Calvin Makelky

A01872013

STAT 5810 Final Project

**Data Set Information**

This data via UC Irvine details student performance in secondary education mathematics of two Portuguese schools. The data attributes include student grades, demographic, social, and school related features. The data was collected using school reports and questionnaires. The response variable is G3, with 0 being the worst possible grade, and 20 the best possible grade in math. The other attributes are a mix of continuous, binary, nominal, and ordinal variables. The goal is to predict G3 using these variables.

**Attribute Information**

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2 sex - student's sex (binary: 'F' - female or 'M' - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: 'U' - urban or 'R' - rural)

5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)

9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13 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)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

31 G3 - final grade (numeric: from 0 to 20, output target)

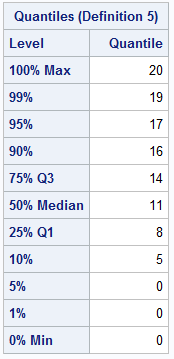
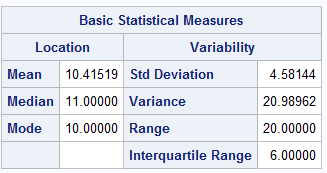
**Random Forests for G3 unchanged**

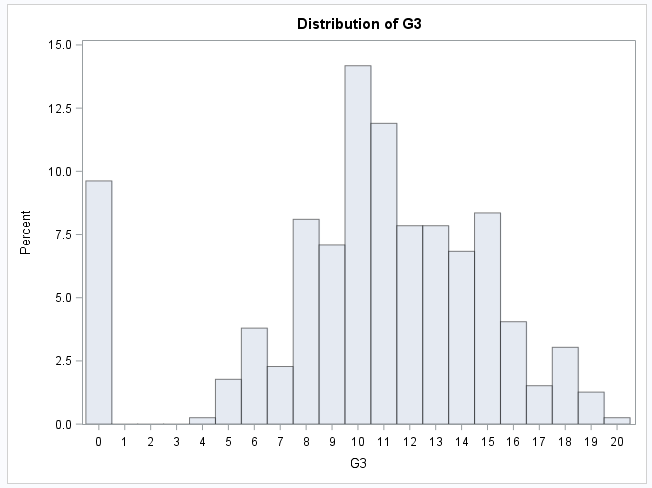
With 21 different possible final grades, it's extremely unlikely there is a good parsimonious model to predict for all the final math grades. Consequently, leaving G3 unchanged, I ran a random forests on the data to see if any variables are important with G3 having so many levels. If any are important, I will do classification trees and random forests with the suggested variables. Else I will skip to slimming down the levels of G3 into more easily interpreted grade groups.



Unsurprisingly, none of the predictor variables are important with G3 unchanged. They all had negative, and thus unimportant, out-of-bound Gini and margin values. I will now look at the distribution of final math grades to see how best to group them.

**Response Variable G3 for Math**



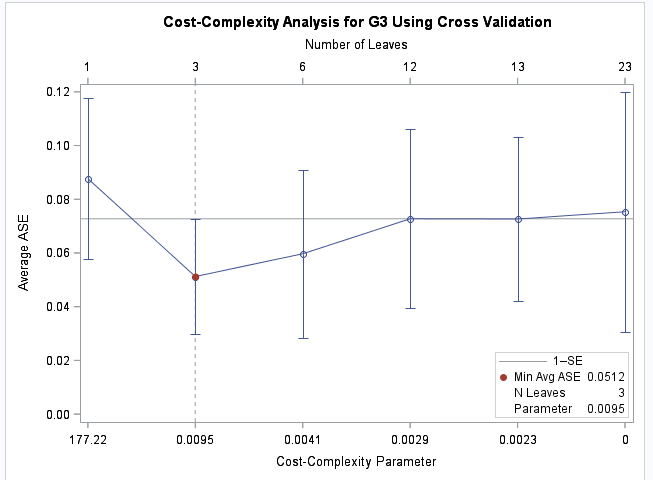


Most of the final grades for math class (G3), the response variable, seem to be between 8 and 16. It looks very normally distributed if it were not for the students who received a final grade of 0. The middle 50% of final grades are from 8 to 14 (not including endpoints). The average grade was 10.4, and the median grade was 11. Due to the response variable being ordinal with 21 possible outcomes, it makes sense to try group the grades into fewer groups in order to improve the prediction and interpretability of the data. The high frequency of 0 final grades being so distinct from the rest of the grades is very interesting. Thus, part I will group by zero and nonzero grades. Part II will have three groups: one from 0-8, one from 9-13, and one from 14-20 since the middle 50% (average grades) are from 9-13. This will leave below average and above average to each get a quarter of all observations

