Reason Behind Good Grades

Saiful Alam
Math 4397 Intro to Data Science & Machine Learning
Prof. Cathy Poliak & Andrey Skripnikov
12/5/2018

Introduction :

In some points of life, we all are curious to know the secret recipe of doing well in school. Obviously, as college students, we should as of now have an idea of some critical qualities of doing great in school. In our dataset we look at how different circumstances affect a student's grade. Our data set describes various types of information of a student and also grades of students taking a Portuguese class from two Portuguese schools. Other attributes in this dataset include student demographics, families information, activities outside of school and in schools.

Our objective for this Project is to figure out what factors influence a student's grade in a class. We might want to see whether factors, for example, study time, travel time, free time and past failures will have an effect on anticipating how well a student does in school. For this project, we will explore some regression methods in order to be able to accurately predict final grades based on various predictors.

Question and Methodology :

For this project, we will try to identify the significant factors that affect a student's grade in a class and determine which of the two methods that will be used can more accurately predict our response variable, G3. We will be using two models in this project: A linear regression model and a regression tree model.

Linear regression is an attempt to model the relationship between two or more variables by fitting a linear equation to the data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. From this dataset, we may be interested in whether there is a relationship between two or more variables, with one of the variables being the

final grade a student has in the class. Every model has some advantages and drawbacks, here we will discuss it briefly. The linear regression model tries to learn best fit line (equation) that will have the least mistakes in predicting a response variable. The model has high efficiency however there are situations where this high efficiency is actually a disservice where it might be inclined to become sensitive to certain data (i.e some noisy data also considered as useful data). Furthermore the performance of this approach may not be good if the relationship between the predictors and the response variables are not linear.

Regression tree is similar to a decision tree, except it is returning a continuous value. Briefly, we will go over advantages and disadvantages of decision trees. Regression trees implicitly perform variable screening or feature selection. Regression trees require relatively little effort from users for data preparation. Nonlinear relationships between parameters do not affect tree performance.

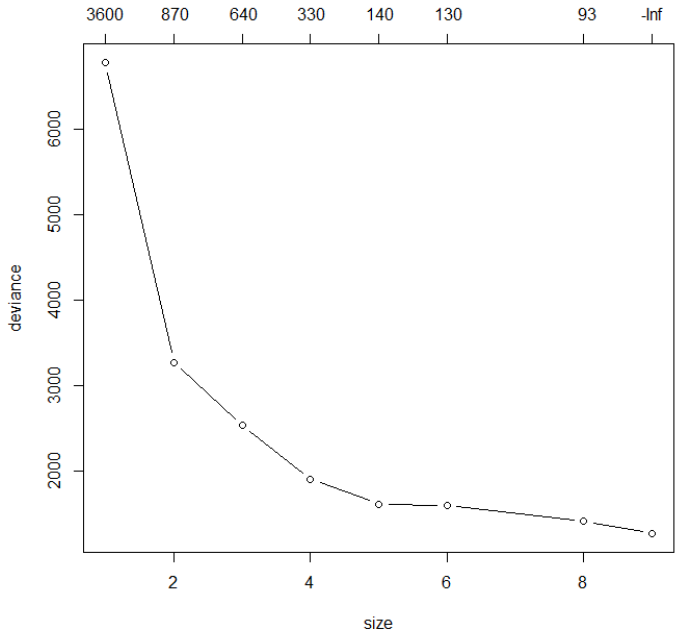
Disadvantages: Preparing regression trees, especially large ones with many branches, are complex and time-consuming affairs. Overfitting occurs when the algorithm captures noise in the dataset and the model becomes too complex. The prediction model can get unstable if there is very small variance in data that it is being fed with. A highly complicated regression tree tends to have a low bias which makes it difficult for the model to work with new data.

Regression Tree Model:

Final grade (G3) ~ school + age + address + famsize + Pstatus +Medu + Fedu + Mjob + Fjob + traveltime + studytime + failures + reason + absences + guardian + G2 + traveltime + studytime + schoolsup + famsup + paid + activities + nursery + higher + internet + romantic + famrel + freetime + goout + Dalc + Walc + health +G1

We built a tree using the formula above and we got a tree of size 9. Next we tried to determine whether we can prune the tree created and get a better performing tree based on the

error calculated for each pruned tree. After trying out different sized trees we get the following plot.



