Multivariate Modeling and Predictive Inference of Win Outcomes in NCAA Division I Men's Basketball: A Statistical Analysis of Gameplay and Recruiting Metrics

Alexandra Lostetter, Ryan Brown, Ava Konopka, Amy Thalangsy, Heywood Williams

Table of Contents

1. Introduction

Attending arguably one of the most historically notable basketball schools has inspired Group 5 to investigate college basketball statistics during the period from 2013 to 2021. Coaches, players, and fans are primarily concerned with the percentage of games their favorite team wins and their progression into the postseason tournament. Higher win percentages increases the likelihood of a teams' entry into the NCAA tournament and their shot at claiming the title of National Champion. To assess the chances of these outcomes in the future, Group 5 has devised 2 questions after having completed an initial investigation of our data. One of our main goals for these questions is to determine which combinations of factors in our data best predict win percentage during the season and performance in the March Madness tournament. Another goal was to introduce new variables to potentially create more effective models which better predict win percentage. In other words, to make our insights more innovative, we incorporated a new dataset with recruiting statistics to analyze several more independent variables that may be affecting our dependent variable.

Our first proposed question states "What combination of variables best predicts win percentage?" As mentioned prior, Division 1 basketball team coaches, players and fans are eager to know how likely their team is to win a certain percentage of games as well as how likely they are to advance to later stages of the postseason tournament. Discovering which combination of variables best predicts win percentage would assist coaches in determining how to optimize certain skillsets of the team to achieve success. For instance, it may be discovered that shooting two pointers is a more effective strategy at winning games than risking lucky three-point shots.

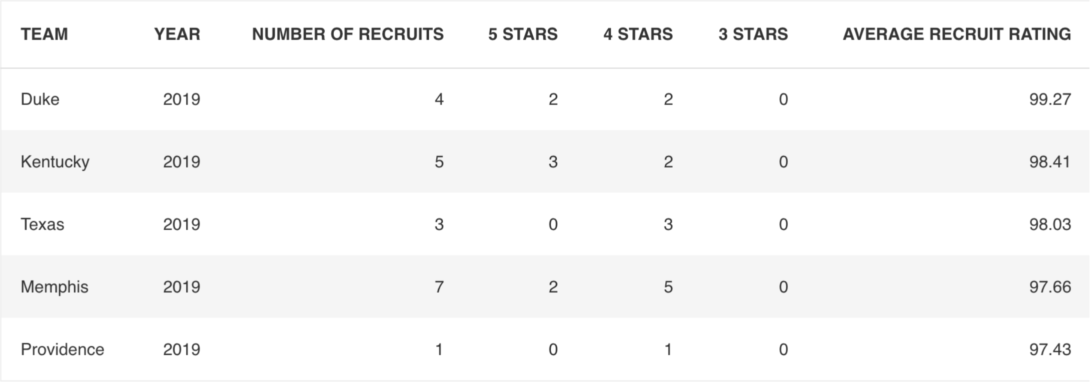
Our second proposed question states "Can you use recruiting data to accurately predict win percentage?" While completing our initial data analysis, we noted how our original dataset disregards some variables that may tell a more complete story of how a team reaches success in a given season. To take into consideration some of these additional variables, we decided to introduce a new dataset that lists college basketball recruits for each team across all conferences. This would be interesting to the owner of the original data as it introduces outside variables that, as mentioned prior, may greatly contribute to a team's win percentage. The answer to this question may also give the basketball program's recruiting team or coach greater insight for the future regarding what details about a player to focus on when recruiting, especially with limited athletic scholarships available per program. In other words, the roster of a college basketball team changes frequently due to players graduating, getting injured, entering the transfer portal, walking on, etc. Since recruiting is constantly occurring, will focusing on recruiting data or number of those recruited yearly affect win percentage? Our paper strives to answer this very question.

2. Data Description

Debatably one of the most well-known companies in the sports industry, Bart Torvik, has provided Group 5 with a plethora of sports data. In a blog post, Torvik states that his data originates from a variety of sources including a subscription with natstat.com and stats.ncaa.org. His data was then compiled by Andrew Sunberg, a student at Iowa State University, who later presented it on Kaggle. Group 5 combined all the individual .csv files in order to conduct our own data analysis. The final .csv file we brought into R Studio contains 3,155 rows and 25 columns. To answer our first proposed question using our original dataset, we examined many different qualitative and quantitative variables. Group 5's qualitative variables include the `TEAM` variable referring to 355 Division 1 college basketball schools. Another qualitative variable is called `CONF`, noting which of the athletic conferences each team belongs to. The last qualitative variable we used in our analysis is `POSTSEASON` which indicates the specific round a given team was eliminated from in the postseason tournament.