**Part I: Zero vs Non-Zero Final Grades**

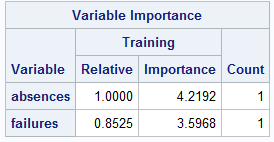
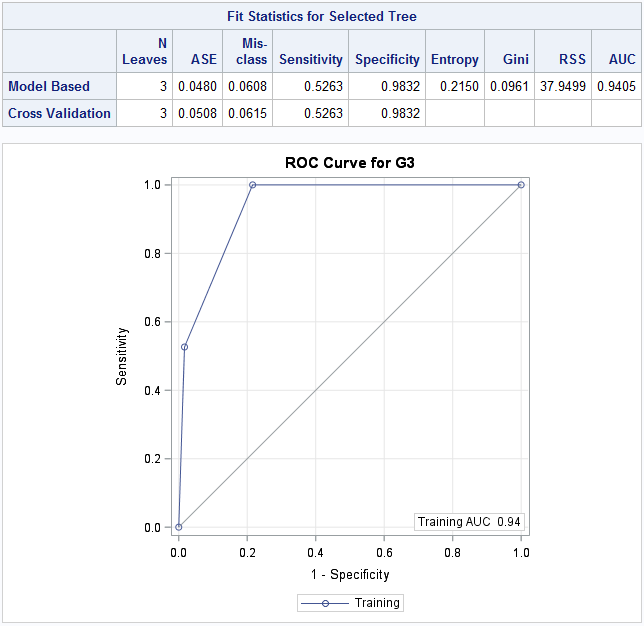
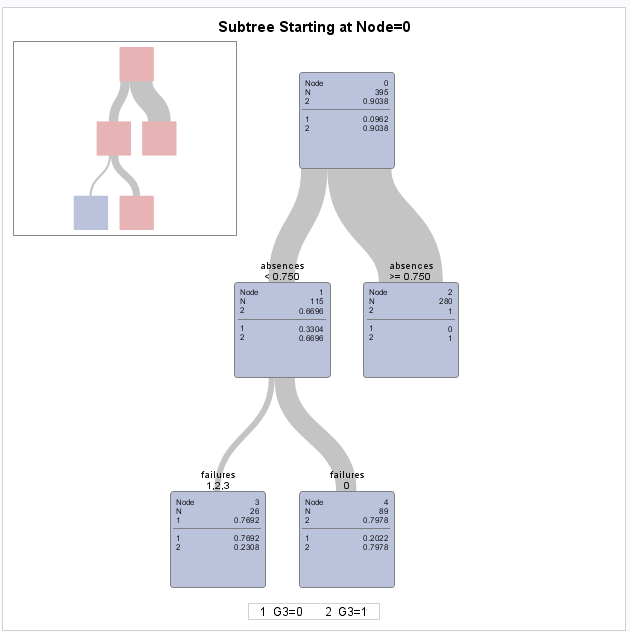
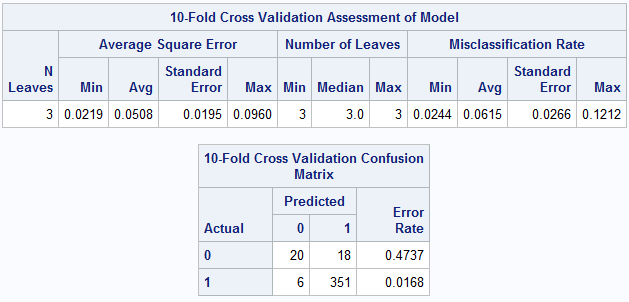
Grouping by nonzero and zero is the easiest, I’ll start with that. Here I will try to answer the question of “**what variables are important in predicting that a student in these two Portugese schools will absoutely fail in math?”** and **“ what model is the best for predicting this?”** First, I will use classification trees and then random forests to anwer these questions.

**Classification Trees with Failures Variable**



Using the 1-SE rule, I choose to do a classification tree with three leaves.

**Three Leaf Classification Tree with Failures variable**

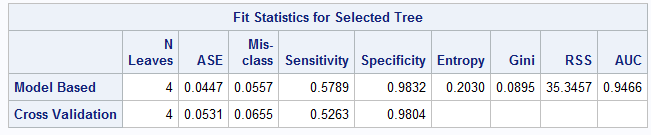


This three leaf classifiation tree, with 10-fold crossvalidation, predicting whether the student would get a final of grade of zero or not wonderfully well. The misclassification trate was only 6.15%, which means the PCC was 93.85%. That means the tree correctly predicted whether a student would get a zero or not nearly 94% of the time. That’s extremely good, anything above 90% PCC is excellent. Additionally, the AUC was an excellent 0.9405. The confusion matrix shows 351 of the 357 students who did not get a zero were correctly predicted to not get zero. This is incredibly good. However, the much smaller amount of students who got zeros, 38, were not predicted well. Only 20 of the 38 students were correctly predicted as receiving zeros. This results in a very high misclassifcation rate of 47.37%. This goes with the incredibly high specificity, 98.32%, but very poor sensitivity, 52.63%.

