Julia Deaver(jrd7ny)
Caden Moses(cpm7sa)
Max Pilloff(mcp9af)
Jenna Mulvihill(jnm9aba)

STAT 4630 Project: Final Report
Group 27

Executive Summary

Regression Question: Can we predict the win percentage of a team in the 2019 season? If so, what factors can be used to predict this?

Classification Question: Are we able to predict if a college basketball team made it to March Madness in the year 2019?

We are interested in exploring which factors are most important to predict a team win percentage because that information could be implemented in a coaching staff's practice schedule and game strategy. While the ultimate goal of every team in college basketball is to win the March Madness tournament, this is incredibly difficult, so the next best measure of success for a season is a team's win percentage. Understanding which predictors are most influential in impacting a team's win percentage can impact what aspects of the game a coach emphasizes in practice, or even change how they recruit players out of high school. Millions of dollars every year are spent by college basketball programs to gain even the slightest advantage, and this regression model can provide another tool they utilize to gain useful insights for what factors most influence a team's win percentage.

Regarding our classification question, it is worth exploring which factors are the strongest predictors for a team to make the March Madness tournament, because whether a team makes March Madness or not has massive implications for a schools' team, athletic program's budget, and fan morale. Every year, after the regular season, a committee selects 68 teams that they think should be included, basing their decisions on a myriad of factors. For the committee it would be incredibly useful to have a model that they could use to predict whether a team should qualify for March Madness. March Madness generates roughly $900 million in revenue for the NCAA, and some of this money trickles down to individual programs through increased merchandise sales, ticket purchases, and increasing total number of applicants to the school. Due to the financial implications of the tournament it is critical that the selection committee stays consistent in their decision criteria, and ultimately makes the correct call. No team that deserves to be in the

tournament should be left out, it is unfair for the players, coaches and fans and exploring our question can ensure that this does not happen.

Our models from previous milestones are very helpful for answering the question of whether a team will make it to the March Madness tournament or not. We viewed our questions as if we were the selection committee deciding which teams made it into the March Madness tournament. After inputting a team and their statistics into our logistic regression equation, we were able to pretty accurately predict whether a team made it to the postseason in 2019. We found the most important predictors for if a team will make March Madness to be home win percentage (HomeWPct), away win percentage (AwayWPct), and strength of schedule (SOS). The selection committee bases their decision strictly on how a team performs and the competition they play against, not the team's style of play. Overall, our model was successful in answering our question of interest, as we were able to accurately predict whether a team made March Madness or not in 2019 with 94.35% accuracy. The in-game statistics will help us predict the win rate of the team, which is a significant predictor for whether a team will make March Madness or not. When using shrinkage methods, we were able to use almost all predictors which shows us that there is more than one way to win a basketball game and be a successful team with a high win percentage. Of course, the best teams dominate all facets of the game, but a winning team can focus on increasing total rebounds or increasing total assists. However, the predictors that have the largest regression coefficients are turnovers per game and steals per game. So, regardless of height or skill-level, teams that prioritize playing good defense and limiting turnovers tend to be the most successful. The regression tree model helped answer our questions by predicting teams' win percentage given their other percentages. Specifically, it is very helpful in understanding the variables that impact win percentage the most. This model showed us that the most important predictor is clearly turnover rate. Teams with a lower turnover rate tend to have a much higher win percentage. Our classification trees were helpful for answering our classification question of interest because we could see how different variables affected whether or not a team made it to the playoffs. HomeWPct appeared to be the variable which affected this value the most, and strength of schedule was also pretty important. The other variables included in our tree (FGApg, AwayWPct, and FTApg) didn't have as much of an impact on whether or not a team made it but they still did have some effect.

Data Processing and Cleaning

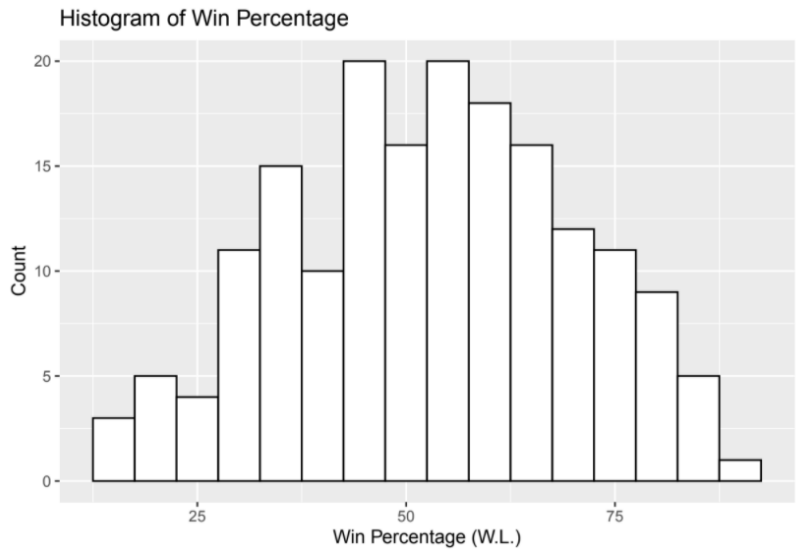
All data was scraped from https://www.sports-reference.com/cbb/. Each Division 1 college basketball team during the 2019 season is an observation in the dataset. The biggest problem we encountered in the data was that all statistics were counted over the season and were not counted as per game metrics. We decided we could not use these as predictors because they would be so strongly correlated with the number of total games a team plays. So, we appended per game statistics for each observation in the dataset by dividing the total number of assists, rebounds, field goals, etc. by the number of games played (G). This was done for the following predictors: FGpg, FGApg, X3Ppg, X3PApg, FTpg, ASTpg, BLKpg, TOVpg, PFpg, TRBpg, ORBpg, STLpg.