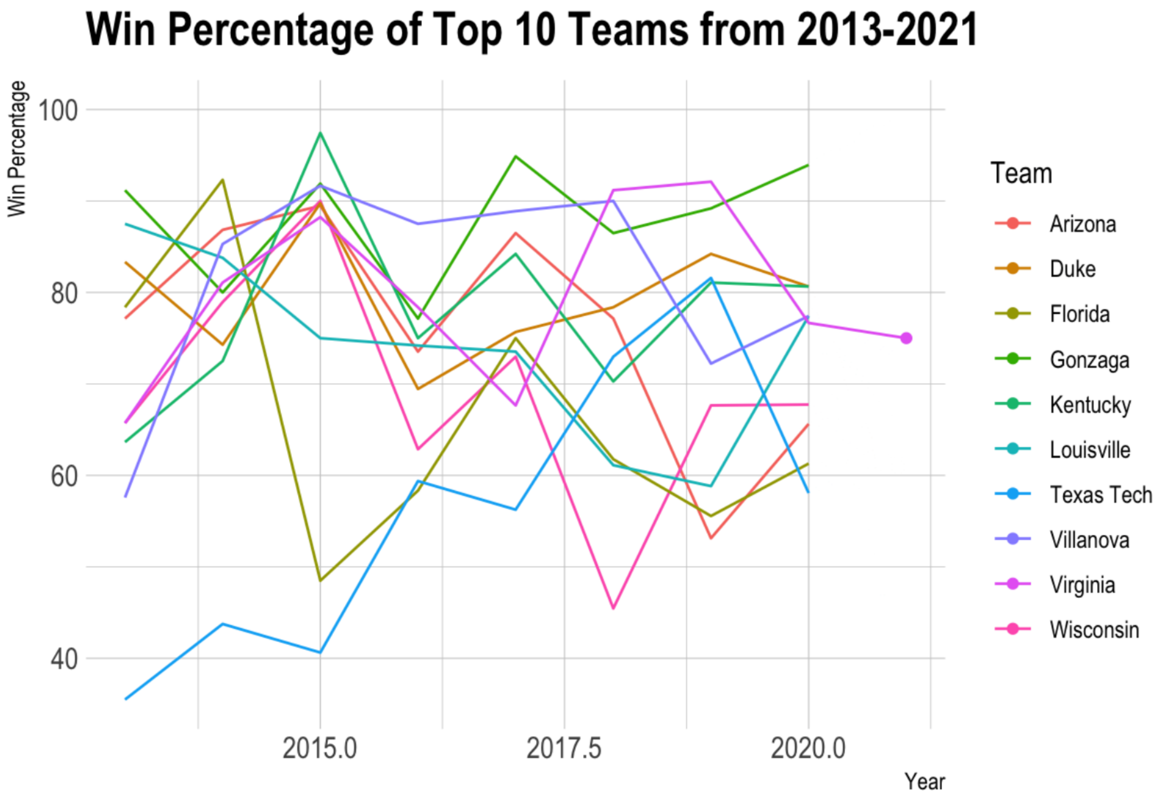
Apart from descriptive, non-numerical values, Group 5 used several quantitative variables to analyze college basketball data and answer the proposed questions. Firstly, we used `YEAR` to indicate a specific year from 2013 to 2021 and `SEED` to indicate a team's seed in the NCAA March Madness Tournament. The NCAA Tournament selection committee determines each team's seed, which is essentially their placement in the bracket and their potential path in the championship. By dividing the variable 'W' indicating the number of games won by the variable 'G' indicating the number of games played and multiplying by 100, we were able to create a new variable called `WIN PERCENTAGE`. This variable was critical to our analysis as we predicted `WIN PERCENTAGE` using either original variables or additional ones added later. Further, `ADJOE` abbreviates adjusted offensive efficiency – the number of points a team scores per 100 possessions. This variable demonstrates how effectively a team can score points while in possession of the ball without turning the ball over to their opponents or getting shots blocked. On the other hand, `ADJDE` abbreviates adjusted defensive efficiency – the number of points a team allows per 100 possessions. This variable measures how effective teams are at defending or, in other words, preventing their opponents from scoring against them. Some may argue that success in college basketball boils down to offensive and defensive efficiency, so these variables were essential in our analysis. `BARTHAG` represents power rating, or the chance of beating an average Division I team. `2P_O` stands for two-point shooting percentage and `2P_D` stands for two-point shooting percentage allowed. On the other hand, `3P_O` stands for three-point shooting percentage and `3P_D` stands for three-point shooting percentage allowed. Effective Field Goal Percentage Shot (`EFG_O`) is a metric calculated by dividing the total points scored by the total field goal attempts. It takes into account the fact that three-point field goals are worth more than two-point field goals. Effective Field Goal Percentage Allowed (`EFG_D`) is closely related to `EFG_O` but examines the opponent’s metric instead. Turnover Percentage Allowed or Turnover Rate, represented as `TOR` in our dataset, is the number of turnovers, or lost possessions of the ball to the opposing team, made by a team per 100 possessions. Turnover Percentage Committed or Steal Rate, known as `TORD` in our dataset, represents the number of turnovers a defensive player causes using aggressive actions per 100 possessions. Offensive Rebound Rate (`ORB`) and Offensive Rebound Rate Allowed (`DRB`) both are good indicators of a team's rebounding performance. The former is a rate of offensive rebounds recovered by the team that shot. The latter is a rate of defensive rebounds recovered by the team guarding the basket. Free Throw Rate (`FTR`) refers to a team's ability to draw fouls and secure points for themselves through free throws. Free Throw Rate Allowed (`FTRD`) refers to a team’s ability to prevent their opponents from getting to the free throw line. Finally, Adjusted Tempo (`ADJ_T`) is an estimate of the tempo (possessions per 40 minutes) a team would have against a team that plays at an average Division I tempo.