The classification tree only used and split off two different variables: absenses (number of school absenses) and failures (number of past class failures). The first split was off absenses. If the student had greater or qual to 0.75 (1 or more) absenses, the student was predicted to not get a zero. This is where most of the students ended up, as all of the 280 studsents with an absense did not get zeros. This goes against what one would expect completely, maybe it is just a coinedence that all 38 students who received zeros did not have an absesnse. The next and final split is on failures. If the student had greater or equal to 1 or more past failures, the student was predicted to receive a zero. If the student did not have any past failures, the student was predicted to not get a zero. This makes a lot more sense, students who have failed a class in the past are more likely to fail this math class than those he have never failed a class. Each of these final two terminal nodes of the tree correctly predicted the remaning students with just under 80% accuracy. The variable importance table shows that absences was the mot important variable in classifiying zero grades, with failures slightly less important. **To sum up the tree’s results, if a student doesn’t have an absenses and has failed a class in the past, that student is likely (about 80% chance) to get a zero in math, otherwise the student is not likely to get a zero in math in these two secondary schools in Portugal.** The model overall is very good good, but the poor sensitivity (predicting zeros when actually zeros) could use much improvement.

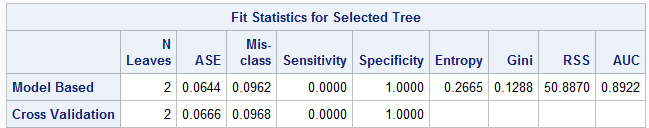
**Six Leaf Classification Tree with Failures variable**

Fitting different leaved classification trees did not improve the sensitivity problem. The 6 leaf tree below, which the costcomlexity analysis regarded as the 2nd best option, has a slightly higher misclassification rate, 6.55%, yet the exact same sensitivity as the three leaf model.

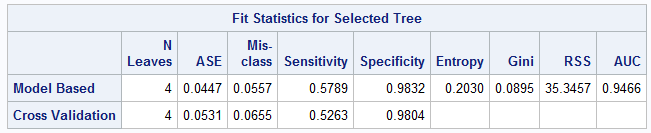
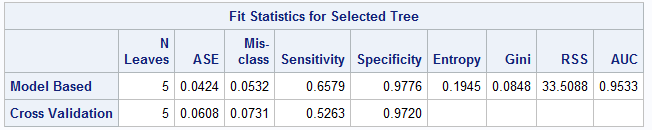


**Two Leaf Classification Tree with Failures variable**

The two leaf tree below was significantly worse than the three and six leaf trees. It’s misclassification rate was nearly 10% while the sensitivity dropped from roughly 53% in the previous trees to 0% in this model.

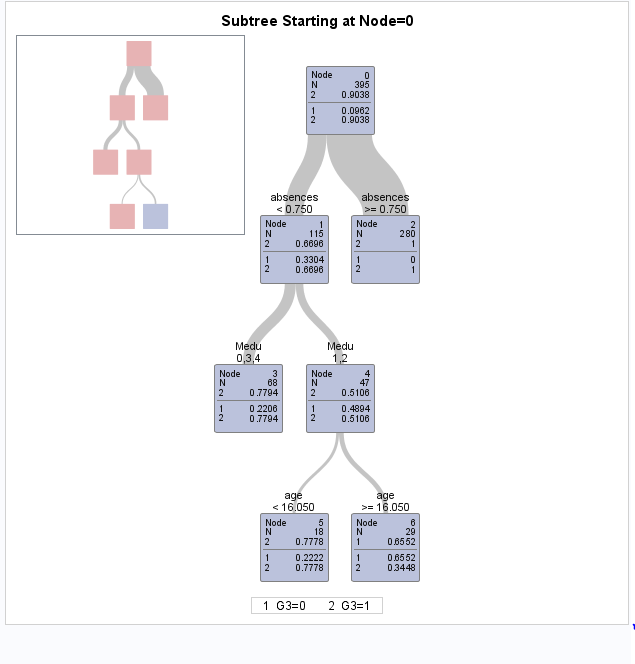
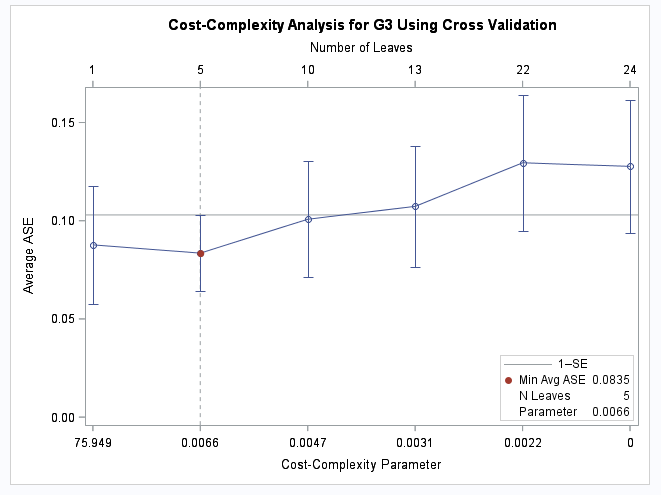
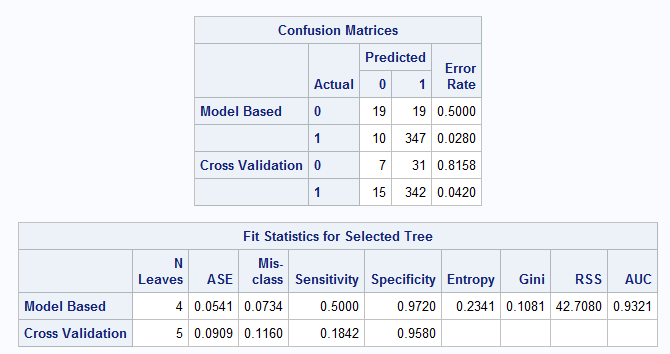
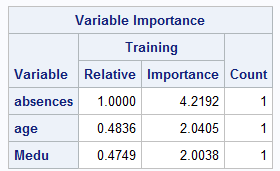
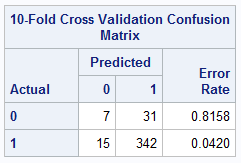


