MATH 4397, Intro to Data Science & Machine Learning

UNIVERSITY of HOUSTON

Final Project Report

Student Performance On Two Secondary Classes

Fall, 2018

Instructors:

Cathy Poliak

Andrey Skipnikov

Team Members:

Joshua Lara

Austin Metcalf

Jennifer Pernia

Bao Tran

Gina Tran

Introduction

In this project we will take a look at two subjects, math and portuguese, taught at a secondary school in Portugal. Our inspiration for choosing this data set was, throughout our time in college we've encountered different people with different lifestyles. Within those lifestyles, student/school success has varied. It'd be interesting to get an inside look of how lifestyle affects student/school success in another country.

Our final goal is Inference. More specifically, we want to see what factors, if any, play a role in students' success. We will determine how each predictor affects the response. What predictors play the most significant roles in students' grades. What model best fit the data. For determining how significant each predictor and and its effect on the students final grade (G3), we will use the linear regression model. This model will also allow us to determine if we need to throw out any predictors that do not affect the response variable and we do that by looking for high p-value. By running linear regression first, it'll give us a better starting point to run the decision tree. With a disadvantage of decision trees being they do not have the same level of predictive accuracy as some other regression approaches. We will use them as a visual aid, which will allow us to see where the student went "wrong" or "right" in terms of their lifestyle choices and how they affected their final grades, as trees can easily handle quantitative and qualitative predictors.

Before running any models we noticed within our data there were instances that some students had zeros for either G1, G2, G3 which didn't affect the other grades. So we removed those rows by

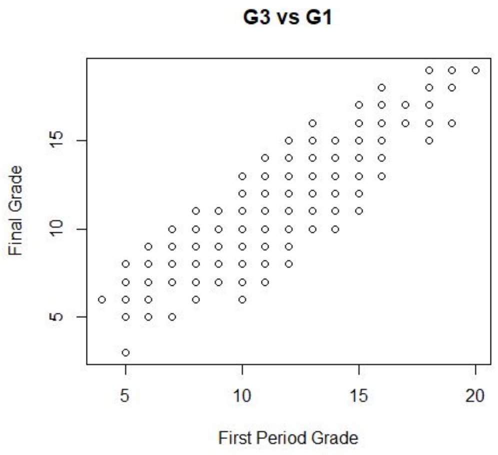
```
student <- student[-which(student$G3 == "0" ), ]
```

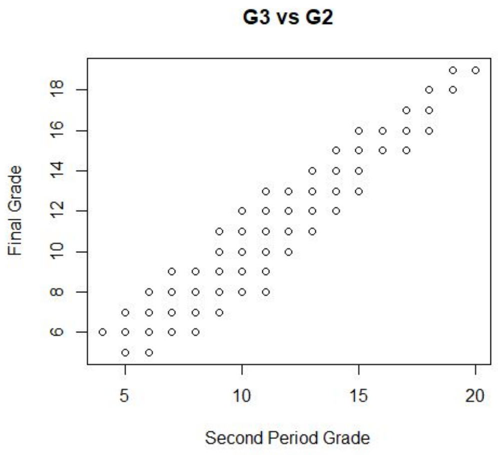
deleting the rows that included zeros for G3 in both data sets, took care of the problem. Before deleting those rows our significant predictors for the math students were age, family relationship, and absences. After, gave us a couple more significant predictors such as failures, freetime, goout, traveltime, studytime. Also since passed zeros for the grades such as G1 and G2 didn't affect the students final grade (G3) we took that as G1 and G2 not having any input on the calculation of G3. So we included G1 and G2 as significant predictors as well.