Additionally, Before we created our graphical summaries we cleaned our data by removing the first column, and transforming two variables. We divided free throw attempts (FTA) by total number of games (G) in order to get the average number of free throws a team shoots per game (FTApg). We also multiplied the win percentage of each team (W.L.) by 100 in order to make the regression results easier to interpret since it is our response variable.

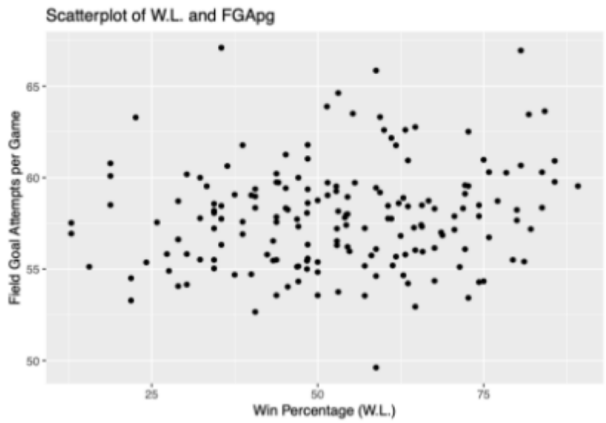
Regression Question

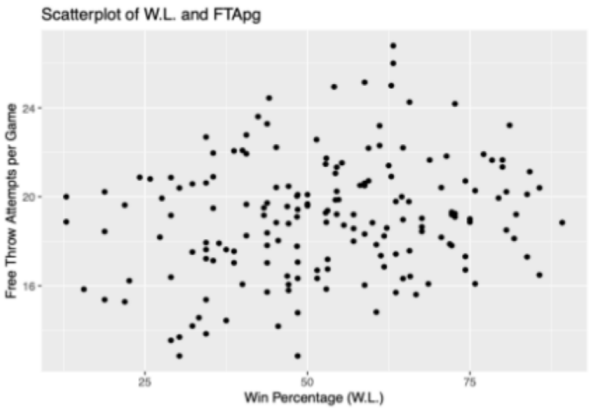
Exploratory Data Analysis

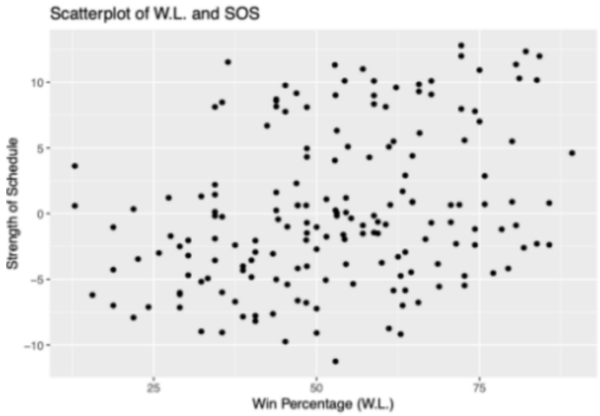
The first plot that we created was a histogram of the variable win percentage (W.L.). We created this plot to see how all of the teams performed in the 2019 season and help us better understand the response variable in our regression analysis. This plot shows us that the distribution of win percentage among all college basketball teams in the 2019 season is nearly normal, with the average win percentage being around 50%. This is interesting because outside of conference play, teams are allowed to schedule their games and we would have expected teams to pick easier competition in order to increase their win percentage. This helps to address our regression question because we are able to get an idea of the distribution of win percentage which is the variable we are focused on in our question.

