**College Basketball Analysis**

-------------------

Andrew Willett

May 6, 2022

Every year March Madness is one of the most talked about sporting events and millions of people make brackets trying to predict who will win each of the games. I thought it would be a perfect time to look deeper into the numbers and see what leads a team to victory. We will see which basketball stats correlate with winning the most, and by the end I should be able to compare my results with the games that are happening to see if the outputs I got would make better predictions.

To get these outputs I will be using data from Data Camps website. This data has information about all regular season and tournament games from 2003 to 2018. Each of the games have most of the basic basketball stats like team score, assists, rebounds, turnovers, fouls, and other common basketball stats. I will use these to try and make predictions based off the team’s stats and compare how basketball is played today versus back in 2003 to see if there are a lot of differences.

-------------------

**Creating Models**

The first thing I looked at after organizing and cleaning up the data was to try and find which variables would make the best model in correlation to winning games. I found that these values make up the model of best fit, these values are score, fouls, field goals attempted, defensive rebounds, steals, turnovers, offensive rebounds, free-throws attempted, and blocks. This order of values listed is the order of importance to the model. So, I decided to start looking more into these values that have a bigger impact on the model, the first one I Chart, scatter chart

Description automatically generatedinvestigated was personal fouls I choose this one first because it has the second biggest impact on the model behind team score. This graph on the right shows the win percentage at each number of personal fouls in a game. Teams with 5 fouls a game have over an 80% chance to win the game but teams with 30 fouls only have a 20% chance to win. This graph isn’t too surprising to me, the more you foul the more points the other team will get, and the goal of basketball is to get the most points. When I looked at the next most important value in the Chart, scatter chart

Description automatically generatedmodel (field goals attempted) I was surprised with the results. The data shows that you have a higher chance to win the game when you attempt less shots, I would have thought that the more shots you attempt would lead to more made shots which would lead to more wins. This graph shows that teams with less shots have a better chance to win.

-------------------

**Has the game changed over the years?**

Chart, line chart

Description automatically generatedThe next thing I investigated was to see how much the game has changed over the years. The graph from the paragraph above shows you that teams are taking a lot more shots in recent years. If we are going to be predicting which teams are going to win right now, the data from back in 2003 could be hurting the models if the game has changed over the years. I made a model that looks at the last three years to see if it is different than the overall model. Both models are similar, there are a couple small differences in which variables have more correlation to winning but the difference is small. I am surprised that the difference isn’t bigger with how the game has changed, an example of this is how many more three pointers are attempted now verses 15 years ago. This graph on the right shows three-point attempts per game throughout the years, I would have thought since the three-point shot is so much bigger in today’s game it would have had more of a correlation to winning games. I also made a model that looked at the first three years of the data and found that the model is pretty similar to the model from 2016-2018, both models say that all the variables are significant other than made three pointers. Both the model for 2003-2005 and the model for 2016-2018 have Score, field goals attempted, defensive rebounds, and personal fouls as the three most important variables to the models. It is interesting that even after 15 years that the best linear model to predict college basketball games are very similar.

-------------------

**Regular season vs tournament**

Chart, scatter chart

Description automatically generatedThe next thing I wanted to look into was the difference in how the game is played in the regular season versus the tournament. I created a model for both regular season and for the tournament using the same variables in each to see if the different variables have the same significance to the model. In the regular season model, all the variables are significant other than three pointers made per game, but in the tournament data it shows us that field goals made, three pointers made, and assists aren’t significant, and blocks are less significant too. I was surprised to see that field goals made was not significant in the tournament model, this graph above shows the win percentage at each number of field goals made and it looks like it would have more of impact on the overall model.

-------------------

**Predictions**

Now that I have made different models on the regular season and the tournament, I will use those models to see how accurate of predictions I can make. First, I will see if I can predict teams’ records based off all the variables, the model I used was WLnum ~ Score + FGM + FGA + FGM3 + FGA3 + FTM + FTA + OR + DR + Ast + TO +Stl + Blk + PF. I took the season average for each of these stats for all the teams that made the tournament and put them into the model and that gave me the teams predicted win percentage for the season. I after finding the predicted win percentage for all the teams in the tournament I went and found what their actual win percentage was for the 2021-2022 season and compared it to the predicted win percentage that the model found for each team. The average difference between their actual win percentage and the predicted win percentage was 0.07. I feel like this model does a good job of predicting win percentages for the season based off the stats of each team. I also wanted to look into predicting the games in the March Madness tournament. For this I used the same model as above, I just use the tournament data set instead of the regular season one. Graphical user interface, application

Description automatically generatedThe only variable I added was the seed of the team. This picture shows the bracket that I made using the model. This wasn’t as good of a predictor as the season win percentage, this bracket ranked in the 73rd percentile of all the brackets created on ESPN. I predicted only 53.97% of games right in the bracket, that is when picking all the games before the tournament started. When doing my prediction for the tournament I noticed a couple flaws. One of them being the strength of schedule of the different teams in the tournament. Some teams play easy schedules throughout the regular season so that makes their numbers look better, but other teams that play in harder conferences might not have as good of averages. If you want to make a better prediction for the tournament you would need to find a stat that can take strength of schedule into the equation. This didn’t really affect the regular season prediction because teams are playing against other teams with similar strengths of schedules so you wouldn’t need to worry about this in the regular season. Since the tournament are all one game eliminations it makes it harder to make predictions since all it takes is one player playing well or someone having one bad game to decide games in the tournament. These two problems aren’t as prominent in the NBA, which is why the NBA can be so much easier to predict than college basketball.

-------------------

**Conclusion**

Some of the results of this paper really surprised me, like how three pointers made is not significant in seeing who will win the game. I also thought that we would see more change in the model from 2003 to 2018 but the same values are significant for both, there was a little more of a difference when looking into tournament data, with field goals made and assists not being as important in the tournament as they were in the regular season. This analysis helps show that the regular season is so much easier to predict that the tournament, in the regular season my predicted win percentage was a lot better than the bracket that I filled out using this analysis. That is one reason why so many people love watching March Madness and filling out the brackets is because it is almost impossible to predict.

**Publication:**

<https://journalofsportsanalytics.com/>