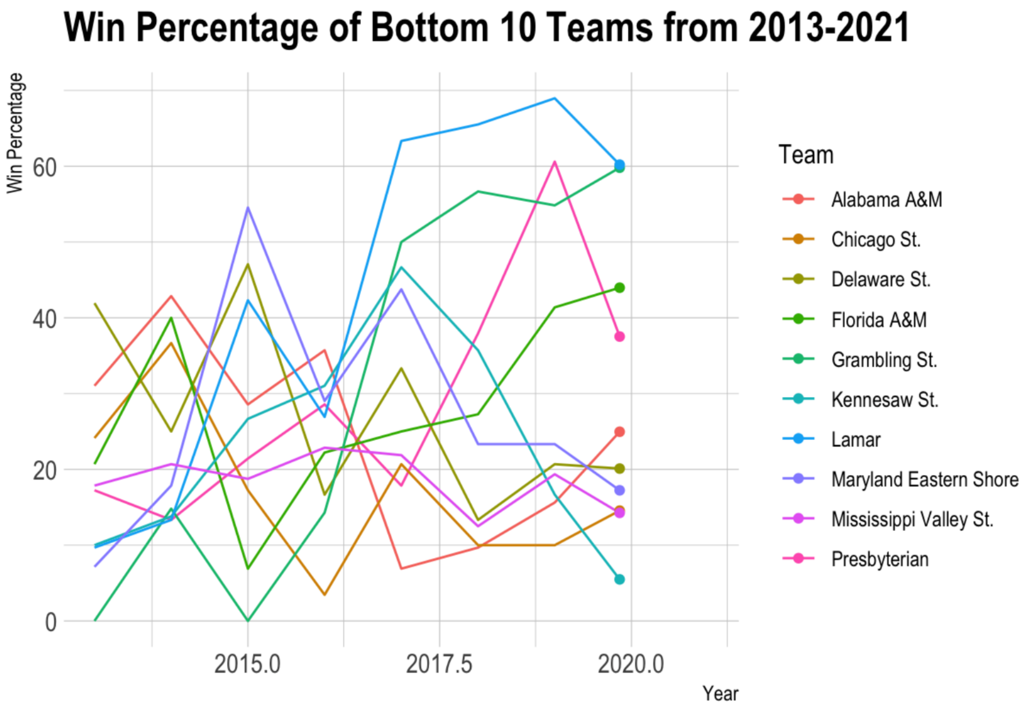
To answer our second question, we web scraped data from 247Sports, a network operated by Paramount that focuses on athletic recruitment in college football and basketball, regarding NCAA men's basketball recruits from 2013 to 2021. From this, we added five new variables to our data set: `NUMBER OF RECRUITS`, `5 STARS`, `4 STARS`, `3 STARS`, and `AVERAGE RECRUIT RATING`. `NUMBER OF RECRUITS` refers to the total number of players committed to each team in a given year, while `5 STARS`, `4 STARS`, and `3 STARS` give the number of recruits with these specific rankings. Each recruit is given a rating out of 100 based on several factors such as talent level, performance in highschool, AAU (Amateur Athletic Union) games, etc. The variable `AVERAGE RECRUIT RATING` refers to the average rating of all the recruits for a given team and is used to gauge overall talent of a program’s incoming players. A high average recruit rate indicates a strong recruiting team overall for a specific team. The following table is a representation of these variables from five teams in 2019:



TEAM	YEAR	NUMBER OF RECRUITS	5 STARS	4 STARS	3 STARS	AVERAGE RECRUIT RATING
Duke	2019	4	2	2	0	99.27
Kentucky	2019	5	3	2	0	98.41
Texas	2019	3	0	3	0	98.03
Memphis	2019	7	2	5	0	97.66
Providence	2019	1	0	1	0	97.43

The main focus of our two questions is to use different variables within our data to predict `WIN PERCENTAGE`. The graphs below display the fluctuation of `WIN PERCENTAGE` for the top 10 and bottom 10 teams between 2013 to 2021. To select the top and bottom teams, we chose the teams that had the highest and lowest average power ratings over the time frame. Since power rating calculates a team's strength relative to other teams in the NCAA, it is a good indicator of team ranking and ability.





As displayed in the two graphs, the top 10 teams tend to have higher win percentages than the bottom 10 teams. Most of the top 10 teams win more than 50 percent of their games during each season. The opposite occurs for the bottom 10 teams, most of which win less than 50 percent of their games. In the rest of our analysis, we dive deeper into the variables that differentiate these teams and impact their win percentages. Determining what factors contribute most to a team's ability to win games, whether its higher two-point shot percentage or better recruits, would help both coaches and players decide what attributes to focus on in order to have the most successful season.

