College Basketball Season Wins

Michael Prodo
MS-DS
University of Colorado
Boulder, CO United States

ABSTRACT

When we talk about college basketball everyone likes to focus on the NCAA March Madness tournament that is held at the end of every season. Which schools are going to make the tournament? Who's going to be the Cinderella team? Who's going to make it far or go out early? Which school is going to win the whole thing? What people don't discuss as much is all the winning games it takes to get there. There are over 350 schools across the United States that have a college basketball program, but only 68 schools can make the tournament. In this study we will use college basketball team stats from the past five seasons (2017-2021) and perform data analysis to see if we can predict how many wins a team will get in a season using the previous year's stats.

A few models were constructed and compared using evaluation metrics. It was found that the best model is a formula called the sports Pythagorean Theorem, created by Bill James, with an error rate of about 11% on average.

INTRODUCTION

How well will a team do in a season? Will a team get a high amount of wins to have the chance of being in the tournament? Young basketball players dream of playing in the NCAA March Madness Tournament, going all the way, and cutting down the nets as champions at the end. Elite high school basketball athletes get recruited by college teams to go play for the school, giving these athletes a difficult choice of deciding where to go play in college. This data analysis can be used by these high school players to help them make that decision on what school to play for that gives them the best chance of living out their dream of playing in the tournament. Another side to this data can be used by sports gamblers to research what college basketball teams will perform better, or worse, than predicted by sportsbooks. This research can help make smarter bets.

Most people only really care about college basketball in March when the NCAA Tournament comes around. That's when the experts discuss the sport and analysts take a hard look at the data. However there isn't a whole lot of talk before the season even begins about a team's overall performance in the upcoming season. When there is discussion and analysis, it's always about the big name schools, such as Kansas or Duke. We will analyze the data to look at a team's overall season performance for all the teams in the NCAA, not just the schools that are always talked about.

RELATED WORK

There doesn't appear to be much already done when it comes to analyzing and predicting college basketball. Much of the college basketball analysis seems to be about the March Madness tournament at the end of each season and not about the season as a whole. The discussion is about what's going to happen in the tournament to teams we already know are in, but what about analyzing teams at the beginning of the season to predict who could be in that position of playing in the tournament.

There is a book written by Wayne L. Winston called "Mathletics" that discussed the work of Bill James on Major League Baseball. Bill James came up with a formula, called Baseball's Pythagorean Theorem (1), used to calculate the number of wins for a baseball team using runs scored and runs allowed in that season.

$$\frac{runs\ scored^2}{runs\ scored^2 + runs\ allowed^2} = win\ percentage \qquad (1)$$

This study will build upon Bill's work and create a similar formula to calculate college basketball win percentage using points scored and points allowed. However, it will be using the previous year's data to predict the win percentage for the current season.

CONDUCTED WORK

Initially, there will be a total of five datasets gathered from sports-reference.com. There is a dataset for each season for the past five years (2017-2021). These datasets will contain a compilation of stats for each team for that season including wins, strength of schedule, points, and many more game

stats. There are over 30 variables for over 350 schools in each dataset.

The programming language that was used for this study is R, a statistical program. The five datasets for the five years of college basketball data has been read into R. These datasets have been cleaned to remove blank columns and repeated rows that contained labels, which had been used for easy readability on their website. Columns were labeled appropriately for easy reference. The school column has been reformatted to remove unnecessary text so that a specific school can easily be analyzed year to year if desired. When the datasets were read into R, all the numbers were read as characters, so they were converted back to numeric. A year column was added to each dataset, and the datasets were row bound to create one large dataset for easy storage and analysis. Rows that contained a missing value were removed as the missing value was in a column that could not be replaced. Additional variables (PPG, Opp PPG, TOV pG) were calculated and added to the dataset as these variables would be used in the data modeling creation. The next year's stats (Wins, SoS, Win%, Games) were also added as these will ne needed for the data modeling.

Now that the dataset has been cleaned and warehoused, we could now create visualizations and learn more about the data, or more importantly, the data we are interested in.

