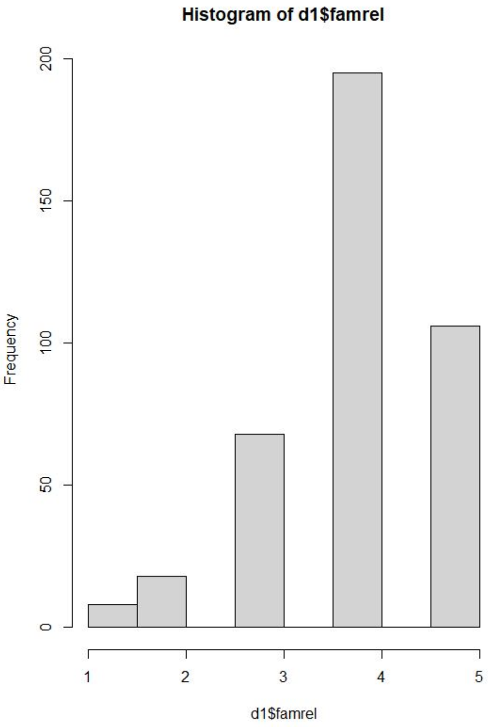


Histogram of d1$absences

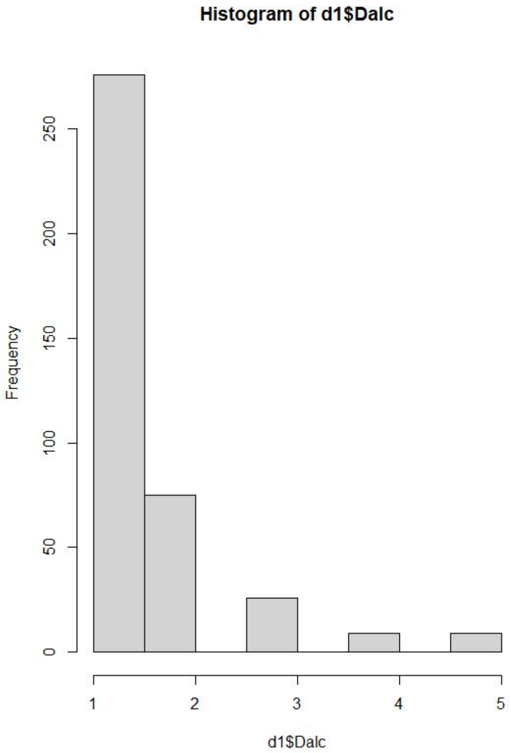
Family Distribution

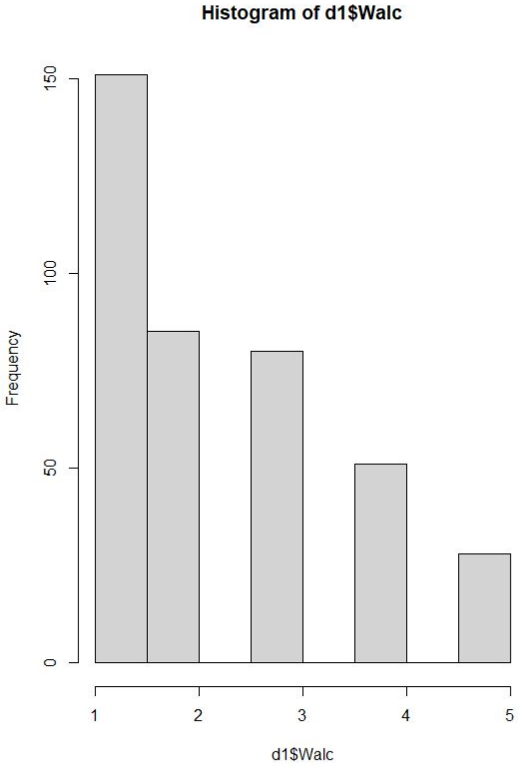


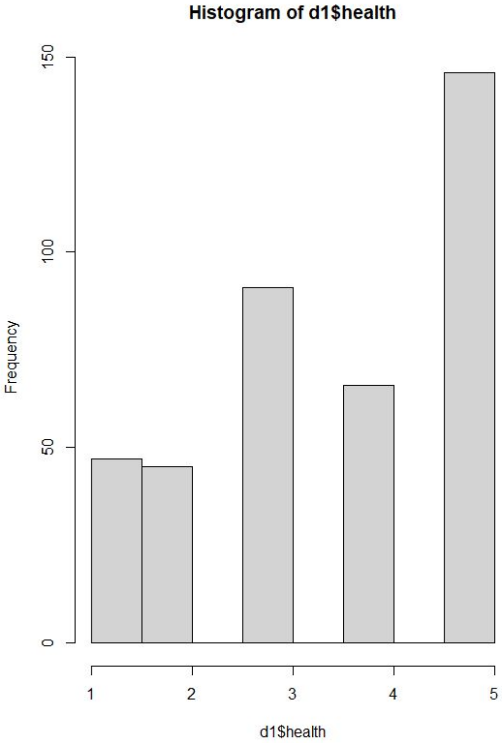




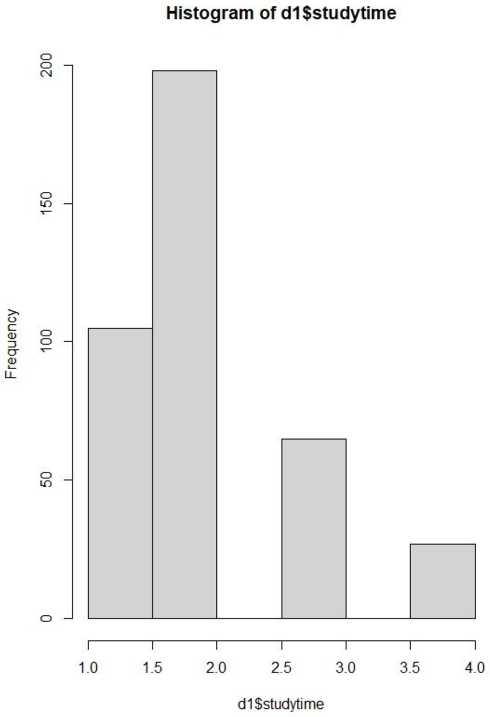
Wellness Distribution

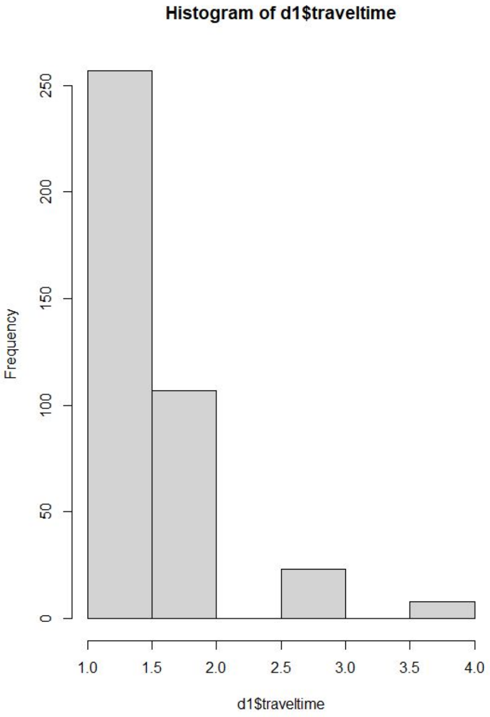


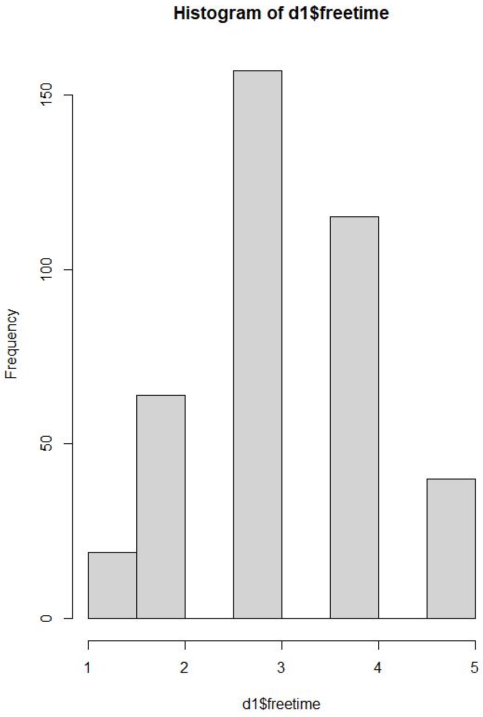




Time Distribution







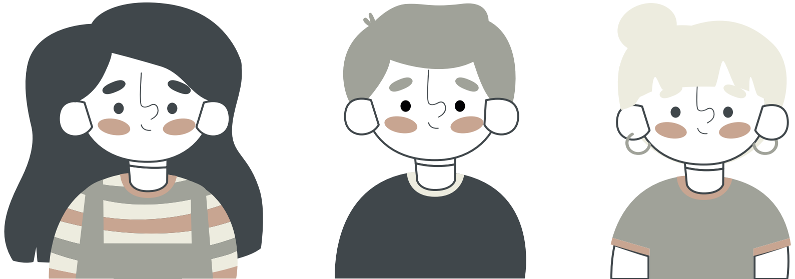
Agenda

Data Exploration

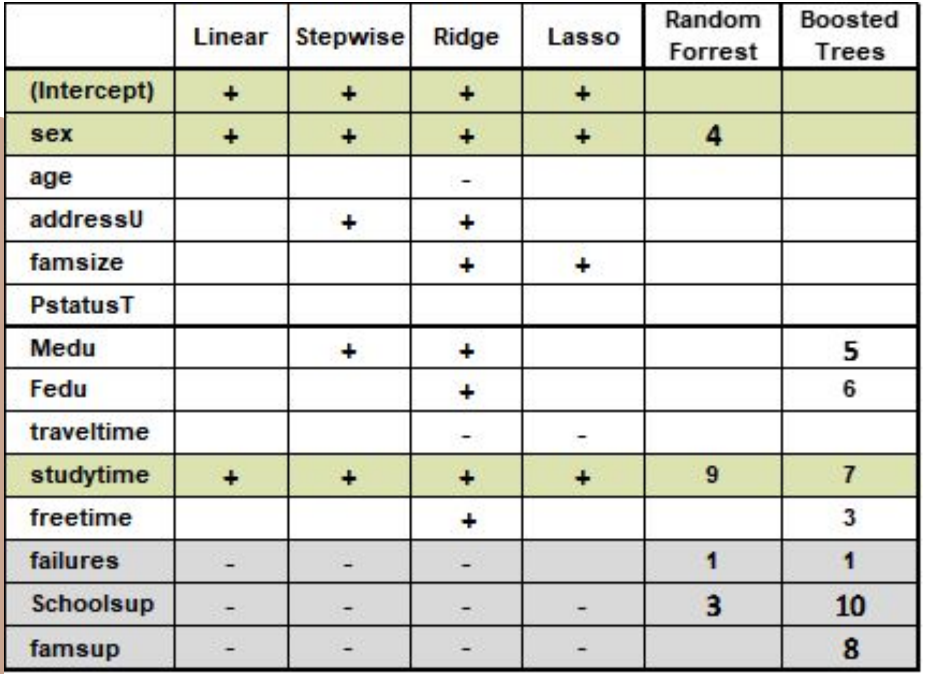
Modeling

Important Predictors

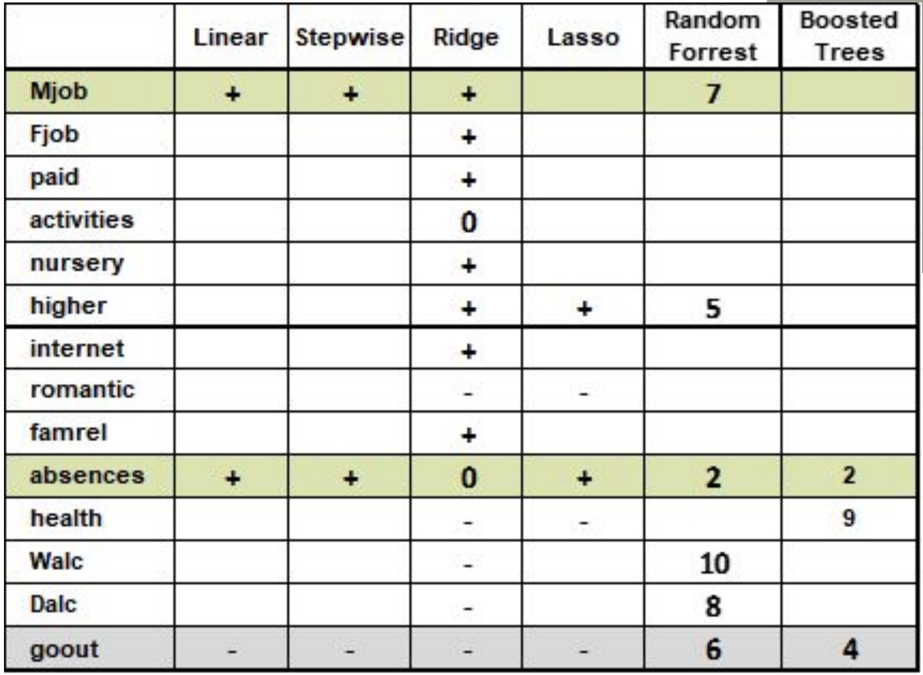
Overall Findings and Recommendations



Comparison of Results