3. Methodology and Results

3.1. Predicting Win Percentage Using Gameplay Variables

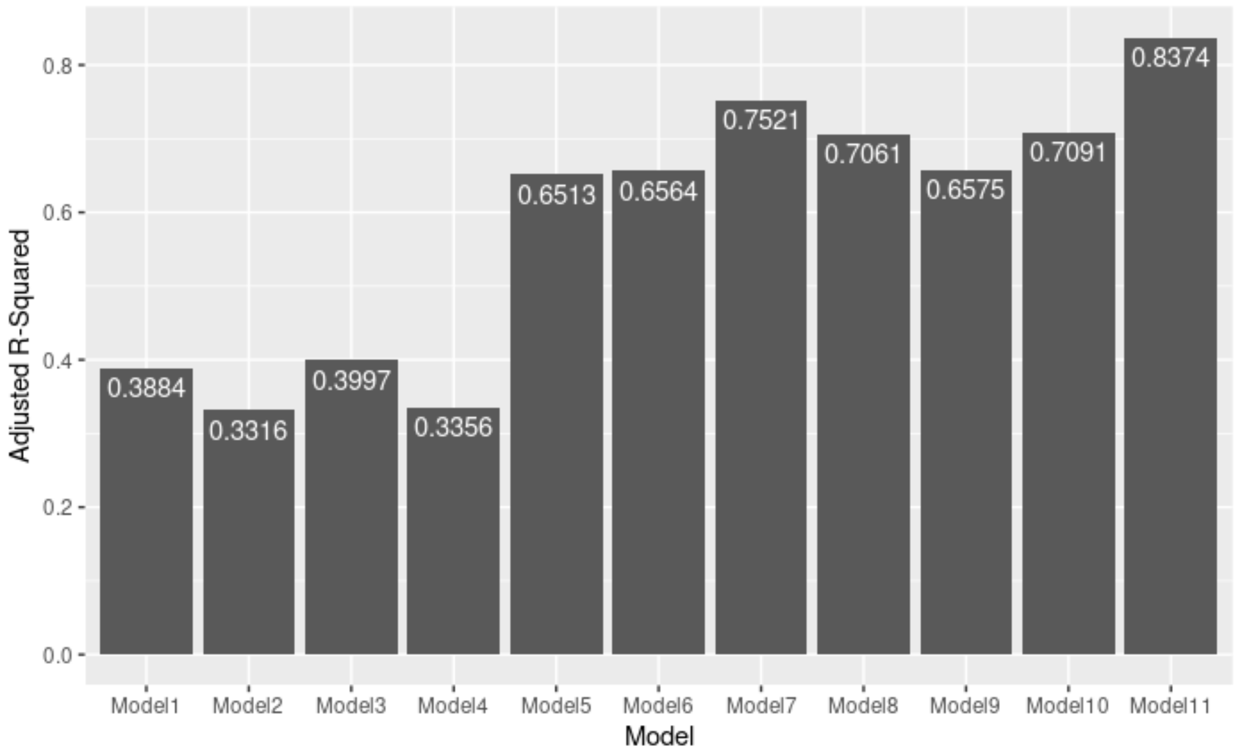
For our first question, we want to find the combination of variables that best predict a team's win percentage. We first split the dataset into a train and test set (70/30) in order to perform cross validation. We then constructed several candidate models in order to determine which regressors best explained the data. In selecting regressors, we chose to focus on gameplay statistics rather than other variables available in the dataset because it would not be particularly useful to know how the year or postseason outcomes relate to win percentage if, for example, a coach wants to know where they should focus their attention in practices in order to improve their team's win percentage.

We began by incrementally adding regressors and noting how they changed the adjusted R-squared of their respective models. We found that simply increasing the number regressors generally increased the adjusted R-squared of the models, up to a ceiling of around 0.83. We made sure not to test models that included both `EFG_O` and `3P_O` or `2P_O` in order to avoid issues with multicollinearity. Similarly, we did not test models with both `EFG_D` and `3P_D` or `2P_D`. We tested the following models:

- Model1: *`WIN PERCENTAGE` ~ EFG_O*
- Model2: *`WIN PERCENTAGE` ~ EDG_D*
- Model3: *`WIN PERCENTAGE` ~ 3P_O + 2P_O*
- Model4: *`WIN PERCENTAGE` ~ 3P_D + 2P_D*
- Model5: *`WIN PERCENTAGE` ~ EFG_O + EFG_D*
- Model6: *`WIN PERCENTAGE` ~ 3P_O + 2P_O + 3P_D + 2P_D*
- Model7: *`WIN PERCENTAGE` ~ EFG_O + EFG_D + TOR + TORD*
- Model8: *`WIN PERCENTAGE` ~ EFG_O + EFG_D + ORB + DRB*
- Model9: *`WIN PERCENTAGE` ~ EFG_O + EFG_D + FTR + FTRD*
- Model10: *`WIN PERCENTAGE` ~ EFG_O + EFG_D + FTR + FTRD + ORB + DRB*
- Model11: *`WIN PERCENTAGE` ~ EFG_O + EFG_D + FTR + FTRD + ORB + DRB + TOR + TORD*

We chose to use EFG_O and EFG_D in later models rather than 3P_O, 2P_O, 3P_D, and 2P_D because they encode very similar information and it minimizes the total number of predictors. The models tested had the following adjusted r-squared values:

```
x = data.frame(Model=c("Model1", "Model2", "Model3", "Model4", "Model5", "Model6",
"Model7", "Model8", "Model9", "Model10", "Model11"), adj_r_squared=c(0.3884, 0.3316,
0.3997,

0.3356, 0.6513, 0.6564, 0.7521, 0.7061, 0.6575, 0.7091, 0.8374))
x$Model = factor(x$Model, levels=x$Model)

ggplot(x, aes(x=Model, y=adj_r_squared)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = adj_r_squared), vjust = 1.5, colour = "white") +
  ylab("Adjusted R-Squared")
```