Math Linear Regression Model

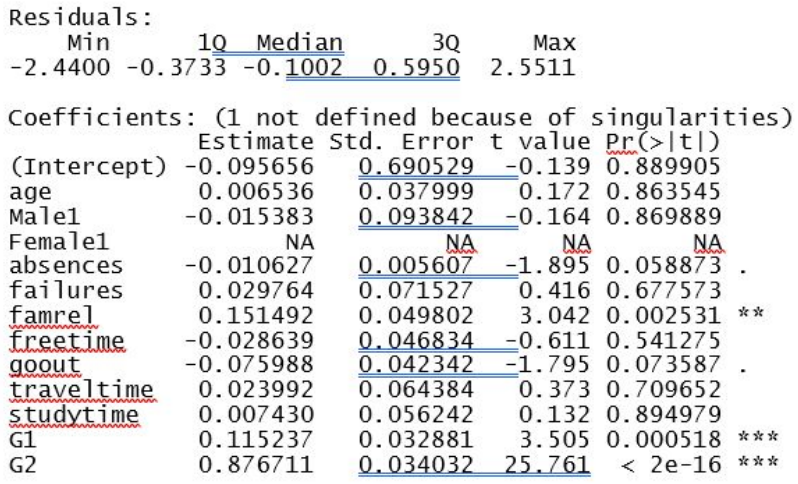
First, I perform some exploratory analysis:

I removed instances whose G3 = 0 because it does not make sense to have G3=0.





This shows that all G1, G2 and G3 are strongly related with each other. This makes sense, a student who does well in first period also do well in the second period. This results in comparable final grade. G3 has a positive linear relationship with G1 and G2.



```
Residuals:
    Min      1Q  Median      3Q     Max
-2.4400 -0.3733 -0.1002  0.5950  2.5511

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.095656   0.690529  -0.139 0.889905
age          0.006536   0.037999   0.172 0.863545
Male1       -0.015383   0.093842  -0.164 0.869889
Female1            NA         NA      NA       NA
absences    -0.010627   0.005607  -1.895 0.058873 .
failures     0.029764   0.071527   0.416 0.677573
famrel       0.151492   0.049802   3.042 0.002531 **
freetime    -0.028639   0.046834  -0.611 0.541275
goout       -0.075988   0.042342  -1.795 0.073587 .
traveltime   0.023992   0.064384   0.373 0.709652
studytime    0.007430   0.056242   0.132 0.894979
G1           0.115237   0.032881   3.505 0.000518 ***
G2           0.876711   0.034032  25.761  < 2e-16 ***
```

$\widehat{Y}$ = 0.095656 + 0.006536 • age + 0.015383 • male - 0.010627 • absences + 0.151492 • famrel - 0.028639 • freetime - 0.075988 • gout + 0.023992 • traveltime + 0.007430 • studytime + 0.115237 • G1 + 0.876711 • G2

Famrel (quality of family relationship), G1(first period grade), and G2 (second period grade) are significant predictors.

As you can see the predictors that had a negative affect were amount of times the student was absent, amount of free time they had, and time spent going out. While the predictors with having a positive affect were the age of the student, travel time to and from school, amount of studytime, G1, and G2.

Portuguese linear regression model equation:

Call:
lm(formula = G3 ~ age + traveltime + failures + higher + goout +
    Dalc + health + G1 + G2 + is.mteacher, data = stdpor)

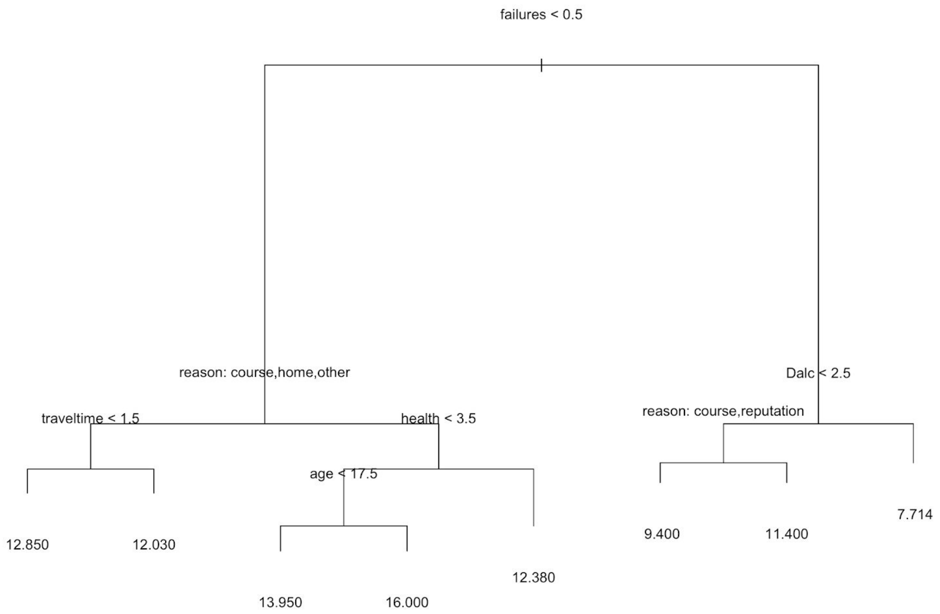
Formula:
$\widehat{G3}$ = -1.067 + 0.1397 • age + .10067 • traveltime - 0.22832failures + 0.26505 • higher - 0.04615 • goout - 0.10169 • Dalc - .04799 • health + 0.20238 • G1 + 0.73929 • G2 + 0.20171 • is.mteacher