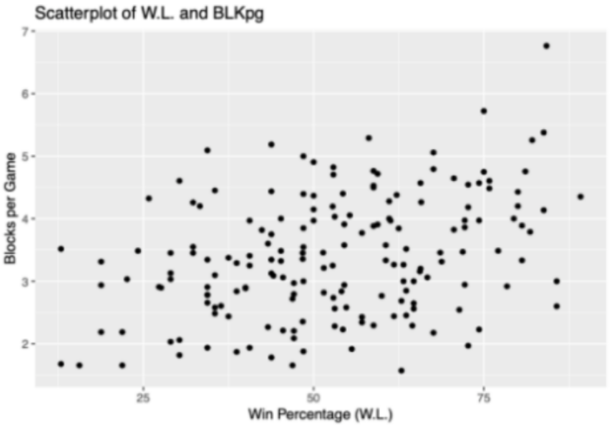


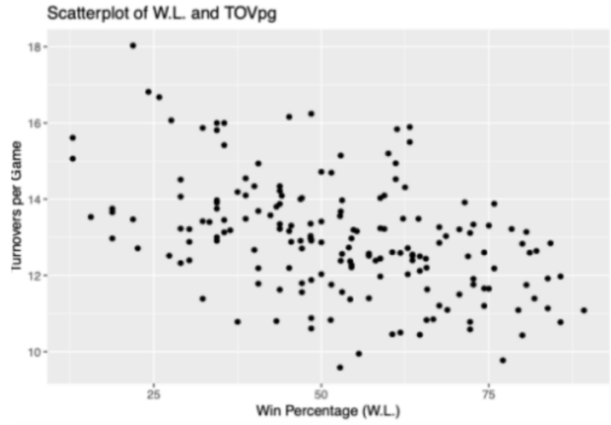
We next created scatter plots analyzing the relationship between the response variable (W.L.) and each of our predictors. We used scatter plots because all of our predictors are numeric variables. It was interesting to see which predictors had a positive relationship with our response variable and which had a negative so that when we got our regression results we could identify predictor values that were different from what we would have expected. These plots help to address our question because we are able to better understand which variables have more of an impact on the win percentage of a team, which we are trying to predict.

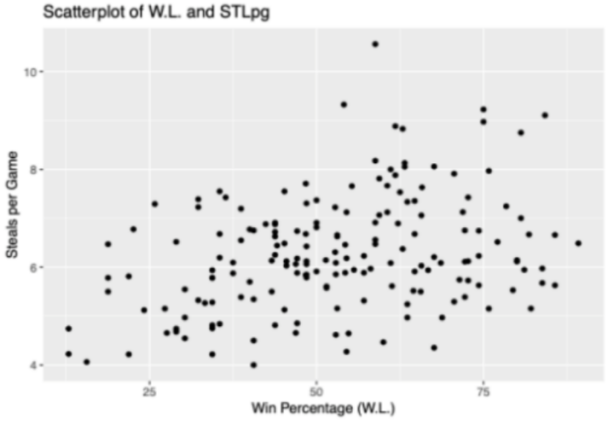

Scatterplot of W.L. and FGApg
Field Goal Attempts per Game
65
60
55
50
25
50
75
Win Percentage (W.L.)

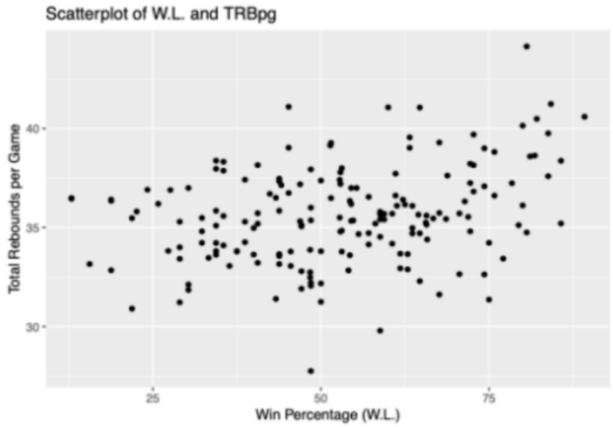

Scatterplot of W.L. and FTApg
Free Throw Attempts per Game
24
20
16
25
50
75
Win Percentage (W.L.)

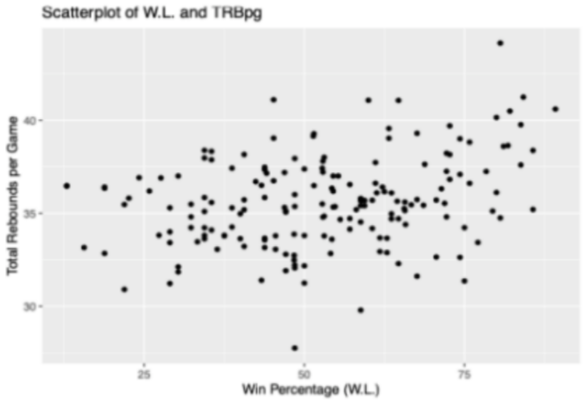

Scatterplot of W.L. and SOS
Strength of Schedule
10
5
0
−5
−10
25
50
75
Win Percentage (W.L.)


Scatterplot of W.L. and BLKpg
Blocks per Game
7
6
5
4
3
2
25
50
75
Win Percentage (W.L.)


Scatterplot of W.L. and TOVpg
Turnovers per Game
18
16
14
12
10
25
50
75
Win Percentage (W.L.)


Scatterplot of W.L. and STLpg
Steals per Game
10
8
6
4
25
50
75
Win Percentage (W.L.)





All regression trees and shrinkage method models performed used the following variables as predictors and a team's win percentage as the response variable:

- Strength of Schedule (SOS)
- FGApg (Field Goal Attempts Per Game)
- X3PA (3 Point Attempts Per Game)
- FTApg (Free Throw Attempts Per Game)
- ASTpg (Assists Per Game)
- BLKpg (Blocks Per Game)
- TOVpg (Turnovers Per Game)
- PFpg (Personal Fouls Per Game)
- STLpg (Steals Per Game)
- TRBpg (Total Rebounds Per Game)