**Four and Five Leaf Classification Trees with Failures variable**

Both the four and five leaf trees below were very similar to the six leaf tree. Slightly worse specificity and PCC than the three leaf tree, and the exact same sensitivity. 

Ultimately, the three leaf tree is clearly the best classifcation tree for this data when predicting whether the student would receive a zero or nonzero final grade in math.

**Classification Trees without Failures Variable**

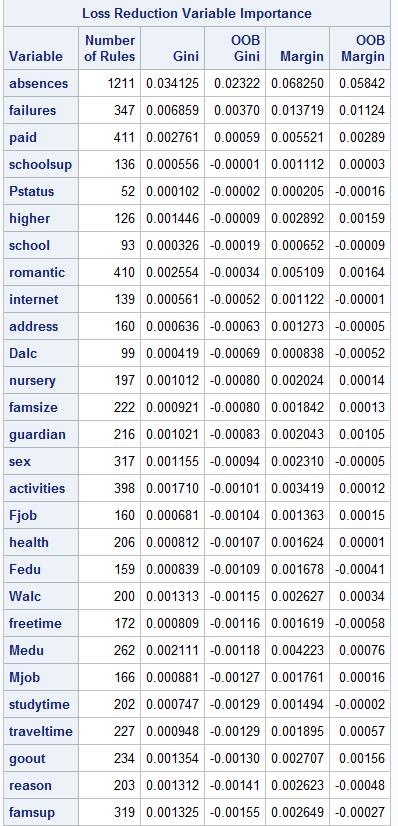
**Three Leaf Classification Tree without Failures variable**  

The cost-complexity graph shows that the best model using the 1-SE rule and ignoring the one leaf tree is the five leaf tree.

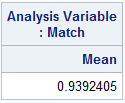
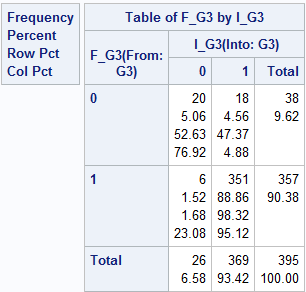
This tree as the exact same first split on absences. Then it splits off students with no absences baed on a student’s mother’s education. Of the students with no absences, those whose mother’s education is none, secondary school, or higher education are predicted to not receive a zero final math grade. Of those students without an absence and who’s mother’s education is between none and secondary school, they are split off by age of the student. If the student’s age of those who do not have absences and whose mother’s education is between none and secondary school is less than 17 years old, then the student is predicted to not get a zero final math grade. If however these same students based on absences and mother’s education have an age 17 or greater, they are predicted to get a zero final math grade. All of these groups correctly predict the remaing students after a split at or above 65%, with only one classified group being less than 77%. This is pretty good.

The 10-fold crossvalidated misclassification rate was 11.6%. Thus, about 88% of the students were correctly classified. The model is very imbalanced though. It correctly classified the students who received nonzero grades almost 96% of the time, but it only correctly classified the actual zero grade students 18% of the time. Thus, the specificity is much higher than the sensitivity. The AUC is very high and good at 0.9321. Absences is the most important variable just like the analysis with the failures variable showed. It’s importance of 4.22 is somewhat important, whereas age and mother’s education are less half as important and thus, not very important.

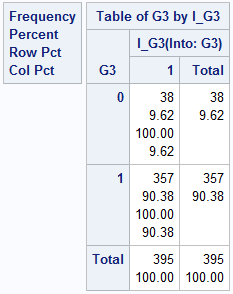
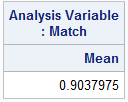
**Random Forests with All Variables**



Now I going to use random forests to see what variables are the most important in classifying whether a student will get a zero or not for a final math grade and to see which parsimonious model is the best in predicting a zero final math grade. The Loss Reduction Variable Importance table shows that absences was the most important predictor variable, failures as the next most important, and paid as the third most important with all variables after paid being not much worse than paid using the out-of-bound Gini statistic. There’s a big drop off in importance from absences to failures, and failures to paid. This suggests that either a model including only absences or absences and failures would be best.