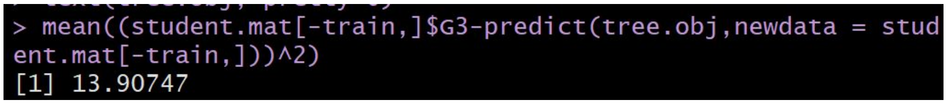
After a stepwise selection on the data of portuguese schools, we find that this subset of predictors best explain the data. I also numericalized the categorical variables so that they could be included in the regression, variable “is.mteacher” corresponds to the mother being a teacher. Variables such as failures in previous classes, frequency of going out with friends, and daily alcohol consumption all negatively influence the final grade. Other variables such as G1, G2, age and the mother being a teacher all positively influence the final grade.

Decision Trees:

While running the decision trees, G1 and G2 dominated the tree so we decided to get rid of them in the main tree. The first tree shown is the unpruned tree for the math students. After deleting G1 and G2 the trees main split is whether or not the student had any past failures.

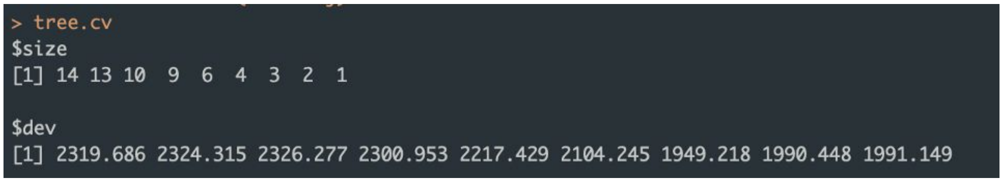
Math:





```
> mean((student.mat[-train,]$G3-predict(tree.obj,newdata = stud
ent.mat[-train,]))^2)
[1] 13.90747
```

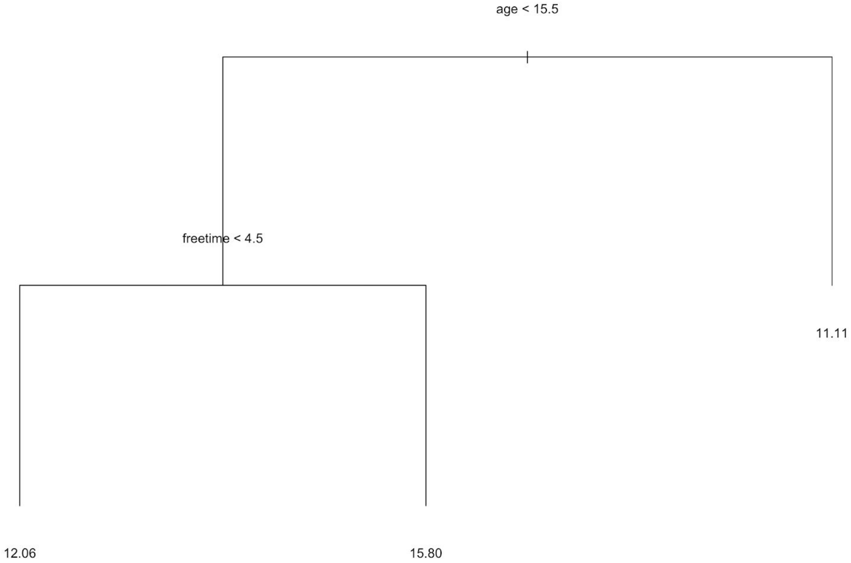
The mean squared error for the unpruned tree is ~14.