Shrinkage Methods

Ordinary Least Squares Regression

The following is the output after performing an ordinary least squares regression:

```
Call:
lm(formula = W.L. ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-23.749  -6.035  -0.113   5.645  24.799

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.8106    16.5250   4.285 3.10e-05 ***
SOS          -0.1470     0.1354  -1.086  0.27911
FGApg        -3.5739     0.3900  -9.163  < 2e-16 ***
X3PApg        0.9342     0.2396   3.900  0.00014 ***
FTApg         0.3956     0.3295   1.201  0.23164
ASTpg         3.0381     0.4436   6.849 1.40e-10 ***
BLKpg         1.8635     0.8689   2.145  0.03345 *
TOVpg        -7.1680     0.5693 -12.591  < 2e-16 ***
PFpg          0.4303     0.5793   0.743  0.45866
STLpg         6.1377     0.7844   7.825 5.79e-13 ***
TRBpg         4.4949     0.4535   9.911  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.147 on 165 degrees of freedom
Multiple R-squared:  0.734,     Adjusted R-squared:  0.7179
F-statistic: 45.54 on 10 and 165 DF,  p-value: < 2.2e-16
```

In the OLS Regression, Turnovers per Game and Steals per Game are statistically significant and have the largest regression coefficients. This means that the more turnovers a team commits per game and the fewer steals they have, the lower their predicted win percentage will be, both predictable outcomes. An interesting result from this regression is that teams with fewer field goal attempts per game, which are commonly ones that play at a slower tempo, have a higher predicted win percentage. This result could mean that teams that take fewer shots often take higher percentage shots and are more likely to win Games.

Ridge Regression

A ridge regression was run using the same predictors and response variable. First, using cross-validation of the training data, the optimal tuning parameter was found. This value was 0.79486. When $\lambda = 0$, the ridge regression acts as a linear regression. So, because this value is so close to 0, the ridge regression acts relatively similar to the OLS Regression. The output from the ridge regression is below:

```
11 x 1 sparse Matrix of class "dgCMatrix"
                     s0
(Intercept) 57.61226322
SOS         -0.05177653
FGApg       -2.79084801
X3PApg       0.78741083
FTApg        0.58113374
ASTpg        2.89385200
BLKpg        2.13808248
TOVpg       -6.36819261
PFpg         0.14255074
STLpg        5.16608915
TRBpg        3.63552786
```

Unsurprisingly, the ridge regression results are similar to the OLS regression. Turnovers per game and steals per game have the largest coefficients. Total rebounds also have a large positive coefficient, meaning that a team with more rebounds per game has a higher predicted win percentage. This result is expected.

Lasso Regression

Finally, a Lasso regression was performed. Using cross-validation of the training data, the optimal tuning parameter was found. This value was 0.06299, so the ridge regression acts very similar to a linear regression

```
11 x 1 sparse Matrix of class "dgCMatrix"
                     s0
(Intercept) 68.9366269
SOS         -0.1151760
FGApg       -3.4253261
X3PApg       0.8913489
FTApg        0.4028151
ASTpg        2.9897492
BLKpg        1.8022060
TOVpg       -7.0043144
PFpg         0.3323834
STLpg        6.0071523
TRBpg        4.3649869
```

The results are very similar to the Ridge and OLS Regression. Most significantly, the Lasso regression did not perform any variable selection, and all 10 predictors were included in the model.

Regression Type	Mean Squared Error
Ordinary Least Squares Regression	94.47264
Ridge Regression	95.40818
Lasso Regression	94.07911

The table above shows the test MSE by regression type. All three regressions had very similar test MSEs, which is understandable given that the tuning parameters were very close to 0. The Lasso regression had the lowest MSE, but all three models were very close in MSE. Thus, the Ridge and Lasso regressions were very similar to the OLS regression. This most likely occurred because there was a lack of multicollinearity among the predictors, which causes higher variance and leads to the Ridge and Lasso (which reduce variance and increase bias) being more effective in reducing test MSE.

The significance of almost all predictors shows us that there is more than one way to win a basketball game and be a successful team with a high win percentage. Of course, the best teams dominate all facets of the game, but a winning team can focus on increasing total rebounds or increasing total assists. However, the predictors that have the largest regression coefficients are turnovers per game and steals per game. So, regardless of height or skill-level, teams that prioritize playing good defense and limiting turnovers tend to be the most successful.

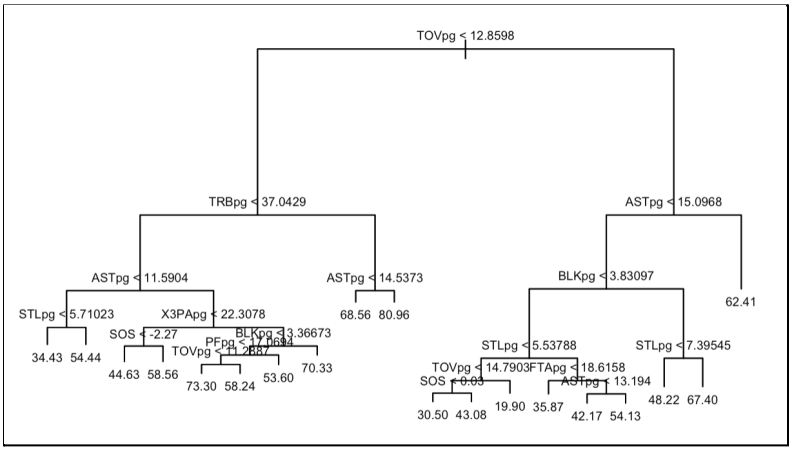
Strength of Schedule was one of the most insignificant predictors in the model. This was a confusing result at first, but makes sense; teams that have a higher strength of schedule might on average have a lower win percentage since their opponents are stronger, which is why in all three models this coefficient is negative. However, teams that are stronger and tend to win more often seek out a difficult schedule in order to boost their chances of making the NCAA Tournament, something we will show in the next section.