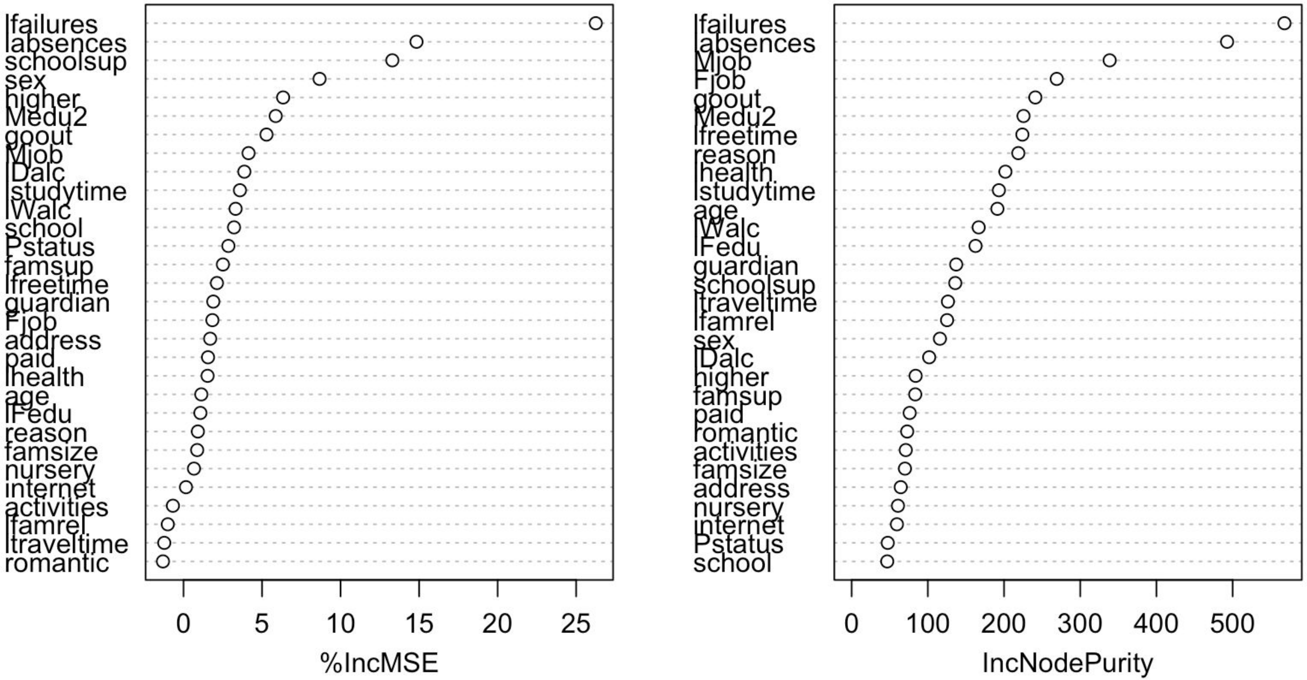


	Linear	Stepwise	Ridge	Lasso	Random Forrest	Boosted Trees
(Intercept)	+	+	+	+		
sex	+	+	+	+	4	
age			-			
addressU		+	+			
famsize			+	+		
PstatusT						
Medu		+	+			5
Fedu			+			6
traveltime			-	-		
studytime	+	+	+	+	9	7
freetime			+			3
failures	-	-	-		1	1
Schoolsup	-	-	-	-	3	10
famsup	-	-	-	-		8



	Linear	Stepwise	Ridge	Lasso	Random Forrest	Boosted Trees
Mjob	+	+	+		7	
Fjob			+			
paid			+			
activities			0			
nursery			+			
higher			+	+	5	
internet			+			
romantic			-	-		
famrel			+			
absences	+	+	0	+	2	2
health			-	-		9
Walc			-		10	
Dalc			-		8	
goout	-	-	-	-	6	4