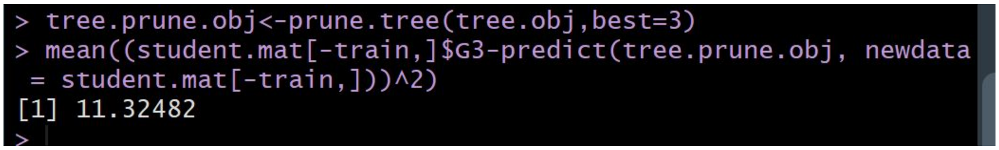


```
> tree.cv
$size
[1] 14 13 10  9  6  4  3  2  1

$dev
[1] 2319.686 2324.315 2326.277 2300.953 2217.429 2104.245 1949.218 1990.448 1991.149
```

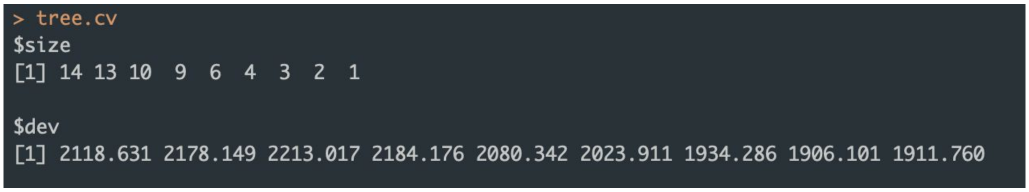
Here we see the optimal tree size is 3.





```
> tree.prune.obj<-prune.tree(tree.obj,best=3)
> mean((student.mat[-train,]$G3-predict(tree.prune.obj, newdata
 = student.mat[-train,]))^2)
[1] 11.32482
>
```

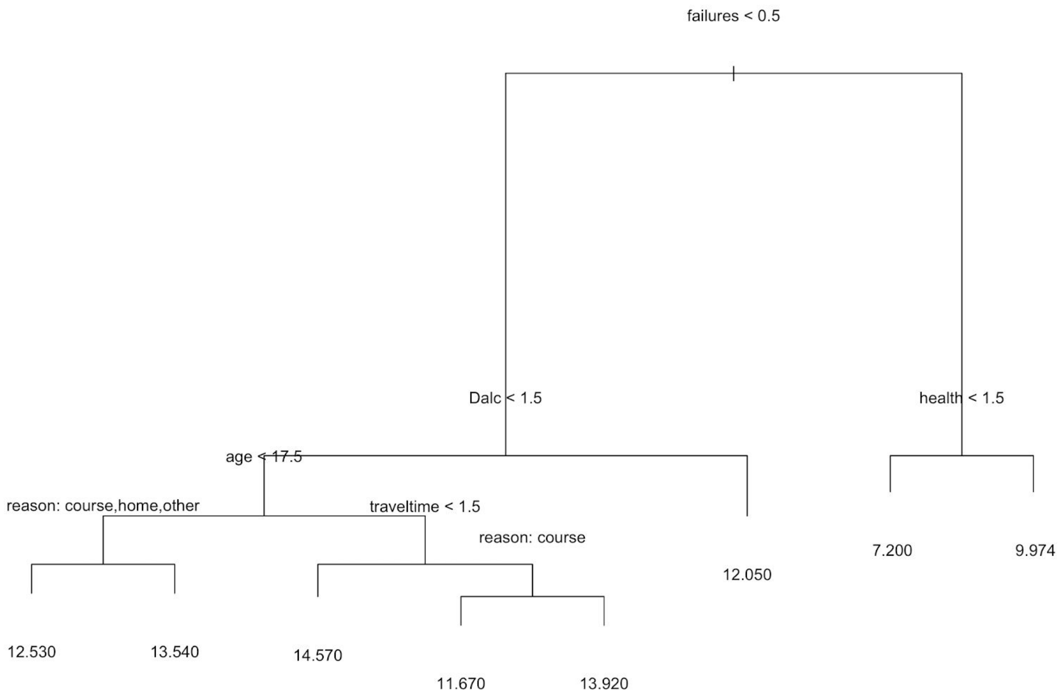
Pruned tree shows much better than the unpruned tree.