Regression Trees

We are choosing to present the tree from recursive binary splitting, because our pruned tree remained the same as our tree from recursive binary splitting.

```
Regression tree:
tree::tree(formula = W.L. ~ ., data = train)
Variables actually used in tree construction:
[1] "TOVpg"  "TRBpg"  "ASTpg"  "STLpg"  "X3PApg" "SOS"    "BLKpg"  "PFpg"   "FTApg"
Number of terminal nodes:  19
Residual mean deviance:  82.83 = 13000 / 157
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-22.4200  -5.6720   0.2154   0.0000   5.8430  23.9800
```

Our tree has 19 terminal nodes:



Our regression tree model tells us that the most important predictor is turnovers per game. Next we can see that total rebounds per game and assists per game are also important. From our model, it looks like teams with the lowest average turnovers per game, higher assists per game and higher average blocks per game have the highest win percentage.

The importance within random forests shows us that turnovers and assists are the most important variable in predicting win loss percentage. It also shows us that field goals and three point attempts are the least important in predicting a team's win loss percentage:

	%IncMSE	IncNodePurity
SOS	4.8225616	4136.736
FGApg	-0.9145192	2540.136
X3PApg	0.3221451	2737.900
TRBpg	12.4414147	5022.950
STLpg	12.2851337	5948.240
FTApg	7.1554490	3767.044
ASTpg	17.2949077	7659.416
BLKpg	8.5119472	5128.904
TOVpg	23.3693262	8755.834
PFpg	4.2335225	3149.396

Below are the test MSEs for each regression tree:

Recursive Binary Splitting <dbl>	Pruned Tree <dbl>	Random Forest <dbl>
230.7726	230.7726	161.6889

Our test MSEs for regression trees are pretty high. This implies that a regression tree is not the best way to model our data, likely because the data is more linear than box shaped. Despite this, the model helps answer our questions by predicting teams' win percentage given their other percentages. Specifically, it is very helpful in understanding the variables that impact win percentage the most. This model shows us that the most important predictor is clearly turnover rate. Teams with a lower turnover rate tend to have a much higher win percentage. In general, this is the most valuable information we received from the regression tree models, as the model in itself is not a great predictor of a team's win percentage.

Summary of Findings

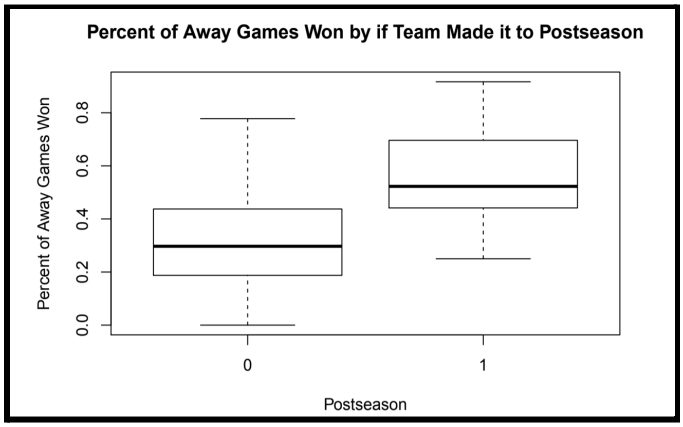
Among all of our models, Lasso regression and ordinary least squares performed the best. Lasso regression has the smallest test MSE of 94.08 and ordinary least squares has a test MSE of 94.47. Next, the ridge regression has a test MSE of 95.41, which is higher than the other linear models but much less than all of the tree models. All of the tree models had very high test MSEs. Random forest has the lowest test MSE of the tree models, at 161.69. Next, recursive binary splitting and pruning have the same test MSE of 230.77.

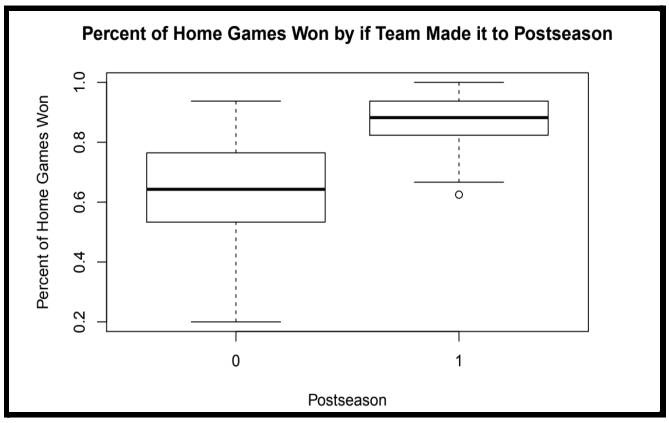
Specifically, we found that the linear regression methods performed much better than the tree methods. This is likely because our data is more linear than box shaped.