This plot shows that the tree we originally created did not need any pruning. We tried to determine whether pruning needed so that we could avoid any overfitting in our regression tree model. Overfitting is a situation where our model becomes complex and fits to the data used to train it, making its prediction training error low. While this may not sound bad, it may cause the model to be unable to perform well in predicting the response variable when data foreign to the training data is introduced. In order to prevent our model from overfitting, we try to perform pruning which is a method where we decrease the size of a decision or regression tree in order to make that model less complex and improve performance with a test data.

Linear Regression Model:

Final grade (G3) = 0.28723 + 0.13193 * G1 + 0.87295 * G2 + 0.37780 * MHealthJob + 0.27589 * MServiceJob+ (-0.27155) * failures + 0.36336 * RCourse + (0.13802) * traveltime

Before creating this model, we converted factor variables into multiple binary attributes For example, we converted the variable MJob into multiple binary attributes, such as MHealthJob and MServiceJob, where each describe whether a student's mother has a particular job like "Health" and "Teacher." This is done so that we can better understand which factors are significant when predicting the G3 variable.

While we were fitting our linear regression model, we first did an initial linear fit using the lm function in R-studio to see the p-values and importance of all the predictors in our dataset. We observed that the above variables are the most important in predicting the final grade. Of course, we already had a feeling that G1 and G2, first period and second period grades respectively, would be two of the most important predictors in finding out the final grade. The variable "failures" is also one of the more obvious important predictors as it indicates how often a student failed in the past, meaning you can use that knowledge to predict their grade. The other variables in this formula also have a significant impact to how well this model can predict the variable G3.

Prediction Comparison (Jinman Cai):

After determining the formulas for each of the models, we will begin training our data with 10 training and calculate a prediction error with 10 test sets. After doing this process, we get the following table of prediction test errors.

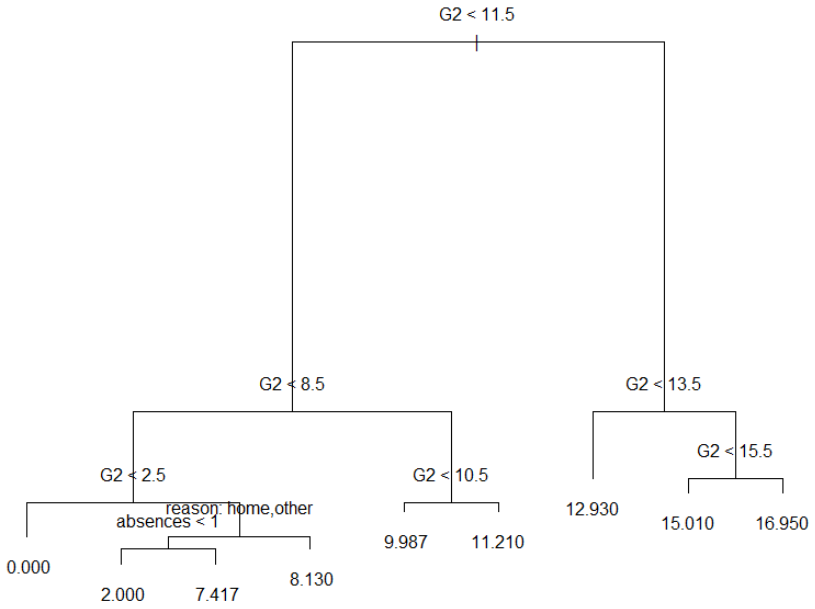
#	Linear Regression test errors	Regression Tree test errors
1	0.9084631	1.578674

2	1.4650270	1.640331
3	1.8278552	2.136596
4	2.2431624	2.853522
5	1.1731068	1.577371
6	3.0047451	3.354380
7	1.2877643	2.153509
8	1.9053144	2.341962
9	1.4954142	2.009846
10	2.3851480	2.515650
Mean	1.7696	2.216184

Based on the table, we can conclude that Linear Regression is more accurate than Regression tree. Looking at the mean errors, we can see that the value for Linear Regression errors is 1.7696 and the value for Regression Tree errors is 2.216184. Also during each iteration, the Regression Tree errors value is higher than Linear Regression errors.

Inference Comparison :

Regression Tree:



This regression tree represents a series of splits starting at the top of the tree. The top split assigns observations having whether G2 < 11.5. We see that the tree is dominated with splits containing the G2 attribute. By looking at the regression tree, many features other than G2 were not used by the tree. The only other features used for this tree other than G2 were reason and absences.