Random Forests with Grade DV



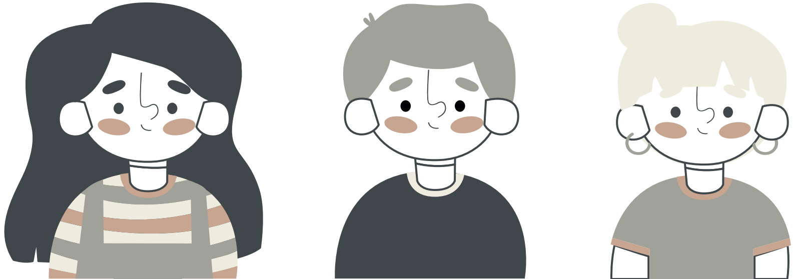
Agenda

Data Exploration

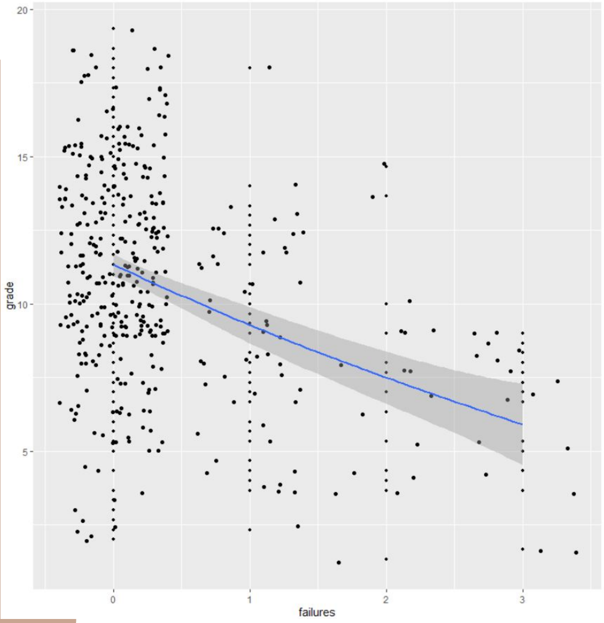
Modeling

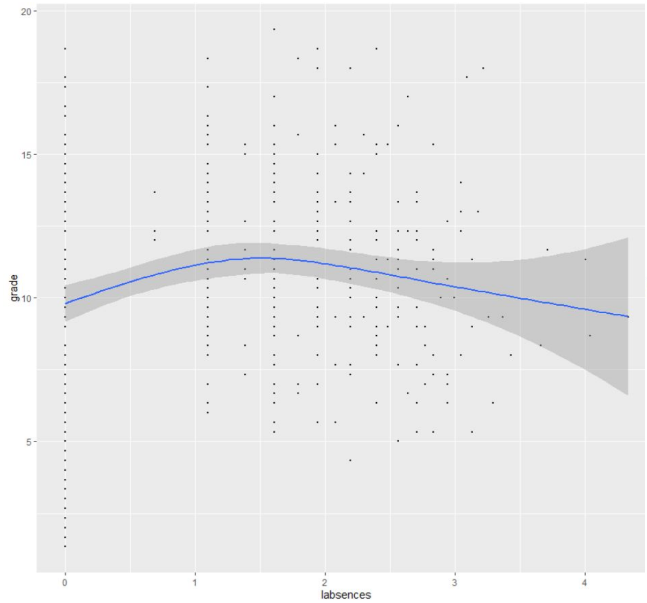
Important Predictors

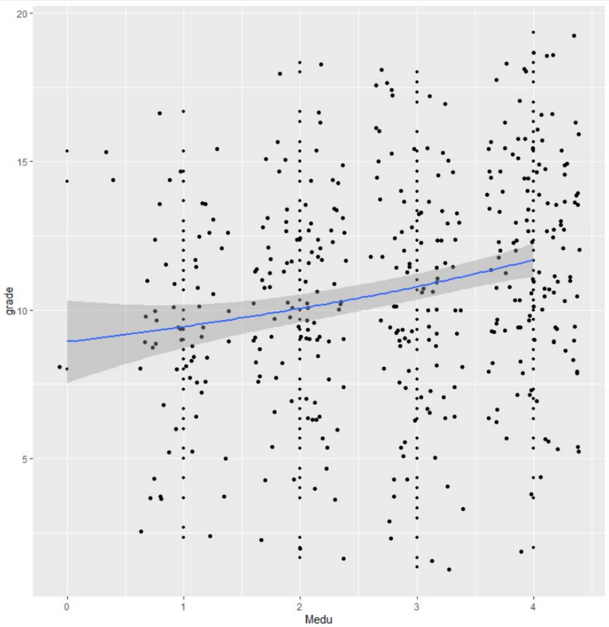
Overall Findings and Recommendations



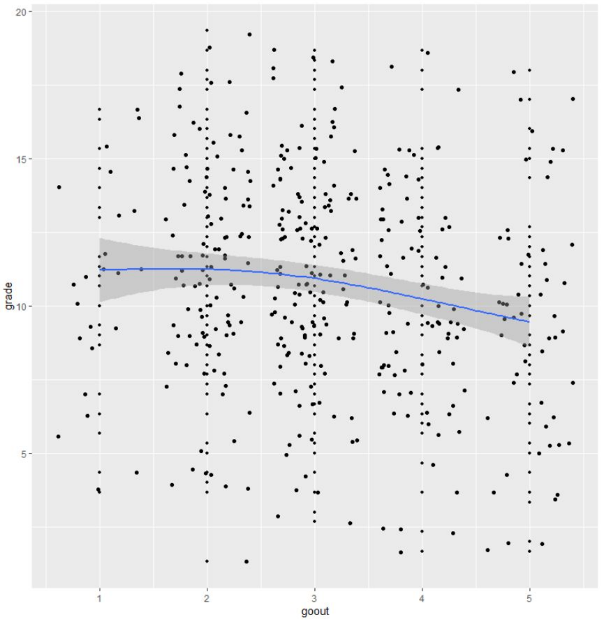
Deep Dive into the important predictors

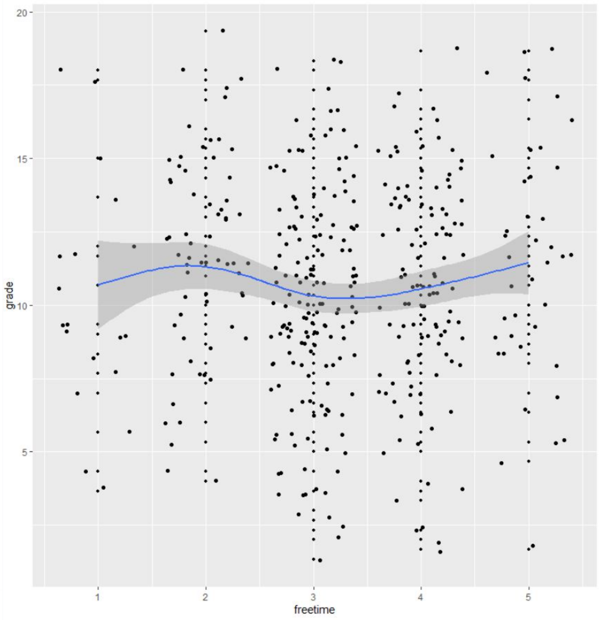


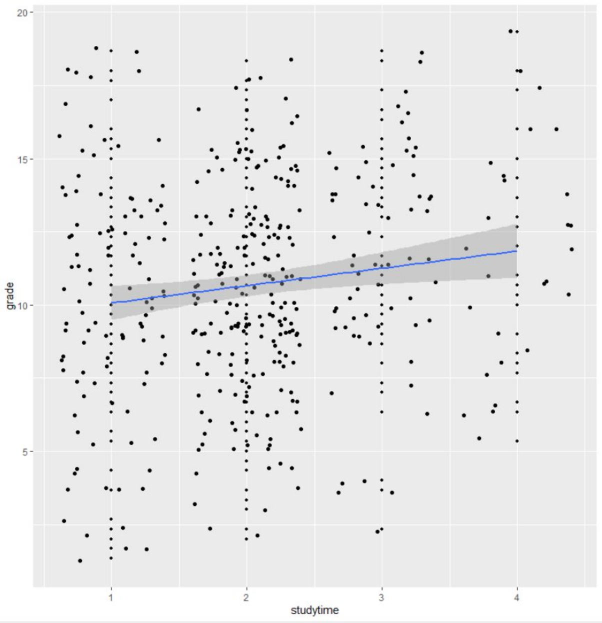




Deep Dive into the important predictors



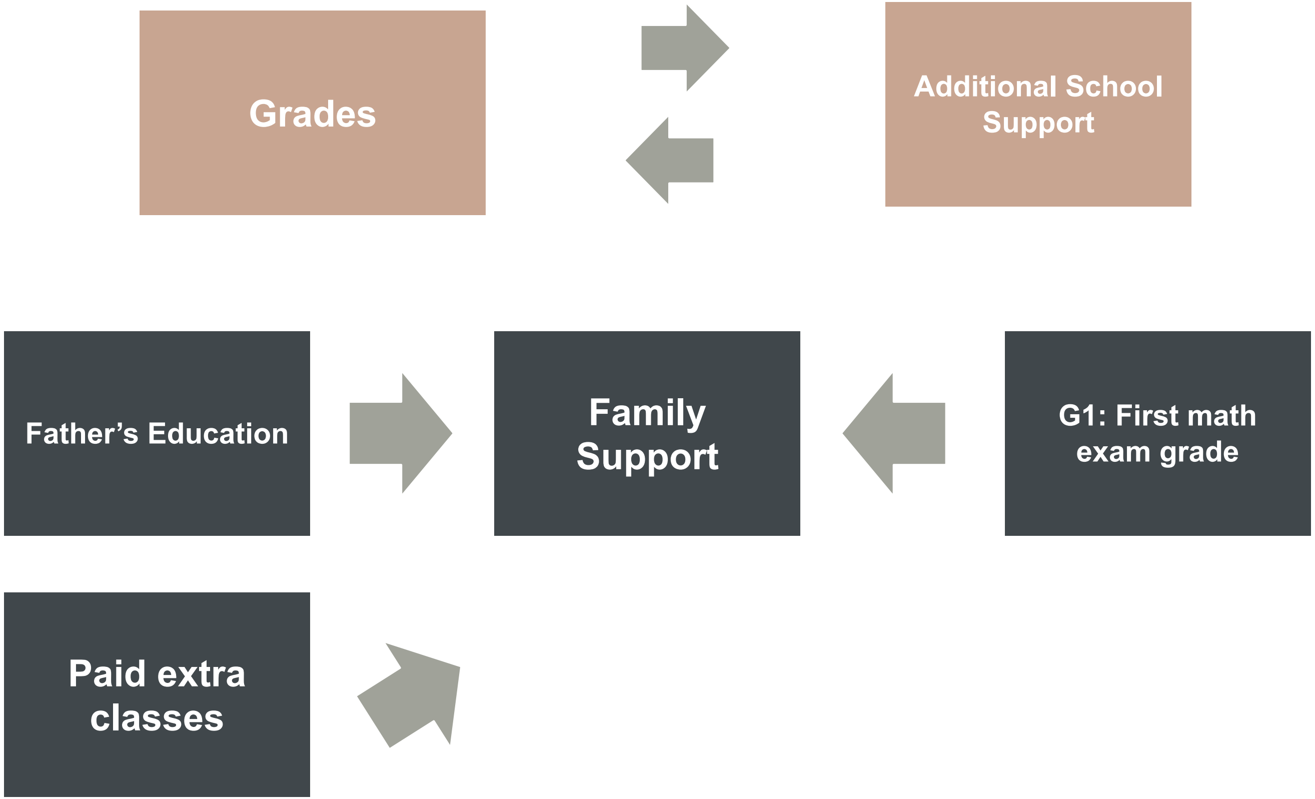




Important Predictors - Summary

- Past Failures has a strong, negative correlation with Math Grades
- Students that receive supplementary school support have a negative correlation with Math scores (opt in bias)
- Students that receive supplementary family support have a negative correlation with Math scores (opt in bias)
- Students who are engaged in extracurricular activities generally have lower Math scores
- The Amount of time spent studying has a positive correlation with Math scores
- Going out with Friends has a negative impact on Math grades

Do the variables affect each other?