Overall, the best model was the lasso regression. It had the lowest test MSE, seeming to perform better than all the other models.

Classification Question

Exploratory Data Analysis





We created boxplots to analyze the percentage of both away and home games respectively won, separated by if a team made the tournament or not. For both predictors, teams that made the tournament clearly had higher winning percentages.

All models used the following variables as predictors and whether a team made the tournament or not as the response variable:

- Strength of Schedule (SOS)
- AwayWPct (Away Win Percentage)
- HomeWPct (Home Win Percentage)
- FGApg (Field Goal Attempts Per Game)
- X3PA (3 Point Attempts Per Game)
- FTApg (Free Throw Attempts Per Game)
- ASTpg (Assists Per Game)
- BLKpg (Blocks Per Game)
- TOVpg (Turnovers Per Game)
- STLpg (Steals Per Game)

Logistic Regression Model

First, a logistic regression model was created using all ten predictor variables above. The primary goal of the logistic regression model was inference; to gather insight into how different variables affect a team's chances of being selected to play in the NCAA Tournament.

```
Call:
glm(formula = tourneydummy ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6897  -0.2466  -0.0528  -0.0020   3.4124

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.195769   7.249934   0.027 0.978458
SOS          0.272986   0.077138   3.539 0.000402 ***
HomeWPct    16.500093   4.515637   3.654 0.000258 ***
AwayWPct     4.364044   2.069975   2.108 0.035009 *
FGApg       -0.290000   0.149333  -1.942 0.052141 .
X3PApg       0.066666   0.127519   0.523 0.601117
STLpg        0.346843   0.317240   1.093 0.274255
FTApg        0.019474   0.140669   0.138 0.889893
ASTpg       -0.008145   0.194226  -0.042 0.966549
BLKpg        0.107796   0.377582   0.285 0.775268
TOVpg       -0.356541   0.268499  -1.328 0.184210
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 188.658  on 175  degrees of freedom
Residual deviance:  72.729  on 165  degrees of freedom
AIC: 94.729

Number of Fisher Scoring iterations: 8
```

The following output was obtained from the full logistic regression model. Only 3 predictors (SOS, HomeWPct, and AwayWPct) were statistically significant at the 5% significance level. The coefficients of these variables were also an expected result. Teams with higher strength of schedules, away win percentages, and home win percentages were more likely to make the NCAA Tournament.

The full logistic regression model performed well on the test data. The Area under the ROC Curve was .9579. Additionally, with a threshold of 0.5, the test error rate was 0.0678. The biggest concern was the 32% false negative rate. Below is the confusion matrix for the test data:

```
   FALSE TRUE
0    146    3
1      9   19
```

Since only three predictors were significant in the full model, we created a reduced logistic regression model, using only SOS, HomeWPct, and AwayWPct as predictors. This model yielded the following output:

```
Call:
glm(formula = tourneydummy ~ HomeWPct + AwayWPct + SOS, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5911  -0.3406  -0.0646  -0.0044   3.2393

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.79916    3.14003  -5.032 4.87e-07 ***
HomeWPct     15.16621    3.86212   3.927 8.60e-05 ***
AwayWPct      4.93240    1.95497   2.523   0.0116 *
SOS           0.29481    0.06885   4.282 1.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 188.658  on 175  degrees of freedom
Residual deviance:  79.496  on 172  degrees of freedom
AIC: 87.496

Number of Fisher Scoring iterations: 7
```

First, a likelihood ratio test was conducted to compare the reduced and full models. In this hypothesis test, the null hypothesis states that there is no difference between the two models in predicting the response variable. The alternative hypothesis is that the full model is statistically significantly more accurate. The full model has a residual deviance of 72.729 and the reduced model had residual deviance of 79.496. Since 7 predictors were dropped, the p-value of this test is 1-pchisq((79.496 - 72.729), 7):

```
[1] "The p-value of the Likelihood Ratio Test is: 0.45353636345008"
```

Thus, because the p-value is greater than 0.05, we fail to reject the null hypothesis. So, there is not a statistically significant difference between the full and reduced models.

The results of the reduced model are very similar to the full model: HomeWPct, AwayWPct, and SOS are all statistically significant predictors and all three have positive coefficients in the logistic regression. This is an expected result, because as a team plays a more difficult schedule and wins a higher percentage of games, one would expect them to be more likely to make the NCAA Tournament.

The reduced logistic regression model also performed well on the test data. The Area under the ROC Curve was .9579. Additionally, with a threshold of 0.5, the test error rate was .0565. Below is the confusion matrix with a threshold of 0.5:

	FALSE	TRUE
0	147	2
1	8	20

The false negative rate was calculated to be 28.6% and the false positive rate was 1.34%. Since the data is unbalanced and many more teams in Division 1 College Basketball miss the tournament than those which make it, we decided to adjust the threshold to lower the false negative rate and more accurately predict teams that make the NCAA tournament. Lowering the threshold means that the model is more likely to predict a team to make the NCAA Tournament, lowering the false negative rate and increasing the false positive rate. Below is the confusion matrix when using a threshold of 0.2:

	FALSE	TRUE
0	128	21
1	3	25

Although the error rate increased from .0565 to .136 when increasing the threshold, the false negative rate decreased dramatically. The new false negative rate is 10.7% and the new false positive rate is 14.09%.

