Professor Moody

By: Rohit Manjunath

Decision Tree for the Entire Data

- First to benchmark our decision tree let us use Rpart with no control functions and all parameters(columns).

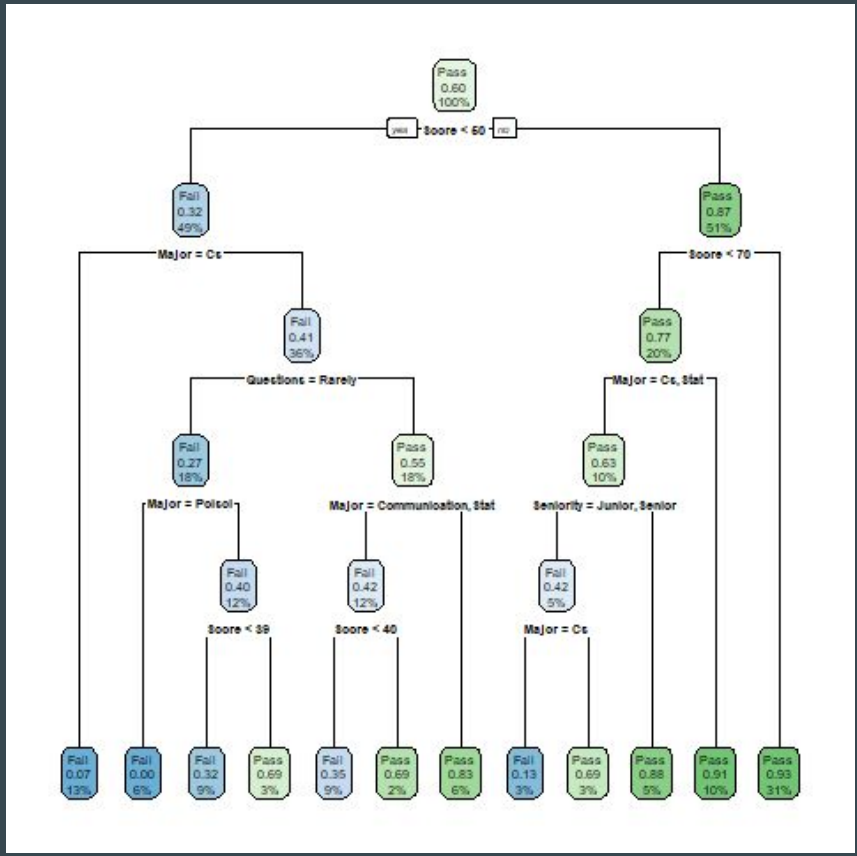
```
tree1 <- rpart(Grade ~ ., data = train)

rpart.plot(tree1)

test$NewGrade <- predict(tree1, newdata = test, type = "class")

error <- mean(test$NewGrade != test$Grade)

error
```

- The ‘error’ output averages around .15 or 15%.

- Since this uses all the columns, cross validation will give us the same accuracy.

Decision Tree for the Entire Data - Output

- As you can see the decision tree in the previous slide completely ignores 3 columns: Student ID, Attendance, and Texting.
- Student ID makes sense as there should not be any correlation with the unique ID and grade, since student ID is just a unique identifier.



Decision Tree for Subsets | Part 1

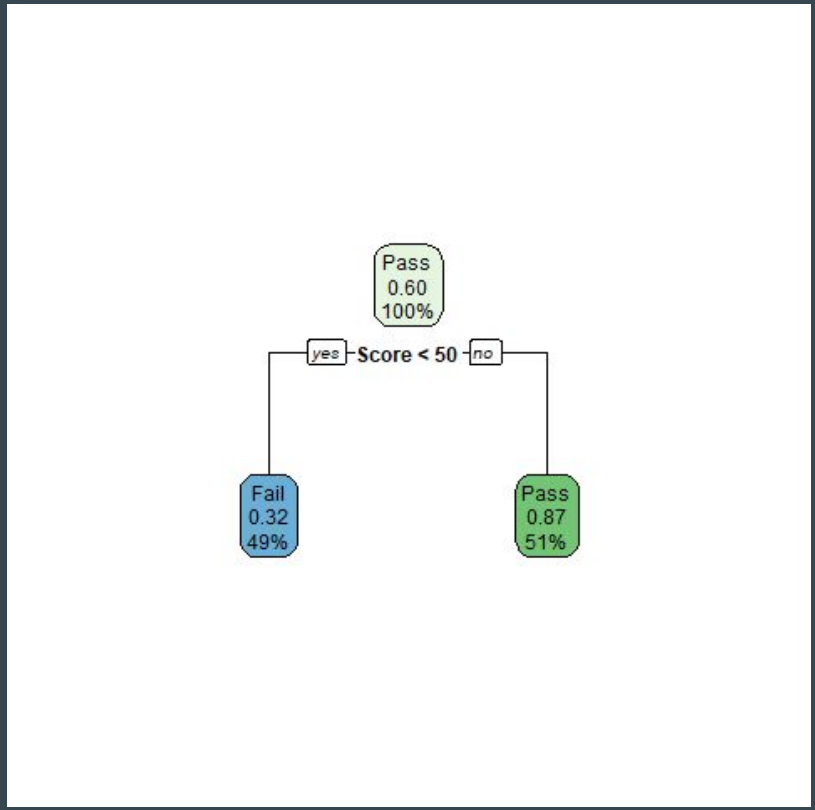
- Let's build a model of the two parameters that was missed by the previous decision tree and score(since it's important) to see if we get a better model.

```
tree2 <- rpart(Grade ~ Attendance + Texting + Score, data = train)

rpart.plot(tree2)

test$NewGrade <- predict(tree2, newdata = test, type = "class")

error <- mean(test$NewGrade != test$Grade)

error
```

- The 'error' output averages around .22 or 22% which is definitely a lot worse than the decision tree with all the columns.