Grades

Additional School Support

Father's Education

Family Support

G1: First math exam grade

Paid extra classes

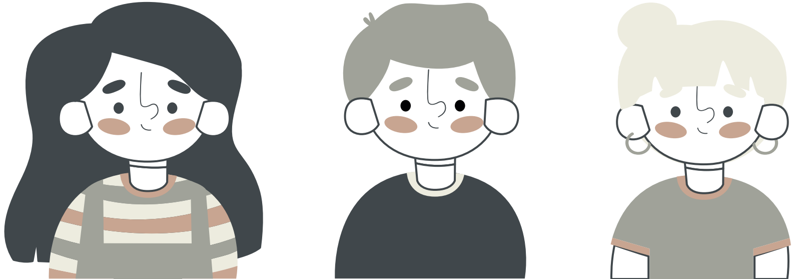
Agenda

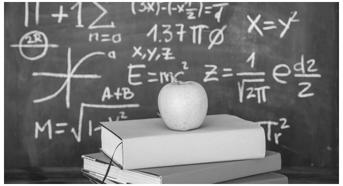
Data Exploration

Modeling

Important Predictors

Overall Findings and Recommendations





Overall Findings

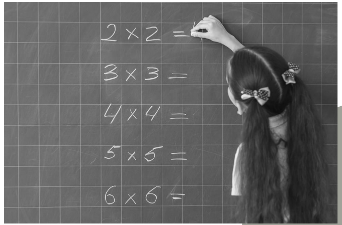
Predictors of a
Stellar Student

- Gender of student
- Study time
- Mother’s Job
- Absences*
- Free time

Predictors of an
Underperforming Student

- Past class failures
- Receiving supplemental school and family support
- Going out with friends

While there are no universal predictors of student performance the above demonstrated a strong relationship with a student’s math grades.



Recommendations

Areas of intervention:

GPA requirements to participate in extracurricular activities

Alcohol ed conferences to decrease usage

Higher education support and resources to help motivate students

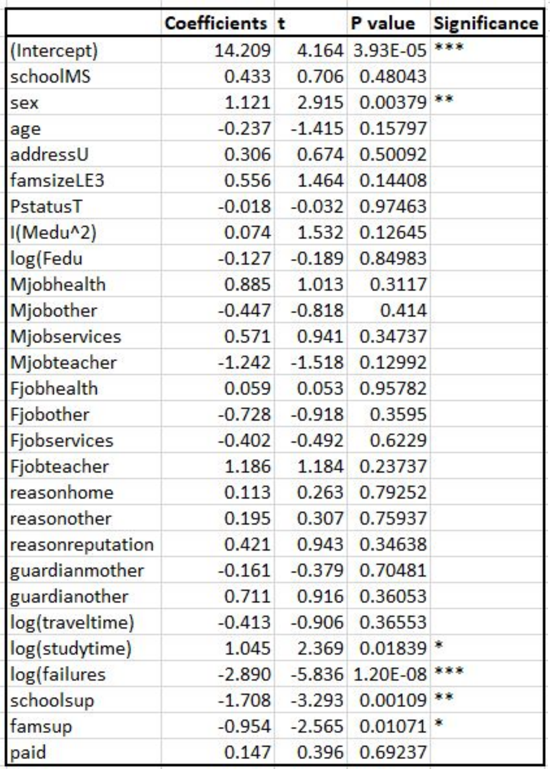
Earlier intervention for school and family support

Implementation of health and wellness days

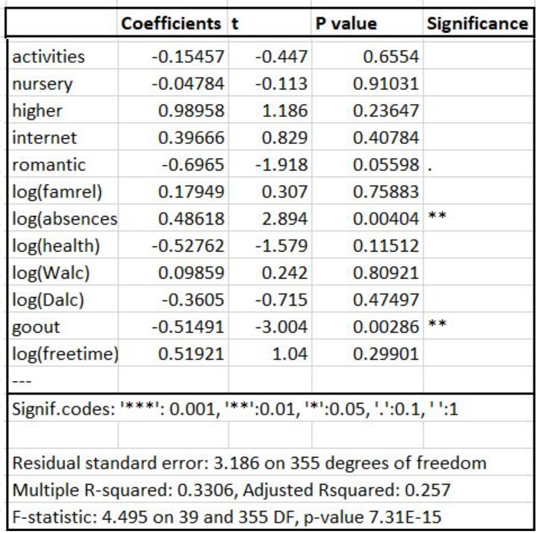
Further research: Collect data over time to construct panel data to monitor students more closely and generate causal inferences.

Appendix

Linear Regression



	Coefficients	t	P value	Significance
(Intercept)	14.209	4.164	3.93E-05	***
schoolMS	0.433	0.706	0.48043	
sex	1.121	2.915	0.00379	**
age	-0.237	-1.415	0.15797	
addressU	0.306	0.674	0.50092	
famsizeLE3	0.556	1.464	0.14408	
PstatusT	-0.018	-0.032	0.97463	
I(Medu^2)	0.074	1.532	0.12645	
log(Fedu	-0.127	-0.189	0.84983	
Mjobhealth	0.885	1.013	0.3117	
Mjobother	-0.447	-0.818	0.414	
Mjobservices	0.571	0.941	0.34737	
Mjobteacher	-1.242	-1.518	0.12992	
Fjobhealth	0.059	0.053	0.95782	
Fjobother	-0.728	-0.918	0.3595	
Fjobservices	-0.402	-0.492	0.6229	
Fjobteacher	1.186	1.184	0.23737	
reasonhome	0.113	0.263	0.79252	
reasonother	0.195	0.307	0.75937	
reasonreputation	0.421	0.943	0.34638	
guardianmother	-0.161	-0.379	0.70481	
guardianother	0.711	0.916	0.36053	
log(traveltime)	-0.413	-0.906	0.36553	
log(studytime)	1.045	2.369	0.01839	*
log(failures	-2.890	-5.836	1.20E-08	***
schoolsup	-1.708	-3.293	0.00109	**
famsup	-0.954	-2.565	0.01071	*
paid	0.147	0.396	0.69237	



	Coefficients	t	P value	Significance
activities	-0.15457	-0.447	0.6554	
nursery	-0.04784	-0.113	0.91031	
higher	0.98958	1.186	0.23647	
internet	0.39666	0.829	0.40784	
romantic	-0.6965	-1.918	0.05598	.
log(famrel)	0.17949	0.307	0.75883	
log(absences	0.48618	2.894	0.00404	**
log(health)	-0.52762	-1.579	0.11512	
log(Walc)	0.09859	0.242	0.80921	
log(Dalc)	-0.3605	-0.715	0.47497	
goout	-0.51491	-3.004	0.00286	**
log(freetime)	0.51921	1.04	0.29901	

Signif.codes: '***': 0.001, '**':0.01, '*':0.05, '.':0.1, ' ':1

Residual standard error: 3.186 on 355 degrees of freedom

Multiple R-squared: 0.3306, Adjusted Rsquared: 0.257

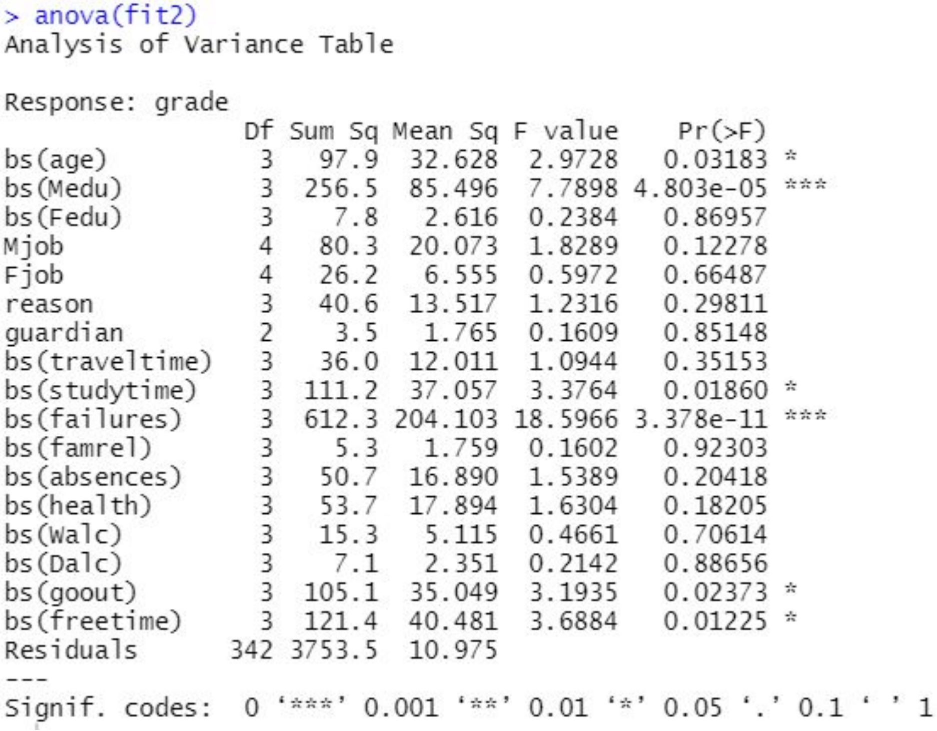
F-statistic: 4.495 on 39 and 355 DF, p-value 7.31E-15

A linear regression of Grade on all variables (after transformation) indicates the following:

- Log Failures is highly significant
- Sex, School Support, Log Absences and Go Out are moderately significant
- Log Study Time and Family Support are significant

However, the adjusted R squared is 25.7%, indicating that this may not be a good model for the data.