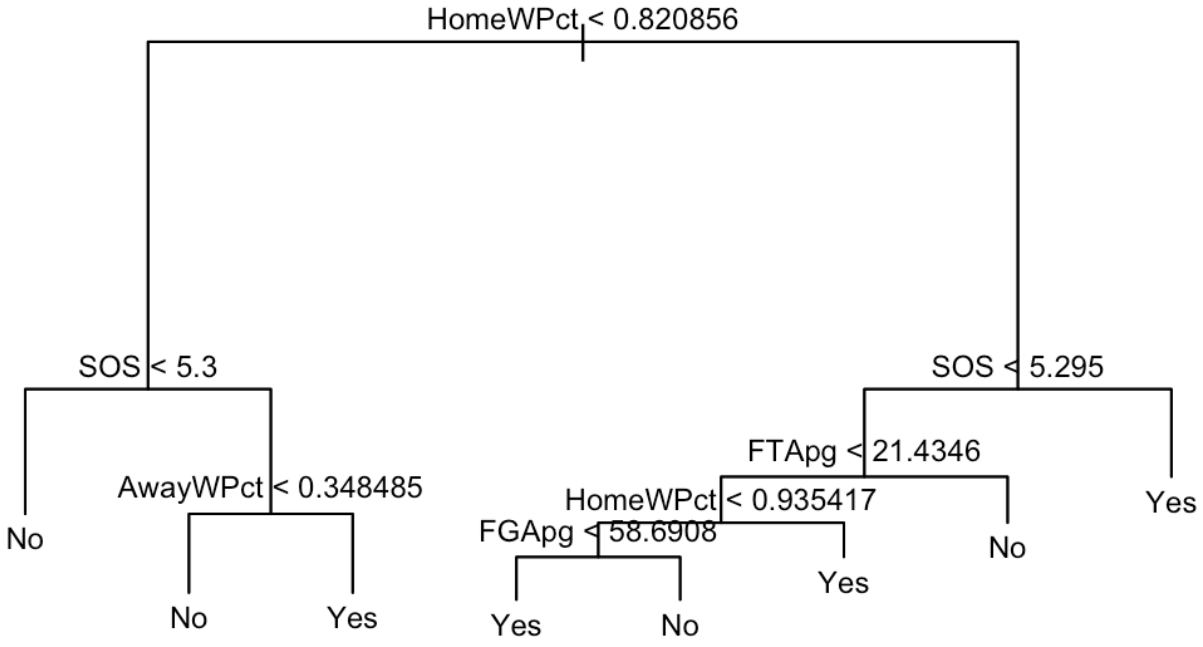
After running our model, taking out variables that had high multicollinearity concerns, we found the most important predictors for if a team will make March Madness to be home win percentage (HomeWPct), away win percentage (AwayWPct), and strength of schedule (SOS). While we expected these predictors to be significant from the beginning of our model building process, we were interested to see that none of the in-game predictors were significant. Initially, we felt that this was cause for concern; however, upon further discussion we concluded that this makes sense for answering our question of interest. The selection committee bases their decision strictly on how a team performs and the competition they play against, not the team's style of play. Overall, our model was successful in answering our question of interest, as we were able to accurately predict whether a team made March Madness or not in 2019.

Classification Trees

For our classification tree, we will be presenting our pruned tree. This is because, although the confusion matrix for both the pruned and recursive binary splitting tree was the same at a threshold at 0.5, when we adjusted our threshold, the pruned tree performed much better. There are also fewer terminal nodes on our pruned tree than our recursive binary splitting tree, making it easier to interpret. This makes answering our question more simple than with the other tree.

```
Classification tree:
snip.tree(tree = tree.class.train, nodes = c(4L, 10L, 48L))
Variables actually used in tree construction:
[1] "HomeWPct" "SOS"      "AwayWPct" "FTApg"    "FGApg"
Number of terminal nodes:  8
Residual mean deviance:  0.2995 = 50.32 / 168
Misclassification error rate: 0.04545 = 8 / 176
```

Our tree has 8 terminal nodes:



This tree is helpful for answering our question of interest because we can see how different variables affect whether or not a team made it to the playoffs. HomeWPct appears to be the variable which affects this value the most, and strength of schedule is also pretty important. The other variables included in our tree (FGApg, AwayWPct, and FTApg) don't have as much of an impact on whether or not a team made it but they still do have some effect.

This output helps to answer our question of interest because we are able to see which variables are most important to answering our question. HomeWPct and SOS appear to be the most important predictors by random forest. AwayWPct also appears to be fairly important.

	No	Yes	MeanDecreaseAccuracy	MeanDecreaseGini
SOS	12.84341038	16.51468038	18.670110392	10.518189
HomeWPct	18.71547836	22.31821673	26.070828856	15.040423
AwayWPct	7.81250470	15.22280766	15.580177123	8.542258
FGApg	0.61460434	-2.20617028	-0.585567464	2.701922
X3PApg	-1.52668941	-0.04774869	-1.433500880	1.877197
FTApg	0.60272548	-0.57778339	0.005363427	2.018385
ASTpg	3.57444977	3.04638713	4.703802709	2.727246
BLKpg	0.57519797	4.20191735	3.190798362	4.385754
TOVpg	3.52443035	11.11069746	10.060255370	5.770409
PFpg	-0.06770426	-3.18328522	-1.992070133	2.192463
STLpg	2.46992103	2.84413001	3.378507291	3.275844
TRBpg	0.48718293	1.11092688	0.924717157	2.763047

Pruning Test Error Rate	Random Forests Test Error Rate
0.1073446	0.08474576

The false positive rate for pruning is 0.08053691. The false negative rate is 0.25. The false positive rate for random forests is 0.02013423. The false negative rate is 0.4285714.