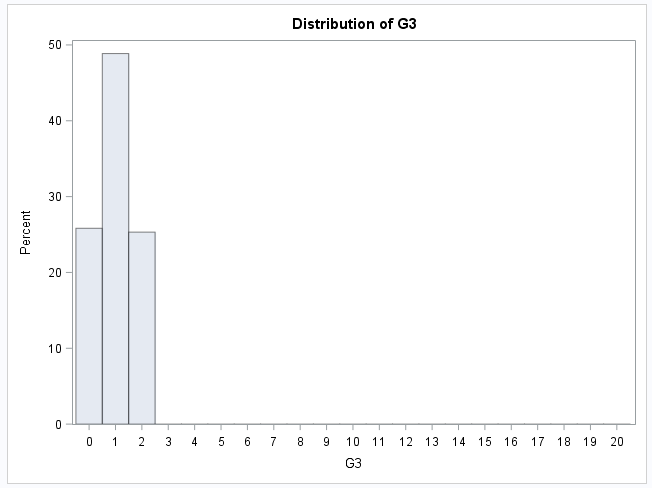


Eliminating every predictor variable other than failures and absenses, the PCC is roughly 94%, which is fantastic. Once again, the sensitivity, was very poor, as only 20 of the 38 zero grades were correctly predicted.

Only using absenses as a predictor variable, the PCC drops to 90%, and the sensitivity nosedives to 0%. This is a signficant drop in PCC for only losing one variable, thus, once again, **the best model includes absenses and failures when prediction zero and nonzero final math grades.** Based on this analysis and the ones done with classificaton trees, **absence is clearly most important variable and only variable not past performance-based that is important in terms of predicting whether a student will receive a zero final math grade.**

**Part II: Group 0 final grades: 0-8, Group 1 final grades: 9-13, Group 2 final grades: 14-20**

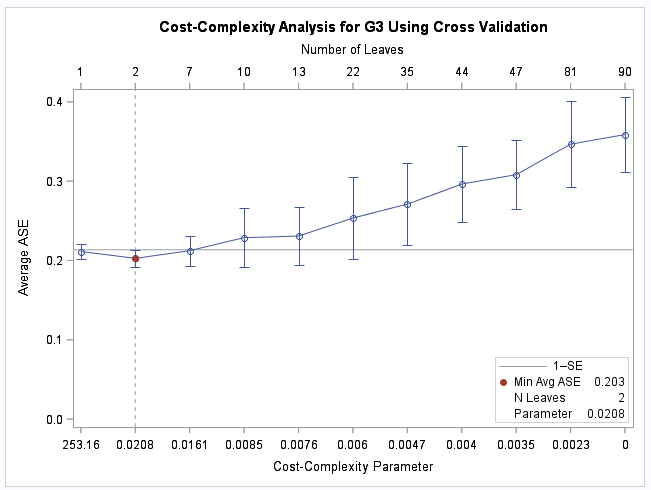


After putting the range of final scores into the three different groups, it’s clear that roughly half of the final grades are in the middle group (“middle50%”), and about 25% in each of the two other groups (“bottom25%” and “top25%”).

In part II will try to answer **“what variables are the most important in classifying students in these math class final grade groups (essentially poor, normal, and great grades in math)?”** and **“what is the best model for correctly classifying these students into these groups?”**

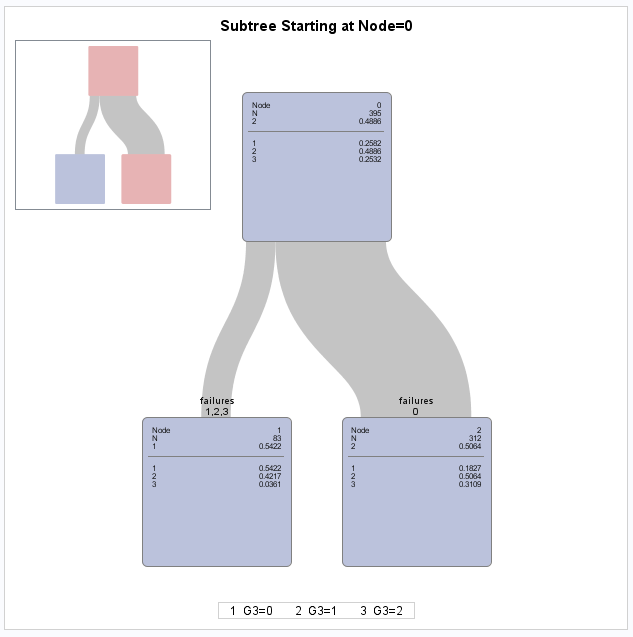
Additionally, the failures variable is the only measure of past class performance. I want to know how well we can predict a student’s math final grade and what variables are important without using past class performance. So I will do an analysis both with and without the failures variable.

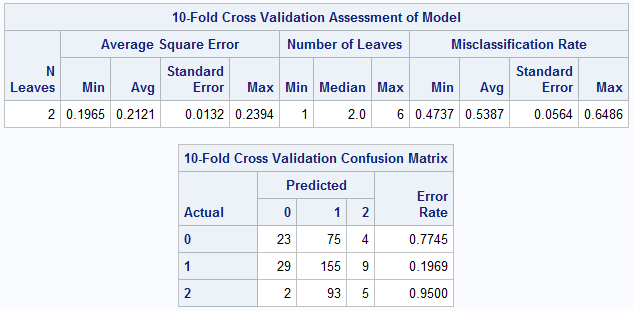
**Classification Trees with Failures Variable**

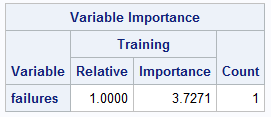


The cost-complexity plot suggests using a tree with two leaves using the 1-SE rule.