ANOVA For Non-Linearity

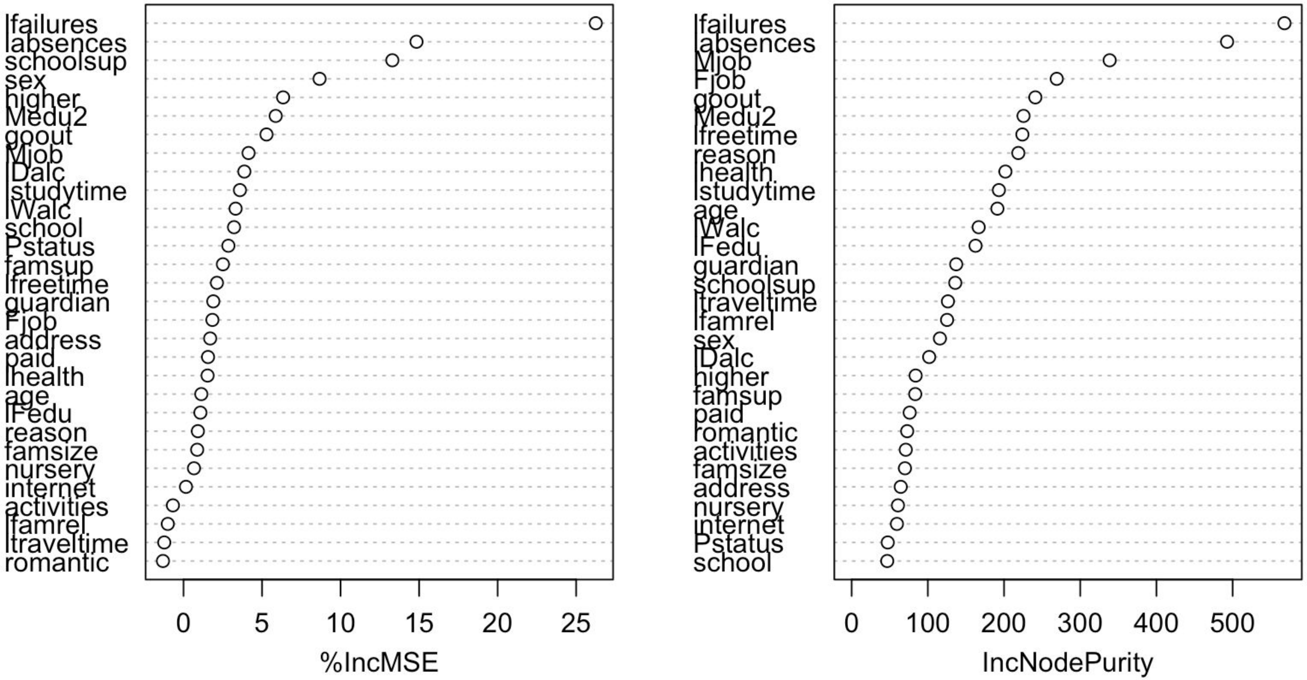


```
> anova(fit2)
Analysis of Variance Table

Response: grade
                Df Sum Sq Mean Sq F value    Pr(>F)    
bs(age)          3   97.9  32.628  2.9728   0.03183 *  
bs(Medu)         3  256.5  85.496  7.7898 4.803e-05 ***
bs(Fedu)         3    7.8   2.616  0.2384   0.86957    
Mjob             4   80.3  20.073  1.8289   0.12278    
Fjob             4   26.2   6.555  0.5972   0.66487    
reason           3   40.6  13.517  1.2316   0.29811    
guardian         2    3.5   1.765  0.1609   0.85148    
bs(traveltime)   3   36.0  12.011  1.0944   0.35153    
bs(studytime)    3  111.2  37.057  3.3764   0.01860 *  
bs(failures)     3  612.3 204.103 18.5966 3.378e-11 ***
bs(famrel)       3    5.3   1.759  0.1602   0.92303    
bs(absences)     3   50.7  16.890  1.5389   0.20418    
bs(health)       3   53.7  17.894  1.6304   0.18205    
bs(Walc)         3   15.3   5.115  0.4661   0.70614    
bs(Dalc)         3    7.1   2.351  0.2142   0.88656    
bs(goout)        3  105.1  35.049  3.1935   0.02373 *  
bs(freetime)     3  121.4  40.481  3.6884   0.01225 *  
Residuals      342 3753.5  10.975                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Age,
Mother’s education,
Studytime,
Failures,
Going out with friends
Free time,

have significant non-linear relationships with grade DV.

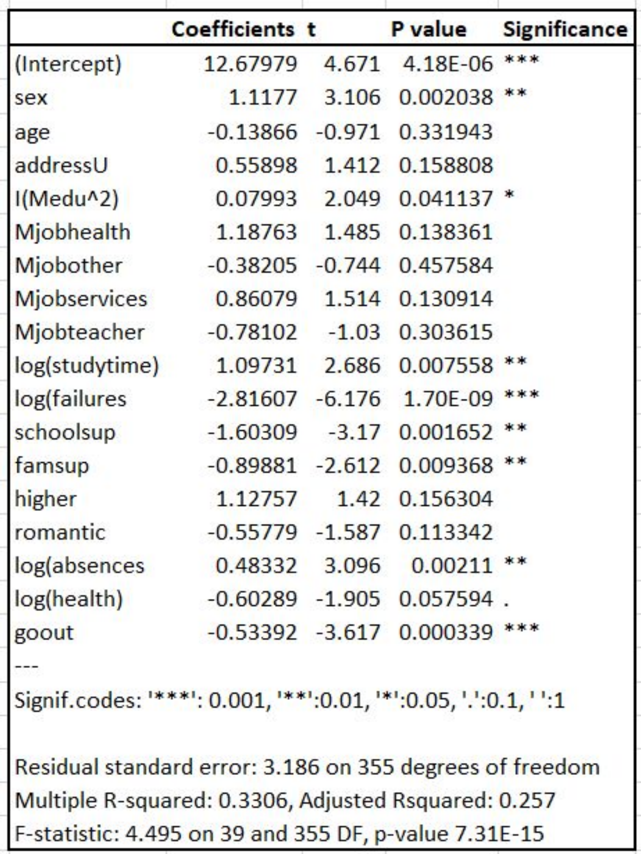
Random Forests with Grade DV



Principal Components Analysis

- In order to determine whether there are underlying latent variables in the data, we conducted a PCA analysis with Promax rotation.
- However, running a PCA analysis with 2, 3, 4, 5, and 6 factors did not yield satisfactory results, with very low communality values/high uniqueness values, and low overall cumulative variance explained.
- This makes sense as the majority of our original variables were not highly correlated.
- We decided not to go forward with PCA.

Stepwise



	Coefficients	t	P value	Significance
(Intercept)	12.67979	4.671	4.18E-06	***
sex	1.1177	3.106	0.002038	**
age	-0.13866	-0.971	0.331943	
addressU	0.55898	1.412	0.158808	
I(Medu^2)	0.07993	2.049	0.041137	*
Mjobhealth	1.18763	1.485	0.138361	
Mjobother	-0.38205	-0.744	0.457584	
Mjobservices	0.86079	1.514	0.130914	
Mjobteacher	-0.78102	-1.03	0.303615	
log(studytime)	1.09731	2.686	0.007558	**
log(failures	-2.81607	-6.176	1.70E-09	***
schoolsup	-1.60309	-3.17	0.001652	**
famsup	-0.89881	-2.612	0.009368	**
higher	1.12757	1.42	0.156304	
romantic	-0.55779	-1.587	0.113342	
log(absences	0.48332	3.096	0.00211	**
log(health)	-0.60289	-1.905	0.057594	.
goout	-0.53392	-3.617	0.000339	***

Signif.codes: '***': 0.001, '**':0.01, '*':0.05, '.':0.1, ' ':1