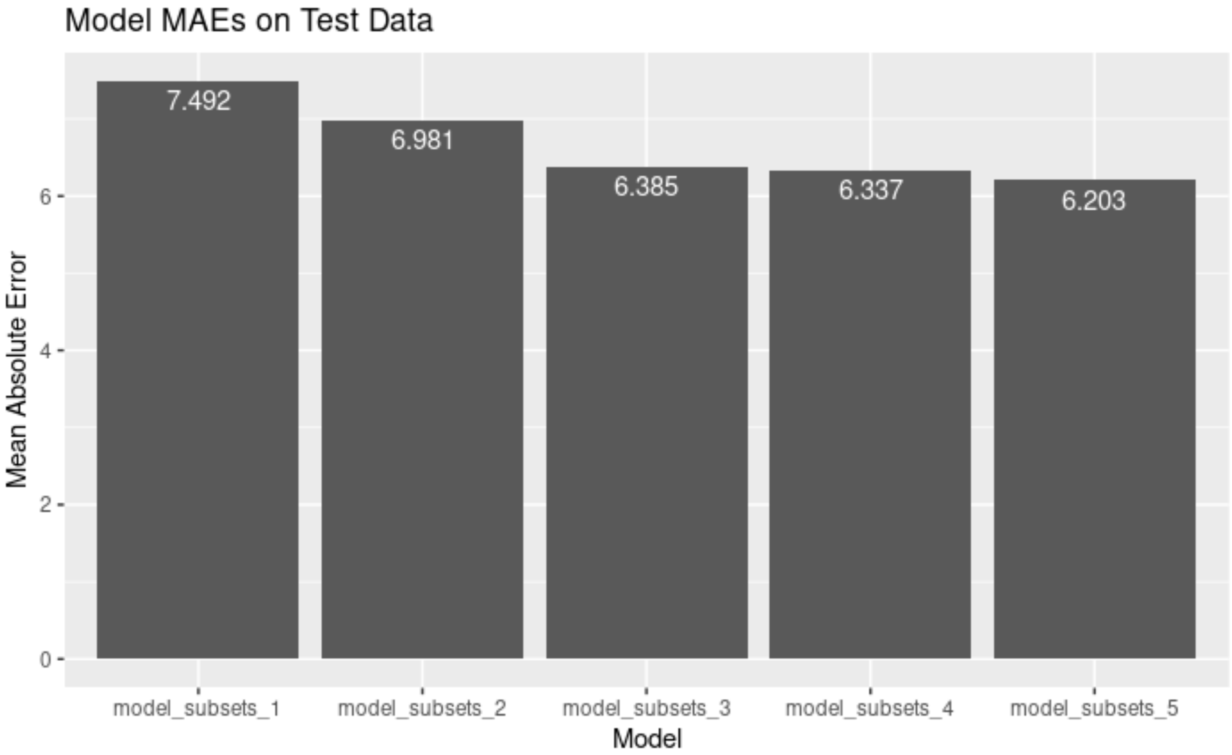
In order to refine our pool of regressors, we decided to use the regular subsets method to perform an exhaustive search over all of the predictors tested in our previous models, in order to find the model that minimizes the Mallows' Cp metric. The top five models produced by the regular subsets method, in order of decreasing Mallows Cp (lower is better), are:

- Model_subsets_1: `*WIN PERCENTAGE*` ~ *EFG_O* + *EFG_D* + *TOR* + *TORD*
- Model_subsets_2: `*WIN PERCENTAGE*` ~ *EFG_O* + *EFG_D* + *TOR* + *TORD* + *ORB*
- Model_subsets_3: `*WIN PERCENTAGE*` ~ *EFG_O* + *EFG_D* + *TOR* + *TORD* + *ORB* + *DRB*
- Model_subsets_4: `*WIN PERCENTAGE*` ~ *EFG_O* + *EFG_D* + *TOR* + *TORD* + *ORB* + *DRB* + *FTRD*
- Model_subsets_5: `*WIN PERCENTAGE*` ~ *EFG_O* + *EFG_D* + *TOR* + *TORD* + *ORB* + *DRB* + *FTRD* + *FTR*

These models had Mallows' Cp values of 1204.66, 735.12, 237.60, 122.42, and 38.16 (in order from Model_subsets_ 1 to 5). Additionally, Model_subsets_1 is identical to Model_7 from before Model_subsets_5 is identical to Model_11 from before.