Linear Regression:

```
Call:
lm(formula = G3 ~ ., data = Student)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6605 -0.5137  0.0163  0.5986  5.4777

Coefficients: (4 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.499e-01  1.004e+00  -0.149 0.881309
age          3.174e-02  4.812e-02   0.660 0.509823
studytime    4.609e-02  6.609e-02   0.697 0.485869
traveltime   1.498e-01  7.369e-02   2.032 0.042544 *
failures    -2.609e-01  9.843e-02  -2.650 0.008247 **
famrel      -2.094e-02  5.446e-02  -0.384 0.700765
goout       -3.484e-02  4.763e-02  -0.731 0.464811
Dalc        -6.230e-02  7.147e-02  -0.872 0.383753
Walc        -6.905e-03  5.505e-02  -0.125 0.900224
health      -5.396e-02  3.613e-02  -1.494 0.135820
absences     1.441e-02  1.168e-02   1.233 0.218041
```

```
G1           1.292e-01  3.760e-02   3.437 0.000629 ***
G2           8.703e-01  3.488e-02  24.953  < 2e-16 ***
MTeacherJob  2.067e-01  1.861e-01   1.110 0.267228
MHealthJob   2.765e-01  2.131e-01   1.298 0.194943
MServiceJob  2.358e-01  1.423e-01   1.658 0.097848 .
MHomeJob     1.253e-01  1.398e-01   0.896 0.370419
MOtherJob           NA         NA      NA       NA
FTeacherJob -1.856e-01  2.362e-01  -0.786 0.432292
FHealthJob  -6.051e-02  2.878e-01  -0.210 0.833576
FServiceJob -1.111e-01  1.212e-01  -0.917 0.359530
FHomeJob     3.329e-01  2.135e-01   1.560 0.119384
FOtherJob           NA         NA      NA       NA
RHome        2.848e-01  1.898e-01   1.501 0.133948
RReputation  1.825e-01  1.964e-01   0.929 0.353240
RCourse      3.599e-01  1.721e-01   2.092 0.036891 *
ROther              NA         NA      NA       NA
GMother     -2.560e-01  2.276e-01  -1.125 0.261228
GFather     -2.138e-01  2.489e-01  -0.859 0.390737
GOther              NA         NA      NA       NA
school      -1.969e-01  1.276e-01  -1.543 0.123241
sex         -1.444e-01  1.167e-01  -1.238 0.216189
famsize     -2.386e-02  1.146e-01  -0.208 0.835135
address      1.153e-01  1.226e-01   0.940 0.347513
Pstatus     -8.695e-02  1.617e-01  -0.538 0.590935
schoolsup   -1.804e-01  1.730e-01  -1.043 0.297517
famsup       9.023e-02  1.065e-01   0.847 0.397404
activities   5.976e-06  1.042e-01   0.000 0.999954
nursery     -1.059e-01  1.266e-01  -0.836 0.403285
higher       2.136e-01  1.815e-01   1.177 0.239652
internet     7.188e-02  1.289e-01   0.557 0.577439
romantic    -5.172e-02  1.076e-01  -0.481 0.630896
paid        -1.963e-01  2.134e-01  -0.919 0.358204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.248 on 610 degrees of freedom
Multiple R-squared:  0.8595,    Adjusted R-squared:  0.8507
F-statistic: 98.16 on 38 and 610 DF,  p-value: < 2.2e-16
```

The output for the linear regression model shows G3 which was the final outcome. The reason as for why some of the variables show NA is because they are dependent on other variables. MOtherJob, FOtherJob, ROther, GOther, are dependent on the other variables in the outcome. The variables that are significant are travel time, failures, G1, G2, MServiceJob and RCourse. The R-square is 0.8595 and we can conclude a well accuracy of our predictions. In most domains explaining 8^ of the variance of a relationship is considered good, but there are contexts where it might be considered barely adequate.

Conclusion :

In this project, we were able to determine many factors that affect student grades and to what extent. From the results we got from the models we created, we saw that G1 and G2 (grades throughout the semester), as expected, has a significant influence in predicting a student's final grade, G3. We also saw some somewhat unexpected attributes be significant such as travel time and the fact whether a student's mother has a job in the service industry.

The two models we used, linear regression and regression tree, had varying prediction errors when trying to predict the final grade of each object in the different test sets used. From the table in the Prediction Comparison section, we saw that overall, the linear regression model had a better performance than the regression tree model.

One model is better for prediction while another model is better for interpretation. The regression tree model is better for interpretation. When there are more than two explanatory variables, the earlier partition diagrams become more difficult to draw, but tree representation can be extended to any dimension. Regression trees imitate the way that humans naturally think about things and make decisions.