Decision Tree for Subsets | Part 1

- Once again, the decision tree ignored these parameters.
- Now I know for sure there is nothing going on with the Attendance and Texting column.



Cross Validation for Subsets | Part 1

- As you can see this subset is definitely a lot worse than the decision tree with all columns.

- I cross validated it 5 times and it was lesser than ‘all accuracy’ all 5 times.

- This tells me, these parameters definitely do not do the job.

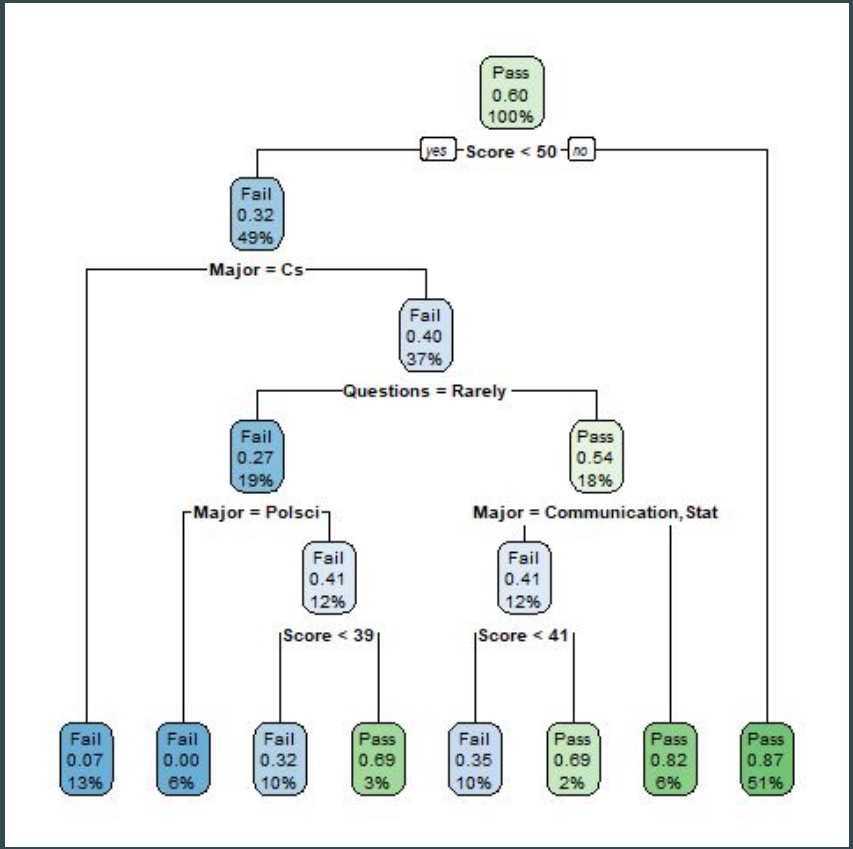
	accuracy_subset	accuracy_all
1	0.7892235	0.8647649
2	0.7654517	0.8314844
3	0.7886952	0.8505018
4	0.7564712	0.8399366
5	0.7723191	0.8277866

Cross Validation for Subsets | Part 2

- Since we have eliminated 3 columns(Student ID, Attendance, Texting) from 7 columns, we only have to worry about Major, Score, Seniority and Questions.
- Since the decision tree that uses all columns only uses these 4, we can try using only 3 columns and eliminate any one column to see if we are overfitting the data.(4C3 = 4)
- After trying all 4 combinations, the decision tree that only takes in Major, Questions, and Score had the lowest error rate of 0.16 or 16%.

```
tree3 <- rpart(Grade ~ Major + Questions + Score, data = train)

rpart.plot(tree3)

test$NewGrade <- predict(tree3, newdata = test, type = "class")

error <- mean(test$NewGrade != test$Grade)
```

Decision Tree for Subsets | Part 2



Cross Validation for Subsets | Part 2

- As you can see this subset is very close to the accuracy of all columns.

- However, it is still lower than the accuracy of all columns.

	accuracy_subset	accuracy_all
1	0.8235605	0.8457475
2	0.8415214	0.8415214
3	0.8225040	0.8452192
4	0.8367670	0.8367670
5	0.8272583	0.8489171

Cross Validation for Subsets with Control | Part 3

- Since we got the subset which is closest to the overall model, let us try to use controls on it and see if we can get it to beat the overall model.

```
tree4 <- rpart(Grade ~ Major + Questions + Score, data = train, control = rpart.control(minsplit = 700, minbucket = 50))

rpart.plot(tree4)

test$NewGrade <- predict(tree4, newdata = test, type = "class")

error <- mean(test$NewGrade != test$Grade)
```

- Using different ratios of minsplit and minbucket(the default is 3), different values of minsplit and minbucket, and by the trial and error method I could get this subset equal or very close to the decision tree with all the columns.
- The ‘error’ output averages around 0.15 or 15%.

Cross Validation for Subsets with Control | Part 3

- As you can see this subset is very close to the accuracy of all columns.

- However, it is still lower than the accuracy of all columns by a slight margin on test numbers 1 and 3.

	accuracy_subset	accuracy_all
1	0.8383518	0.8610671
2	0.8357105	0.8357105
3	0.8383518	0.8515584
4	0.8351823	0.8351823
5	0.8219757	0.8219757

Conclusion

- Through my analysis, I can conclude subsetting and using different parameters only overfits the data and reduces the accuracy of the prediction model.

- Hence, the best model is with no subsets and no controls.

Thank you!