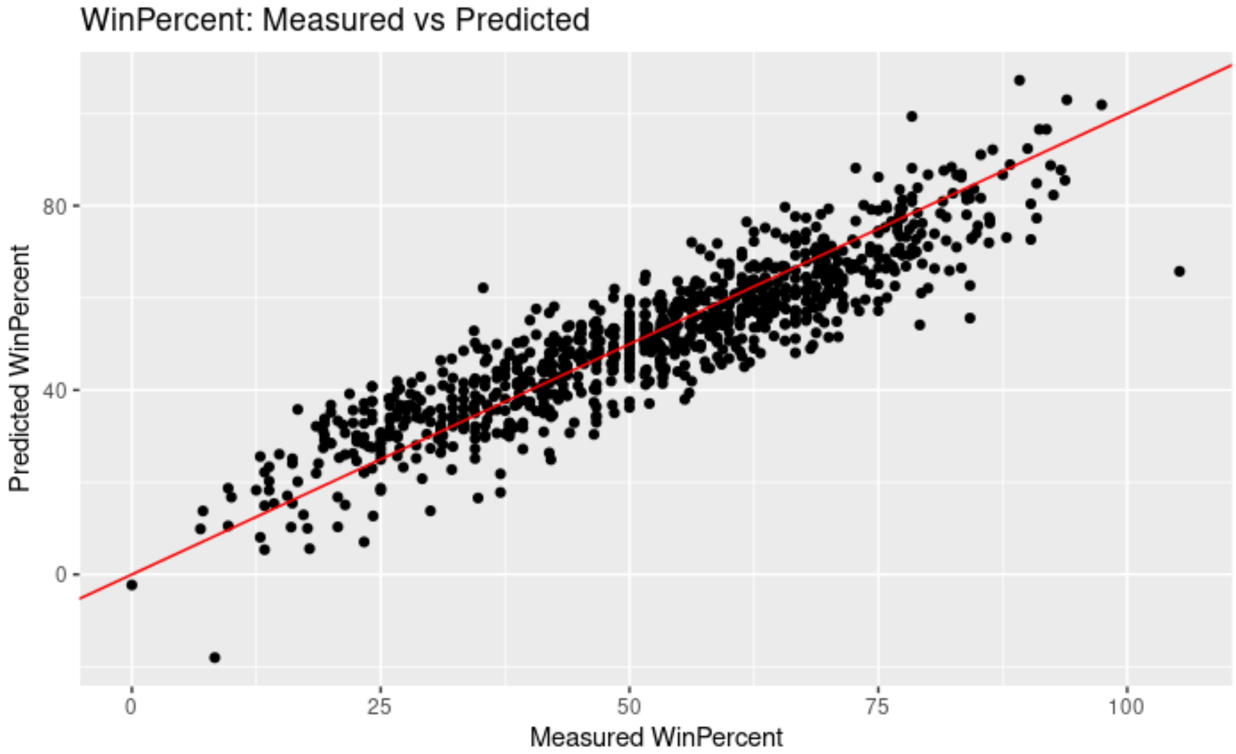
In order to test our new models, we calculated the Mean Absolute Error (MAE) between actual `WIN PERCENTAGE`s in our test data and the predictions made by our new models. The following chart illustrates the MAEs of the top five models generated by regular subsets:

```
out = data.frame(Model=c("Model_subsets_1", "Model_subsets_2", "Model_subsets_3",
"Model_subsets_4", "Model_subsets_5"), MAE=c(7.492, 6.981, 6.385,

6.337, 6.203))

out$Model = factor(out$Model, levels=out$Model)
ggplot(out, aes(x=Model, y=MAE)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = MAE), vjust = 1.5, colour = "white") +
  ylab("Mean Absolute Error") +
  ggtitle("Model MAEs on Test Data")
```



As model_subsets_5 has the lowest Mallows’ Cp value and the lowest MAE, we decided to choose it as the best model for predicting `WIN PERCENTAGE`. We evaluated model_subsets_5 further by plotting actual `WIN PERCENTAGE` values in the test data versus the model’s predicted values. The resulting graph is shown below. We can see that the fit is generally fairly good but the model does tend to slightly underpredict for higher `WIN PERCENTAGE`s and slightly overpredict for lower `WIN PERCENTAGE`s.

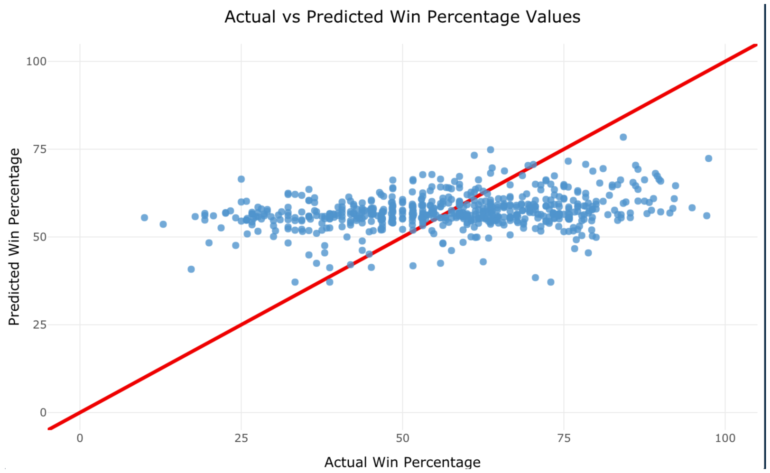
```
mod_subsets_5 = lm(WinPercent ~ EFG_O + EFG_D + TOR + TORD + ORB + DRB +
FTRD + FTR, data=train)

preds = test %>% add_predictions(mod_subsets_5)

ggplot(preds) +
  geom_point(aes(x=WinPercent, y=pred)) +
  geom_abline(color="red") +
  xlab("Measured WinPercent") +
  ylab("Predicted WinPercent") +
  ggtitle("WinPercent: Measured vs Predicted on Test Data")
```



One key aspect to note about this graph is that it displays a few negative predictions for `WIN PERCENTAGE` for values on the lower extreme of measured `WIN PERCENTAGE`. In a real world setting, obviously, this is impossible. Thus, this is a weakness of our model and if it was actually in use, all predictions would need to be constrained to the range [0,100].