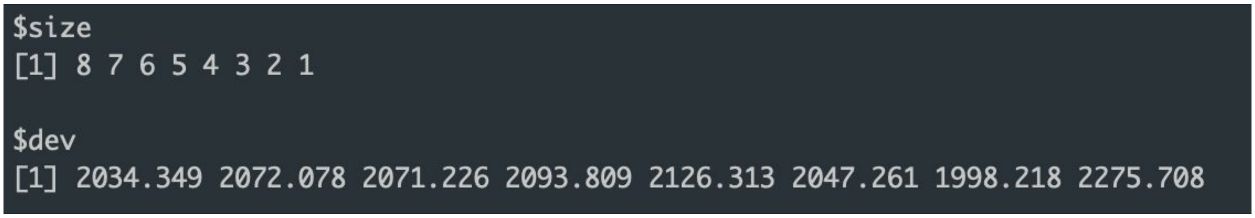


```
> tree.cv
$size
[1] 14 13 10  9  6  4  3  2  1

$dev
[1] 2118.631 2178.149 2213.017 2184.176 2080.342 2023.911 1934.286 1906.101 1911.760
```

Portuguese:

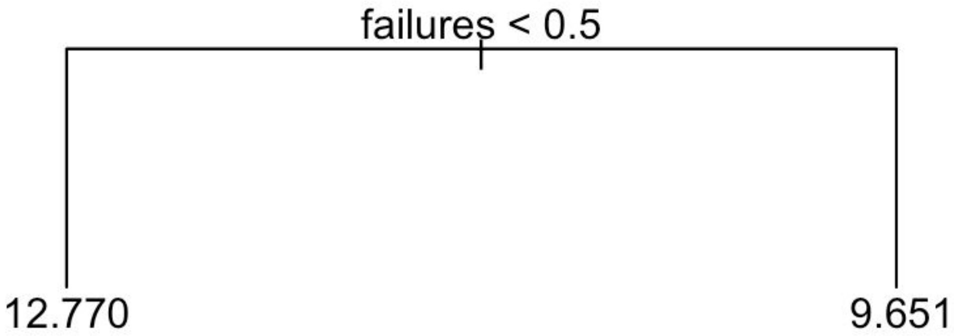


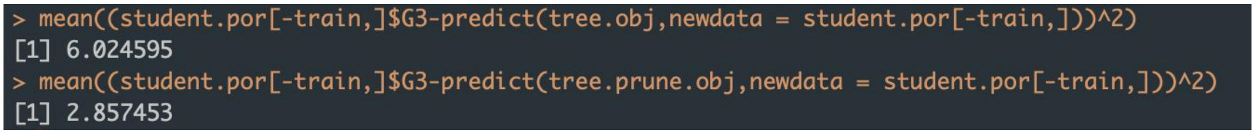


```
$size
[1] 8 7 6 5 4 3 2 1

$dev
[1] 2034.349 2072.078 2071.226 2093.809 2126.313 2047.261 1998.218 2275.708
```

We can see that the optimal tree size is 2.





```
> mean((student.por[-train,]$G3-predict(tree.obj,newdata = student.por[-train,]))^2)
[1] 6.024595
> mean((student.por[-train,]$G3-predict(tree.prune.obj,newdata = student.por[-train,]))^2)
[1] 2.857453
```

After pruning the tree we see that the mean squared error for the pruned tree (6.02) is better than the unpruned tree (2.85).

Conclusion:

After modeling our data with both linear regression and decision trees, we concluded that using linear regression gives us a better fit for our data as well as a more concrete answer to our original question.For the project we used inference to measure the factors that would affect the student's overall performance. After excluding all the predictors that did not affect our results and analyzed only the predictors which caused an effect on the final grade of the student such as travel time to school, number of absences and gender. We then used decision trees for our data to again determine how much each predictor affected student performance.

We started with 2 unpruned trees which were pruned to obtain lower variance and a better interpretation for the data. For the math decision tree we can conclude that the age predictor plays a big role in the students overall grade.For the portuguese decision tree failures of past courses seemed to be the most important factor. While in linear regression we can see both the negative and positive effects of the predictors.

In math the positive predictors were: age,travel time to school, and amount of study time while the negative factors included being absent, amount of free time and time spent going out.In portuguese the positive predictors included   age and having a mother as a teacher,while negative predictors included failures in previous courses, alcohol consumption and time spent going out.

Overall linear regression turned out to be better in terms of performance when our interest is inference. The decision trees proved to be better for interpretation purposes.