Although the test error rates are pretty low, the threshold needs to be adjusted. The threshold should be decreased because the false negative rates are both very high. This means that the tree is often saying that a team didn't make it to the playoffs when they actually did. We can't have this because not many teams make it so we want to predict those who did as best as possible.

Pruned Confusion Matrix with Threshold of 0.1	Random Forests Confusion Matrix with Threshold of 0.1
pred.test FALSE TRUE No 130 19 Yes 3 25	pred.test FALSE TRUE No 93 56 Yes 2 26

Tree-based methods are helpful for answering our group's question of interest because they are easily interpretable and we are able to see the impact that each variable has on our response. It was also important because we were able to see that a lot of the variables that we once thought would be important for our question actually were not because about half of the variables weren't in a single tree. We were able to see that SOS and HomeWPct were as important as we had thought and that they would be able to really have an effect on whether a team made it to the playoffs or not. If these were the only values we had, we could still get a pretty good idea of how a team did.

Summary of Findings

The test error rate for our reduced logistic regression model was 0.0565. The test error rate for our pruned classification tree was 0.1073. The test error rate for random forest was 0.0847. This means that our logistic regression model had the lowest test error rate.

Before adjusting the threshold, the false positive rate for our reduced logistic regression model was 0.0134 and the false negative rate was 0.286. The false positive rate for pruning is 0.0805. The false negative rate is 0.25. The false positive rate for random forests is 0.0201. The false

negative rate is 0.4286. The reduced logistic regression model had the lowest false positive rate, however, the pruned tree had the lowest false negative rate.

In both our reduced logistic regression model and our classification trees we found that decreasing our threshold would be important. This is because with the threshold of 0.5, the false negative rate for all of our models was very high. This means that the models were often predicting that a team did not make it to the March Madness tournament when they really did. With few teams making it to the tournament, it is really important that our model predicts more of those teams to have made it. With our reduced logistic regression model, when we changed the threshold to 0.2, although the error rate increased from .0565 to .136, the false negative rate decreased dramatically. The new false negative rate was 0.107 and the new false positive rate was 0.1409. When we changed the threshold for our pruned tree and random forests to 0.1, we saw for our pruned tree that the test error rate changed to 0.1243, the false positive rate changed to 0.1275, and the false negative rate changed to 0.1071. For the random forests, we saw the test error rate increase to 0.3277 and the false positive rate increase to 0.3758, but the false negative rate decreased to 0.0714.

The findings in subsections 4.1 to 4.3 were fairly similar. In each of our models in sections 4.2 and 4.3, we found that the test error rate appeared to be low, but when looking more closely at our false positive rate and false negative rate, the models weren't great. This was because the false negative rates were so high. In each of the subsections, we found that decreasing the threshold helped the models significantly with the issue of the high false negative rate. In our EDA and the models in sections 4.2 and 4.3, we found that Home Win Percentage and Strength of Schedule were very important variables for predicting if a team made it to the March Madness tournament.

Our pruned tree with a threshold of 0.1 appeared to be our best model to answer our question. This model had the lowest test error rate of any of the models after the change of threshold, but it also kept the false positive and false negative rates pretty low. Although the false positive rate was a bit higher than that of the reduced logistic regression model, this wasn't as important because the false negative rate is the rate we should care about the most when not many teams make it to March Madness and we want to predict those who do well. The random forest model with a new threshold had a huge false positive rate which is also bad for our model, so it is best to stick with the model that had a fairly low test error rate, false positive rate, and false negative rate, like our pruned tree. With this model being the best to answer our question, we are able to easily interpret it by looking at the small tree and figuring out where different teams would fall. This way, the predictions of whether a team makes it to the tournament will be most accurate.

Further Work

With more time on this project, we would have investigated a few different things. First, we would have subsetted the data by type of college. The easiest way to do this would be to separate Power 6 Schools (those in the ACC, Big Ten, Big East, SEC, Pac 12, and Big 12) from all other colleges in the country. This would be especially useful for the classification question. Power 6 conferences usually send at least 4-5 teams to the NCAA tournament, while smaller conferences only send the conference tournament winner. It's possible that strength of schedule is less important for small schools, since the only way for them to qualify for the NCAA Tournament is to win their conference tournament.

Additionally, we could also only observe "at-large bids" to the NCAA Tournament, and remove all teams that win their conference tournaments and automatically qualify for March Madness.

Investigating if the same results hold in another era of college basketball or in Women's College Basketball would be another interesting future project. We could see in the 1970s, before the 3-point line was added, if there was a much different way to win basketball games. Additionally, we could look at Women's College Basketball and determine if certain statistics led to a more successful team compared to Men's College Basketball.