**Two Leaf Classification Tree with Failures Variable**







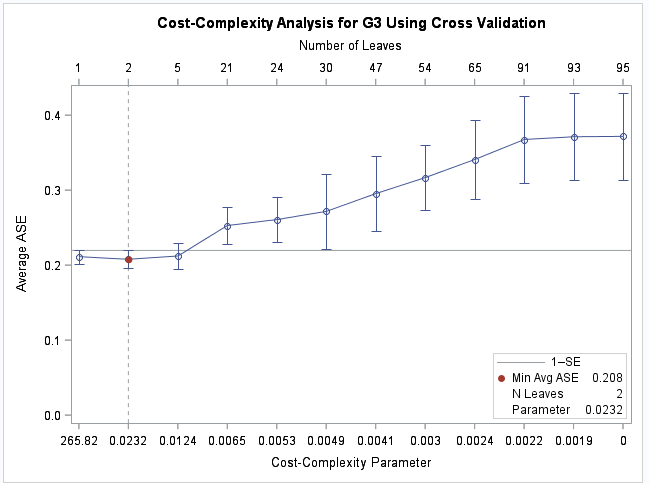
The two leaf tree splits on the number of past class failures. If the student had no past class failures, that student is predicted to be in group 1, or the middle 50% of final grades in math class. Only 50% of the students who have no past class failures were in the middle 50% of math final grades though, which is not very good prediction wise. 31% of these students received a final grade in the top 25%, while only 18% received a final grade in math in the bottom 25%.

If the student has failed a class before, then the student is predicted to be in the bottom 25% of final grades in math class. Only 54% of students who have failed a class received zeros though, so this is not that good of a prediction. 42% of these same students were in the middle 50% of math class final grades, yet only about 4% of students who have failed a class in the past were in the top 25% of final grades in math class.

These two terminal nodes suggest that the strongest conclusion from this model is that the best (top 25% of students) students in math have not failed a class in the past. It’s hard to make any other strong statements. In general, the students with zero past class failures did better than the students with past class failures.

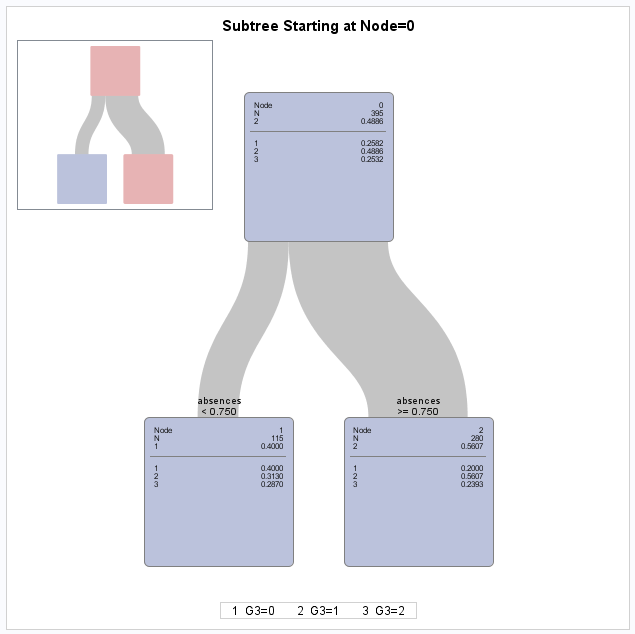
The model is very poor in terms of predictive accuracy. The misclassification rate was a whopping 53.87%. This means the model was more likely to wrongly predict which of the three final grade groups the student was in than to correctly predict it. The 10-fold crossvalidation confusion matrix shows that the vast majority of students who received a grade in the top or bottom 25% were wrongly predicted, while about 80% of the middle 50% group was correctly predicted. The variable importance table shows that failures was the only important variable, as it was the only one used in the model, and an importance of 3.72 is only somewhat important.

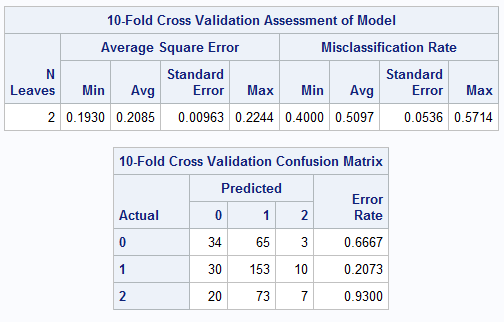
**Classification Trees Without Failures Variable**

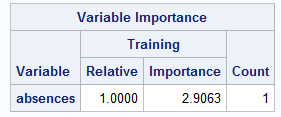


The cost-complexity graph above suggests using a 2 leaf tree using the 1-SE rule if we are excluding 1 leaf trees.