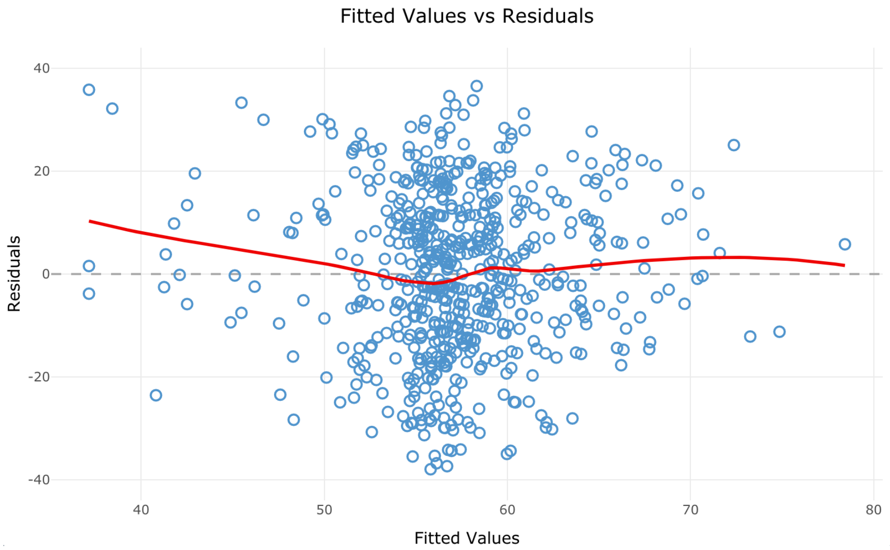
3.2. Predicting Win Percentage Using Recruiting Data

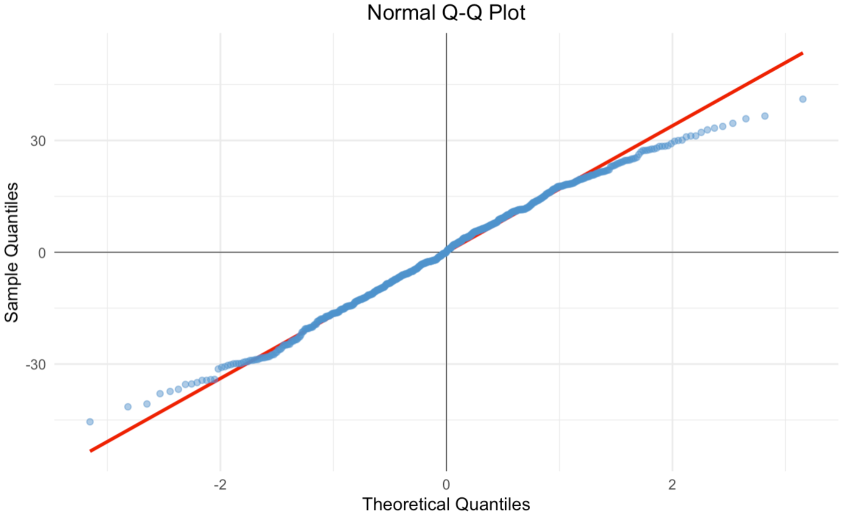
To answer our second question, we compared models using different combinations of `NUMBER OF RECRUITS` and `AVERAGE RECRUIT RATING` to predict `WIN PERCENTAGE` in order to determine if recruiting factors could be used to accurately predict how well a team will do during a given season. Since the 2020-2021 season was cut short due to COVID-19, we decided to only examine data from 2013 to 2019 when answering this question. We first randomly assigned 70% of the data to a train set and the other 30% to a test set for the sake of cross validation. The first model used `NUMBER OF RECRUITS` to predict `WIN PERCENTAGE`, but the R-squared value was only 0.00537, indicating that `NUMBER OF RECRUITS` alone is not a good predictor for `WIN PERCENTAGE`. The next model we created predicted `WIN PERCENTAGE` using `AVERAGE RECRUIT RATING`, which resulted in a R-squared value of 0.0530. Although higher than the previous model, this R-squared value was still very low, meaning this model also did not explain the data well. Finally, we created a third model that used both `NUMBER OF RECRUITS` and `AVERAGE RECRUIT RATING` to predict `WIN PERCENTAGE`, finding an R-squared value of 0.0535, which was only a 0.0005 improvement from the second model. When looking at the summary for this model, `AVERAGE RECRUIT RATING` had a p-value less than 0.05 while `NUMBER OF RECRUITS` had a p-value greater than 0.05, meaning `AVERAGE RECRUIT RATING` was significant as a predictor in this model while `NUMBER OF RECRUITS` was not. Since it had the highest R-squared value, we further examined the residuals for the third model. For the train set, the root mean squared error was 19.38 and the mean absolute error was 15.64, both of which are high error values, indicating that the model could be improved. Similar large error values were found when using the test set, with a root mean squared error of 18.99 and a mean absolute error of 15.35. After attempting various transformations of the model, we found that adding an interaction term between `AVERAGE RECRUIT RATING`and `NUMBER OF RECRUITS` resulted in the most improvement. Since the presence of highly rated recruits influences the number of players that commit to a school and vice versa, adding an interaction term to reflect the influence these variables have on each other improved the model. This model had an R-squared value of 0.087, which was still low but an increase from the previous model. Both the interaction term and `NUMBER OF RECRUITS` term had p-values below 0.05, meaning that they were significant predictors for the model. However, the `AVERAGE RECRUIT RATING` term had a p-value of 0.821, meaning there is no evidence that this term has a nonzero regression coefficient and therefore it is not statistically significant as a predictor in this model. This is the opposite of the previous model, in which `AVERAGE RECRUIT RATING` was significant and `NUMBER OF RECRUITS` was not. The interactive graph below displays the actual `WIN PERCENTAGE` values versus the values predicted using the transformed model.