Residual standard error: 3.186 on 355 degrees of freedom

Multiple R-squared: 0.3306, Adjusted Rsquared: 0.257

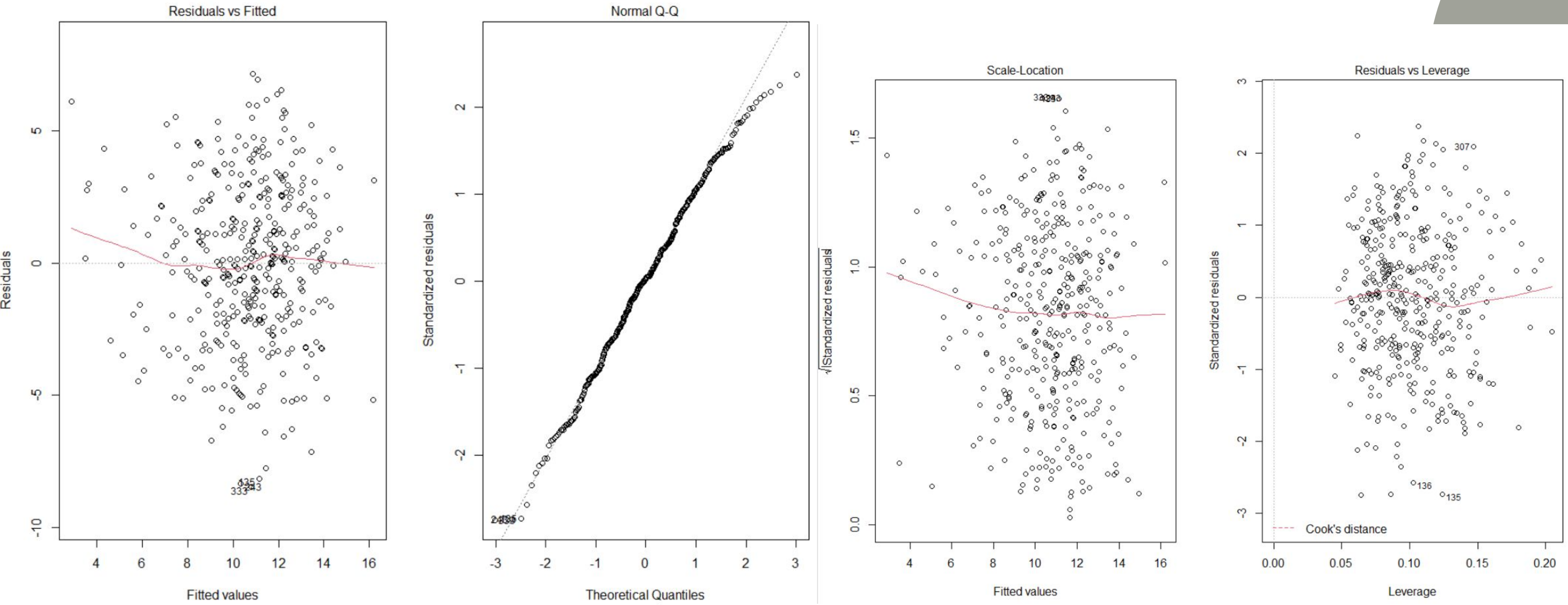
F-statistic: 4.495 on 39 and 355 DF, p-value 7.31E-15

A linear regression of Grade on stepwise selection on the previous model is highly significant.
The following is indicated by the model:

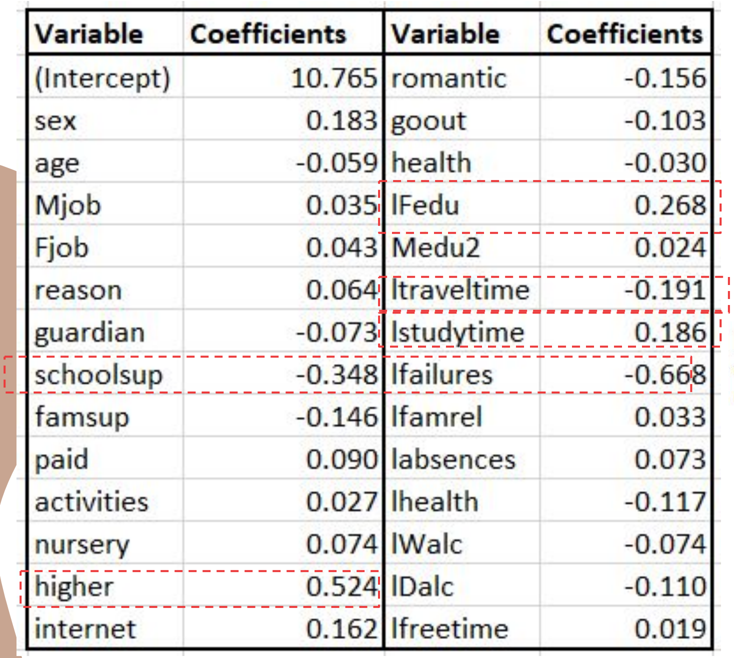
- Sex, Log transformed study time, number of failures, absences are significant.
- Mother's education significantly affects the student grade.
- Going out with friends, negatively impact the grades.
- Having an extra family and school support has a negative impact on grades.

Adjusted R squared is 25.7%

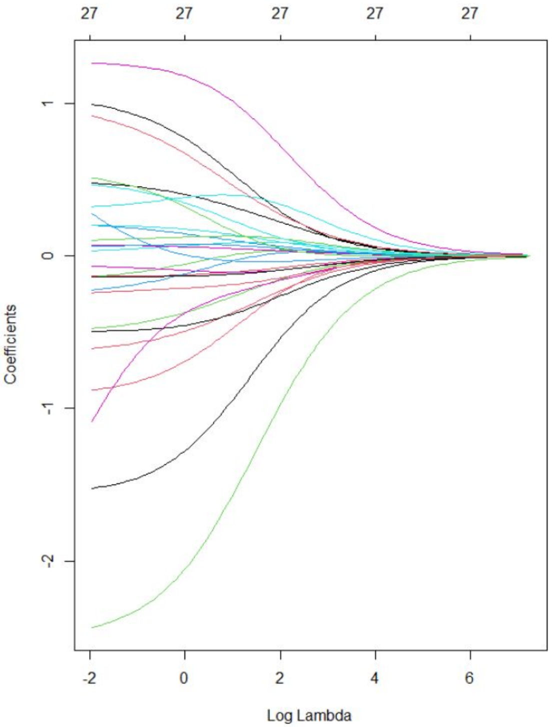
Diagnostics



Ridge

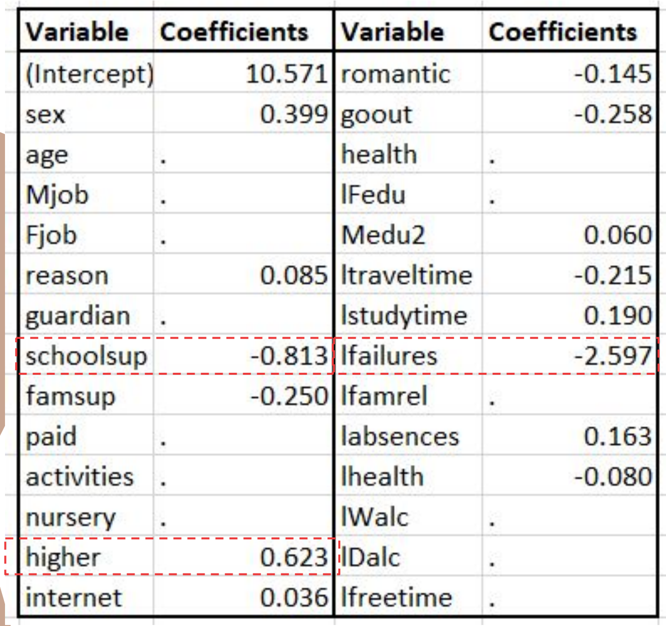


Variable	Coefficients	Variable	Coefficients
(Intercept)	10.765	romantic	-0.156
sex	0.183	goout	-0.103
age	-0.059	health	-0.030
Mjob	0.035	lFedu	0.268
Fjob	0.043	Medu2	0.024
reason	0.064	ltraveltime	-0.191
guardian	-0.073	lstudytime	0.186
schoolsup	-0.348	lfailures	-0.668
famsup	-0.146	lfamrel	0.033
paid	0.090	labsences	0.073
activities	0.027	lhealth	-0.117
nursery	0.074	lWalc	-0.074
higher	0.524	lDalc	-0.110
internet	0.162	lfreetime	0.019

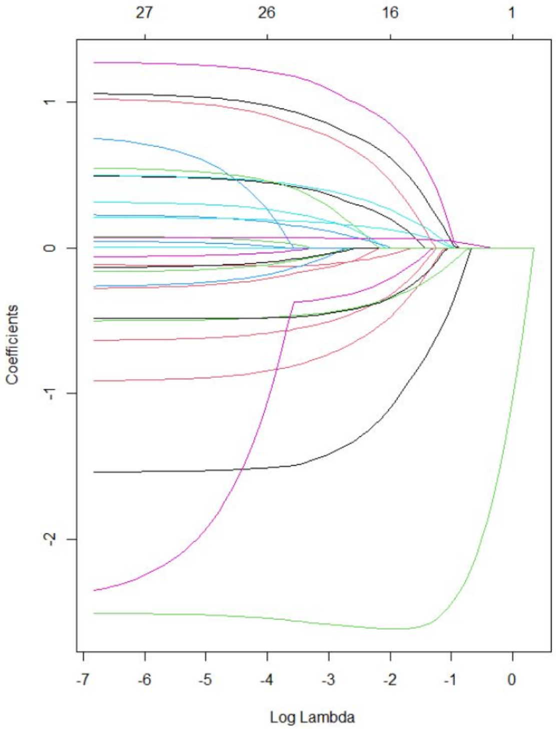


- Log of past number of class failures has the most significant effect on Grade
- This is closely followed by the motivation to pursue higher education
- Log father’s education has a moderately positive impact on grade, as does the log of study time
- Log travel time has a negative impact on Grade, perhaps limiting the amount of time students have to study

Lasso



Variable	Coefficients	Variable	Coefficients
(Intercept)	10.571	romantic	-0.145
sex	0.399	goout	-0.258
age	.	health	.
Mjob	.	lFedu	.
Fjob	.	Medu2	0.060
reason	0.085	ltraveltime	-0.215
guardian	.	lstudytime	0.190
schoolsup	-0.813	lfailures	-2.597
famsup	-0.250	lfamrel	.
paid	.	labsences	0.163
activities	.	lhealth	-0.080
nursery	.	lWalc	.
higher	0.623	lDalc	.
internet	0.036	lfreetime	.



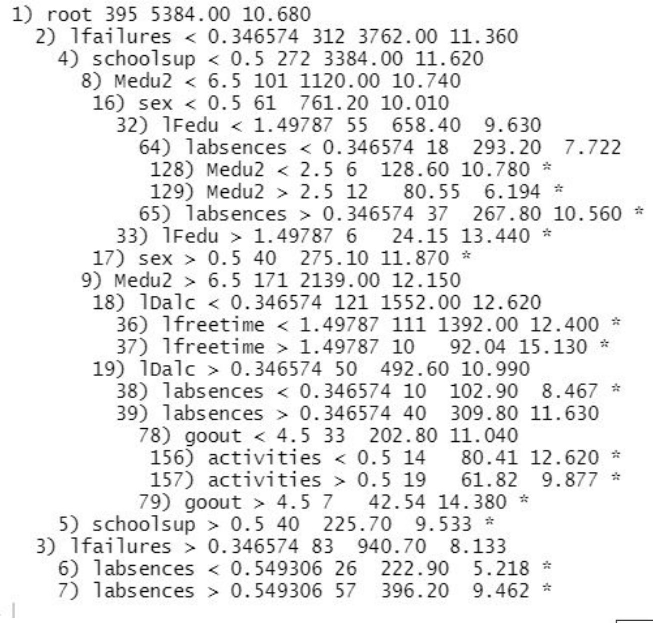
- Log of past failures has the highest negative impact on Grades
- School support also has a negative impact on Grades, likely due to students who have extra support are likely poor performers
- Wanting to pursue higher education has a positive impact on Grades

Tree

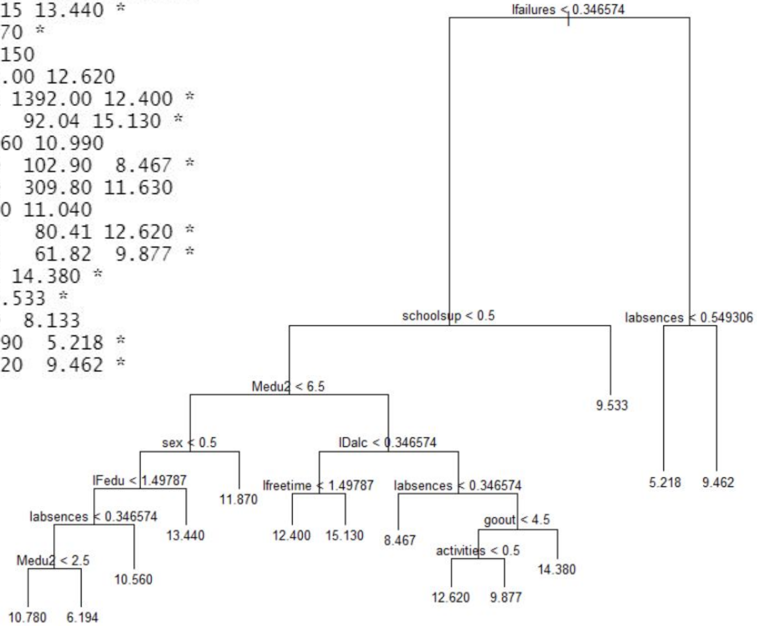
- Failures, school support, absences are the most important variables for grades.

- Mother’s education is the leading variable in terms of household/family related variables. In fact, it is more important than the student’s sex.

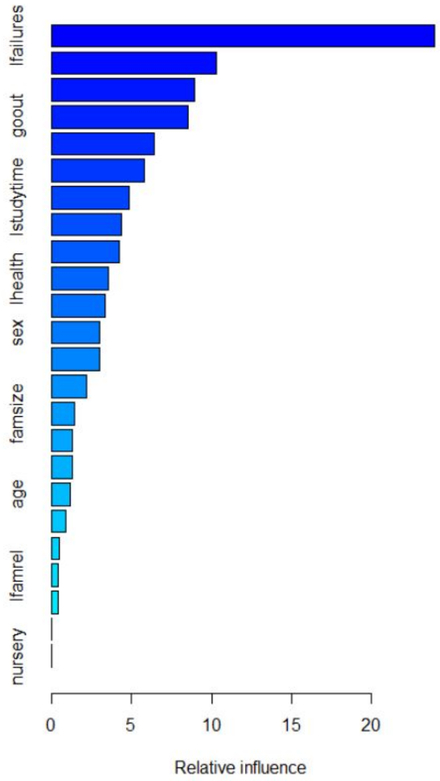
High Score student
Low Score student

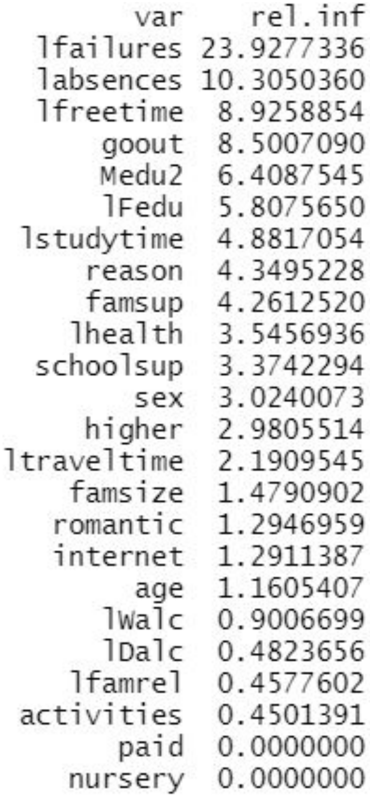