**Two Leaf Tree without Failures Variable**

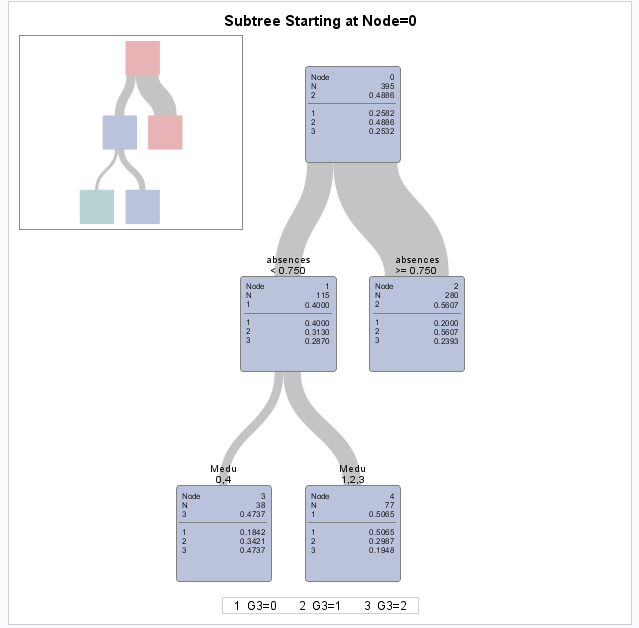


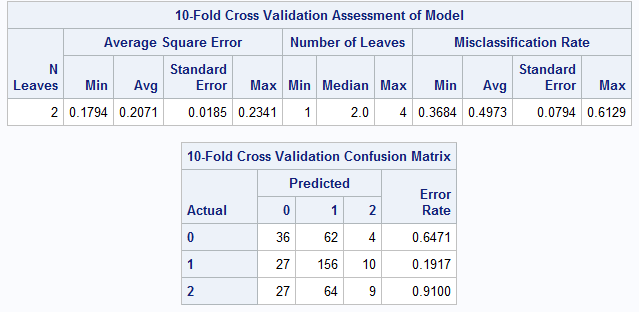


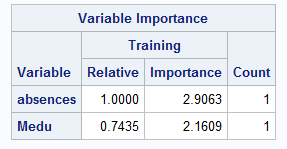


The tree only splits on absences. If the student does not have an absence, the student is predicted to be in the bottom 25% of final math grades. Only 40% of the students without an absence are in the bottom 25% of final grades though, which means this prediction is very weak and inaccurate. If the student has had an absence, then the student is predicted to be in the middle 50% of final math grades. Only 56% of students actually meet both criteria however, so this prediction is weak and not very accurate. The 10-fold crossvalidated misclassification rate was 51%, meaning only 49% of students were correctly classified into the right math final grade group. This is very poor in terms of predictive accuracy. The confusion matrix shows that almost every student in the top 25% was not correctly classified, about two thirds of students in the bottom 25% were also not correctly classified, and about 79% of students in the middle 50% in math final grades were correctly classified. Thus, this model is good for the middle 50% of math final grades, but horrendous for everyone else. The variable importance table shows that absences was somewhat important with an importance of 2.9.

**Three Leaf Tree Without Failures Variable**





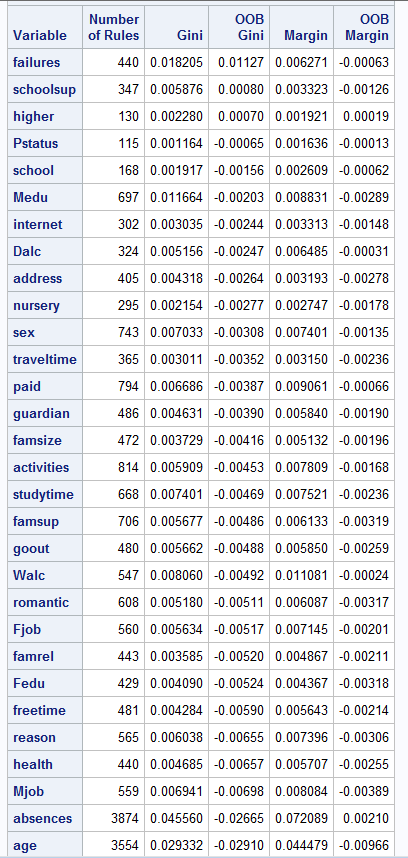


SAS recommended a three leaf model because it had the minimum average square error, so I will compare it with the two leaf model.

The tree has the exact same first split on absences as the two leaf tree. The difference is that it split off the students that did not have an absence. Of those students who did not have an absence, it split on the mother’s education variable. If the mother had no education or higher than a secondary level education, the students with no absences were predicted to be in the top 25% of math final grades. If the mother was in between the two extremes of education level, the students without an absence were predicted to be in the bottom 25%. Both of these two final terminal nodes do not correctly predict more than 50%, so the relationship and predictive accuracy is very weak. It’s hard to draw conclusions on this model not only because of its poor predictive accuracy, but due to one terminal node including the lowest and highest measure of a mother’s education level.