(graph is interactive in R so you can see the values for Actual vs Predicted for each point)

The points do not closely follow the reference line, which represents perfect prediction, but instead fan out from the center of the line. This means that the model does not do a good job of predicting `WIN PERCENTAGE`, as indicated by the low R-squared value. Also, the win percentages that the model predicts are all between 37 to 85 percent, but the actual win percentages range from 10 to about 98 percent. This indicates that the model does not do a good job of predicting extreme values. This is further illustrated by the figures below. In the fitted versus residuals graph, the red line that models residual behavior tends to move away from the reference line that represents zero error as the fitted values move away from the center, indicating that residuals tend to increase as the values become more extreme. This is further emphasized in the Normal Q-Q plot, as the residuals skew from both the left and right sides of the reference line. Also, the residuals are widespread on the fitted versus residuals graph, ranging between -40 and 40, displaying that many of the predicted values greatly differed from the actual `WIN PERCENTAGE`.





Overall, we found that recruiting data is not a good predictor for how well a team will perform during a season, even when considering both amount and average rating of recruits. Going into this question, we expected recruiting data to have a high impact on `WIN PERCENTAGE`, but discovered the opposite. The models we created using recruiting variables to predict `WIN PERCENTAGE` all had low R-squared values, meaning they did not explain the data well. The best model we discovered had high residuals and did not predict extreme values well. A possible explanation for why recruiting data is a bad predictor is that playing basketball at a college level can be a lot more pressure than playing in high school, so some recruits may not reach the high level they were predicted to. Also, these recruits are playing with experienced college athletes and facing tougher opponents, which could cause them not to perform as well as they did when they were the top players in high school.

4. Conclusion

The first focal point of our paper was analyzing which variables from our dataset were best suited to predicting `WIN PERCENTAGE`. We constructed several baseline models and analyzed their adjusted r-squared values on our training data. We then used the regular subsets method to find the model that minimized the Mallows' Cp metric. We found that the model that best satisfied this metric included a majority of the predictors available, indicating that they are all useful in predicting `WIN PERCENTAGE`. Our best model had a MAE of 6.203 on the test data and could be seen to predict `WIN PERCENTAGE` fairly well in the measured vs. predicted graph. The fact that the best model included quite a few predictors makes sense intuitively because it makes sense that a well rounded team that is good at all skills will generally do well in their games. However, our data does not technically measure "all" potential metrics for analyzing basketball play so future work could include collecting and analyzing additional data related to injury statistics or play time for the starting lineup.

Our results for the best predictors for `WIN PERCENTAGE` are quite important for those who aim to understand the performance of college basketball teams and potentially for coaches who aim to improve the performance of their teams. We originally wanted to determine the best subset of predictors for `WIN PERCENTAGE` so that coaches could better optimize their time in practices on the more relevant skills for their players. However, we found that the majority of predictors are relevant, so no one key subset of skills emerged as more directly tied to `WIN PERCENTAGE` than the others. However, as mentioned above, our data does not fully encompass all potential statistics about basketball. Our data focuses on a few key metrics about technical skills, but it would also be very useful to have data about non-technical aspects of the game, such as those mentioned above. Additional non-technical statistics that would be useful for further analysis include NCAA violations, past coach performance, and program budget.

After determining the best combinations of variables to predict `WIN PERCENTAGE` using our original data set, we decided to look specifically at how recruiting data relates to how well a team does during a season. To do this, we posed the question "is recruiting data a good predictor for win percentage?" Going into this question, we expected recruiting data to have a high impact on `WIN PERCENTAGE`, but discovered the opposite. After examining multiple models with different combinations of `AVERAGE RECRUIT RATING` and `NUMBER OF RECRUITS`, we found that recruiting data does not do a good job at predicting win percentage. Most people assume that a larger number of high rated recruits means that a team will perform better. Teams that have the top recruiting classes are expected to win more games, but we discovered that this is not necessarily the case. To take this analysis further, one could look at if recruit tenure impacts win percentage. By staying longer than a year, these recruits could continue to improve and the team would gain more experience playing together, which could ultimately improve their overall performance. In the real world, these findings could encourage teams to pursue recruits that would stay longer as opposed to highly rated recruits that would likely leave after one season. Additionally, the data we used for our models only included the ratings for recruits based on their performance prior to playing in college, so examining the accuracy of these ratings by looking at data on how well these recruits actually played could be useful in further analysis.