```
1) root 395 5384.00 10.680
  2) lfailures < 0.346574 312 3762.00 11.360
    4) schoolsup < 0.5 272 3384.00 11.620
      8) Medu2 < 6.5 101 1120.00 10.740
       16) sex < 0.5 61  761.20 10.010
         32) lFedu < 1.49787 55  658.40  9.630
           64) labsences < 0.346574 18  293.20  7.722
            128) Medu2 < 2.5 6  128.60 10.780 *
            129) Medu2 > 2.5 12   80.55  6.194 *
           65) labsences > 0.346574 37  267.80 10.560 *
         33) lFedu > 1.49787 6   24.15 13.440 *
       17) sex > 0.5 40  275.10 11.870 *
      9) Medu2 > 6.5 171 2139.00 12.150
       18) lDalc < 0.346574 121 1552.00 12.620
         36) lfreetime < 1.49787 111 1392.00 12.400 *
         37) lfreetime > 1.49787 10   92.04 15.130 *
       19) lDalc > 0.346574 50  492.60 10.990
         38) labsences < 0.346574 10  102.90  8.467 *
         39) labsences > 0.346574 40  309.80 11.630
           78) goout < 4.5 33  202.80 11.040
            156) activities < 0.5 14   80.41 12.620 *
            157) activities > 0.5 19   61.82  9.877 *
           79) goout > 4.5 7   42.54 14.380 *
    5) schoolsup > 0.5 40  225.70  9.533 *
  3) lfailures > 0.346574 83  940.70  8.133
    6) labsences < 0.549306 26  222.90  5.218 *
    7) labsences > 0.549306 57  396.20  9.462 *
```



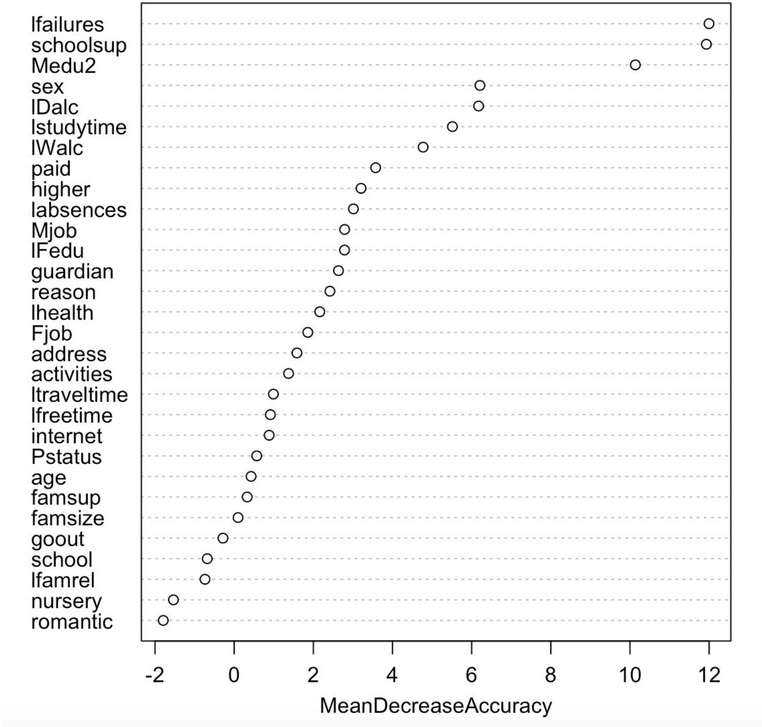
Boosted Tree

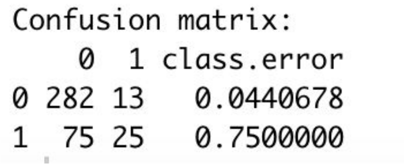




```
        var    rel.inf
  lfailures 23.9277336
  labsences 10.3050360
  lfreetime  8.9258854
      goout  8.5007090
      Medu2  6.4087545
      lFedu  5.8075650
 lstudytime  4.8817054
     reason  4.3495228
     famsup  4.2612520
    lhealth  3.5456936
  schoolsup  3.3742294
        sex  3.0240073
     higher  2.9805514
ltraveltime  2.1909545
    famsize  1.4790902
   romantic  1.2946959
   internet  1.2911387
        age  1.1605407
      lWalc  0.9006699
      lDalc  0.4823656
    lfamrel  0.4577602
 activities  0.4501391
       paid  0.0000000
    nursery  0.0000000
```

Random Forest with Pass DV

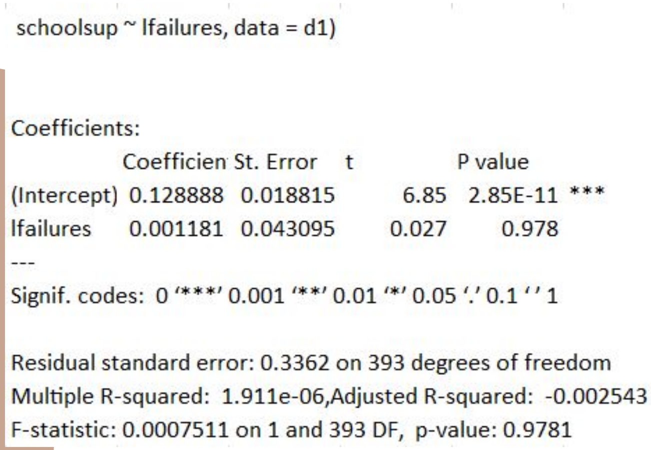




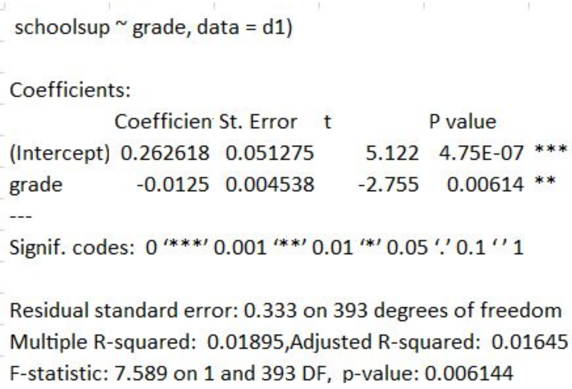
```
Confusion matrix:
    0  1 class.error
0 282 13   0.0440678
1  75 25   0.7500000
```

- False Positive Rate: 4.4%
- False Negative : 75%
- True Positive Rate: 25%
- Precision: 65.7%
- Classification error: 22.28%

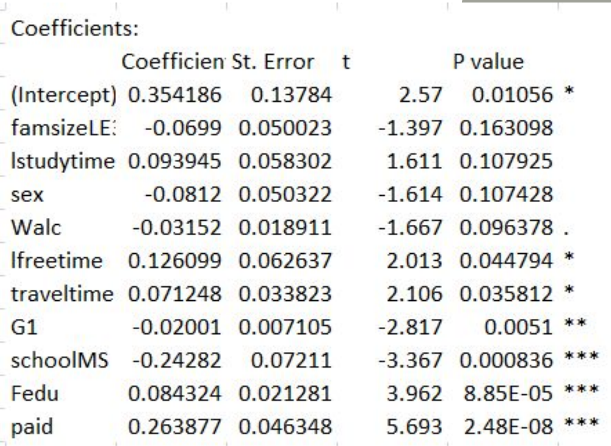
School and Family Support