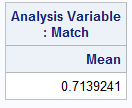
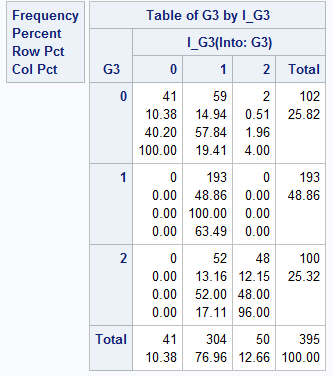
The 10-fold crossvalidated misclassification rate for this three leaf tree is 49%, which is terrible. This means the model is almost as likely to make a wrong prediction than a correct one for which final grade group the students will be in. The confusion matrix shows that 81% of students in the middle 50% of math final grades were correctly predicted, but most of the other students were wrongly predicted into their final grade groups. This means that this model is also only good for predicting students in the middle 50% of math final grades. The variable importance table shows that absences is the most important variable, with mother’s education slightly less important. Both have an importance below 3, so neither is that important in determining which final grade group students will end up in. Despite adding another step in the tree compared to the two leaf model, this tree as almost the exact same PCC, and the mother’s education split is impossible to interpret, so the two-leaf tree is best. **Thus, only using absences to predict a student’s final grade group is the best model.**

**Random Forests with all Variables**

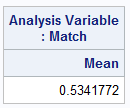
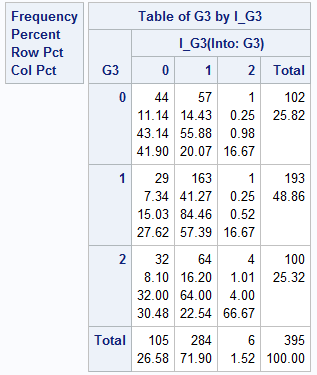


The Loss Reduction Variable Importance table shows that failures is by far the most important variable in predicting the three final grade groups. Mother’s education was not the second best predictor variable, but the sixth best in terms of out-of-bound Gini. There’s a big drop off in importance from absences the next variables in terms of both out-of-bounds Gini and margin. This suggests that clearly absences are likely the only variable worth including in this model. The only other variables to have a positive out-of-bound Gini or margin was schoolsup and higher. To test this, I will compare the confusion matrix of a model with only absences, and a model with failures, schoolsup, and higher.

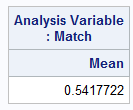
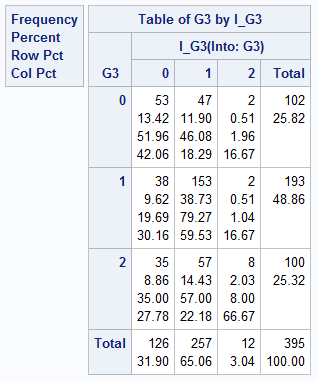
**Full Model**



**Model with failures, schoolsup, and higher variables**



**Model with only failures variable**



The PCC of the full model was decent, at about 72%. The problem with this is that includes 32 predictor variables, so it is not parisimous at all. The PCC of the model with the top three predictor variables was not good, with a PCC of only 53.4%. Incredibly, the model with only absences. This means despite losing two variables, the model improved its accuracy. This shows how entirely insignificant any variable other than absences is in predicting these three final grade groups in math. **Clearly, based on random forests and classification, the model with only absences is the best model for part II.**

**Conclusion**

In conclusion,for part I, **absences and past class failures are the variables important in predicting whether a studnent in these two Portuguese schools will receive a grade of zero in math.** Furthermore, **the best model predicts that students with no absences and at least one past class failure predicts that the student will receive a zero, while all other scenarios predict the student will not receive a zero.** Essentially, **if you have failed a class in the past, students are more likely to get zeros in math class than other students.** The model was very good overall, but needs improvement in predicting zero grades for actual zero grade studennts.

In part II**, failures was the only important variable for predicting a student’s final grade group (bottom 25%, middle 50%, and top 25%).** This is due to random forests and classification trees both showing that the best model only uses failures. This **model for classification trees predicts students with no past class failures are most likely to be in the middle 50% of final math grades, while students with past class failures are most likely to be students who are in the bottom 25% of final math grades.** However, **both the classification and random forests models were very poor overall** at correctly classifying students into one of the three final math grade groups. I think this is due to the fact that failures is the only important variable in predicting final grades, it is hard to classify so many students into more than two groups having only one variable as a predictor. **The one strong conclusion from analysis in part II is that the best (top 25% of students) students in math are extremely unlikely to fail a class in the past.** Eliminating the failures variable resulted in the best model only splitting off absences, but the random forests variable importance showed that no variables other than failures is important in part II. It is unsurprising that failures are so important since it measures past class performance. It is surprising to me though, that none of the other variables were important. I would have thought absences, parent’s education, study time, and the parent’s cohabitation status would be significant factors in predicting a student’s math grade.

The one clear and common interpretation of both parts I and II is that **the more past class failures a student has, the more likely the student is to fail math class.**