```
schoolsup ~ lfailures, data = d1)


Coefficients:
            Coefficien St. Error    t              P value
(Intercept) 0.128888   0.018815          6.85   2.85E-11  ***
lfailures   0.001181   0.043095         0.027      0.978
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3362 on 393 degrees of freedom
Multiple R-squared:  1.911e-06,Adjusted R-squared:  -0.002543
F-statistic: 0.0007511 on 1 and 393 DF,  p-value: 0.9781
```



```
schoolsup ~ grade, data = d1)


Coefficients:
            Coefficien St. Error    t              P value
(Intercept) 0.262618   0.051275        5.122   4.75E-07  ***
grade        -0.0125   0.004538       -2.755    0.00614  **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.333 on 393 degrees of freedom
Multiple R-squared:  0.01895,Adjusted R-squared:  0.01645
F-statistic: 7.589 on 1 and 393 DF,  p-value: 0.006144
```



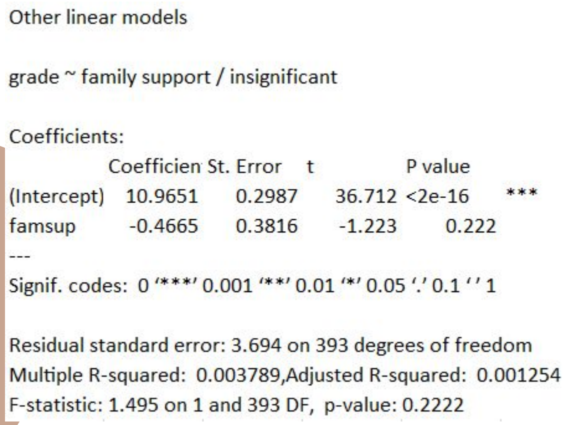
```
Coefficients:
            Coefficien St. Error    t         P value
(Intercept) 0.354186    0.13784      2.57    0.01056  *
famsizeLE:   -0.0699   0.050023    -1.397   0.163098
lstudytime  0.093945   0.058302     1.611   0.107925
sex          -0.0812   0.050322    -1.614   0.107428
Walc        -0.03152   0.018911    -1.667   0.096378  .
lfreetime   0.126099   0.062637     2.013   0.044794  *
traveltime  0.071248   0.033823     2.106   0.035812  *
G1          -0.02001   0.007105    -2.817     0.0051  **
schoolMS    -0.24282    0.07211    -3.367   0.000836  ***
Fedu        0.084324   0.021281     3.962   8.85E-05  ***
paid        0.263877   0.046348     5.693   2.48E-08  ***
```

Failures are insignificant. They do not have any relationship with schoolsupport.

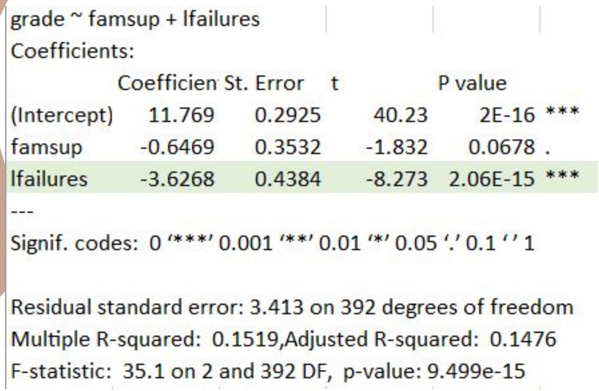
Grade is a significant predictor of school support. Each ten point decrease in grades make a student 12.5% more likely to receive school support.

Family support can be explained with freetime, traveltime, G1, school, Father's education,extra paid classes within the course subject.

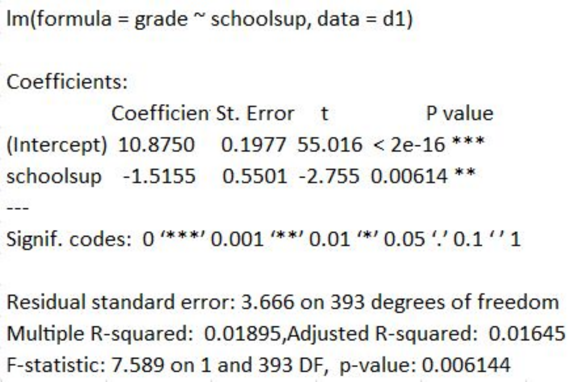
Additional Linear Runs



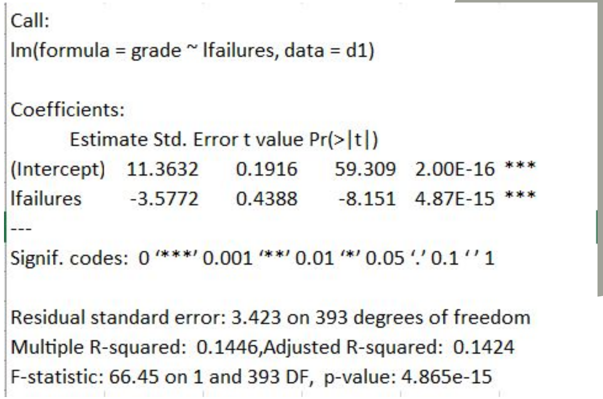
```
Other linear models

grade ~ family support / insignificant

Coefficients:
           Coefficien St. Error    t            P value
(Intercept)  10.9651     0.2987     36.712  <2e-16      ***
famsup       -0.4665     0.3816     -1.223       0.222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.694 on 393 degrees of freedom
Multiple R-squared:  0.003789,Adjusted R-squared:  0.001254
F-statistic: 1.495 on 1 and 393 DF,  p-value: 0.2222
```



```
grade ~ famsup + lfailures
Coefficients:
            Coefficien St. Error    t            P value
(Intercept)    11.769     0.2925      40.23       2E-16  ***
famsup        -0.6469     0.3532     -1.832      0.0678  .
lfailures     -3.6268     0.4384     -8.273    2.06E-15  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.413 on 392 degrees of freedom
Multiple R-squared:  0.1519,Adjusted R-squared:  0.1476
F-statistic:  35.1 on 2 and 392 DF,  p-value: 9.499e-15
```

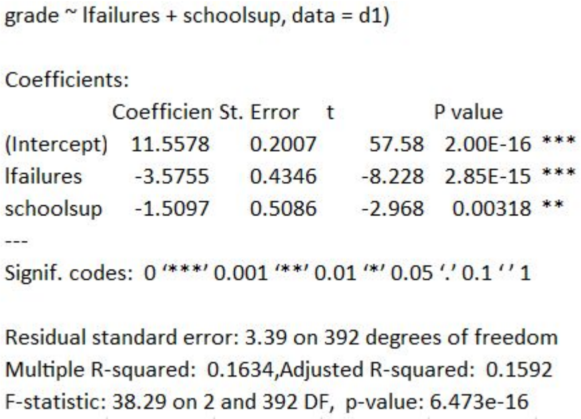


```
lm(formula = grade ~ schoolsup, data = d1)

Coefficients:
            Coefficien St. Error   t            P value
(Intercept)  10.8750    0.1977  55.016  < 2e-16 ***
schoolsup    -1.5155    0.5501  -2.755  0.00614 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.666 on 393 degrees of freedom
Multiple R-squared:  0.01895,Adjusted R-squared:  0.01645
F-statistic: 7.589 on 1 and 393 DF,  p-value: 0.006144
```



```
Call:
lm(formula = grade ~ lfailures, data = d1)

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.3632     0.1916     59.309  2.00E-16 ***
lfailures     -3.5772     0.4388     -8.151  4.87E-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.423 on 393 degrees of freedom
Multiple R-squared:  0.1446,Adjusted R-squared:  0.1424
F-statistic: 66.45 on 1 and 393 DF,  p-value: 4.865e-15
```

Family support either by itself or by controlling failures, doesn’t explain grades significantly. School support and failures both significantly affect grades and their effect do not change much while controlling for either.



```
grade ~ lfailures + schoolsup, data = d1)

Coefficients:
            Coefficien St. Error    t            P value
(Intercept)   11.5578     0.2007      57.58   2.00E-16  ***
lfailures     -3.5755     0.4346     -8.228   2.85E-15  ***
schoolsup     -1.5097     0.5086     -2.968    0.00318  **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.39 on 392 degrees of freedom
Multiple R-squared:  0.1634,Adjusted R-squared:  0.1592
F-statistic: 38.29 on 2 and 392 DF,  p-value: 6.473e-16
```