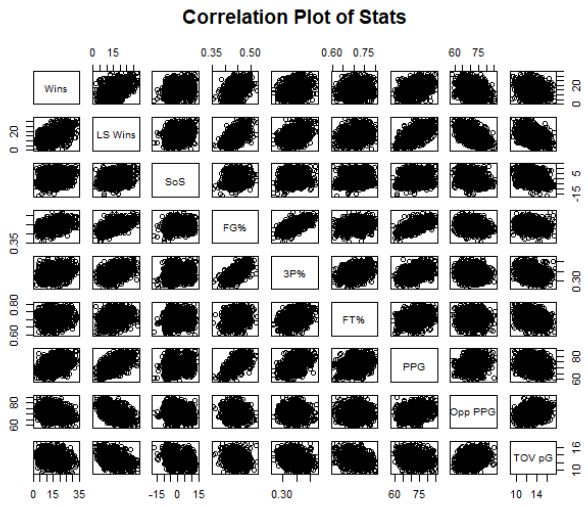

Figure 1: Correlation plot of a handful of stats that are considered as more prominent and important in the sport of basketball.

Figure 1 shows a correlation plot of the main stats we'll be looking at for this study. It shows that there is a strong correlation between the variables 'FG%', '3P%', and 'PPG'. This suggests that only one of those three variables should be used in the model for predicting the total number of wins in a season. We will use 'PPG' and remove the other two as 'PPG' encompasses both 'FG%' and '3P%.' We can also see in the correlation plot that last season's wins is correlated with almost every other variable, therefore we can also remove this variable so that there isn't redundancy in our models. For this study, 'Wins' is our dependent variable, and 'SoS', 'FT%', 'PPG', 'Opp PPG', and 'TOV pG' are our possible predictors.

Now that we've identified the variables we will be using, we will construct boxplots for each of these variables to detect any outliers in the data. In the sport of college basketball, there are a number of factors that can influence the game and cause discrepancies in the stats year to year. For example, rosters are constantly changing as players are graduating or going to play professionally and new players are coming in from high school. Rules can also change year to year as well as referees focus on different things each year that can influence the game. Therefore yearly side by side boxplots will be made for each variable to identify outliers for each year.

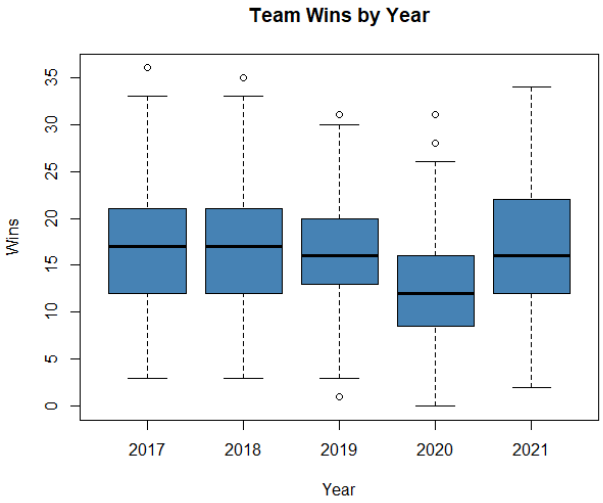

Figure 2: Yearly Side by Side Boxplots for season wins.

In Figure 2 above, we can see there are about one to two outliers each year. 6 of the 7 outliers are on the top end of the distribution: Villanova (2017), Virginia (2018), Gonzaga (2019), Gonzaga (2020), Baylor (2020), and Houston (2020). 4 of those 6 teams either won the entire tournament or made it to the championship game: Villanova (2017 Champs), Virginia (2018 Champs), Gonzaga (2020 2nd Place), Baylor (2020 Champs). It would appear that these teams are not true outliers but just played more games. However, looking into it further, only one of those teams (Villanova 2017) were tied for the second most games played in the respective season. The one team that was an outlier on the bottom end of the

distribution (Kennesaw State 2019) was tied for the seventh fewest games played for that season. All of these outliers are believed to be true outliers and have been removed from the dataset.

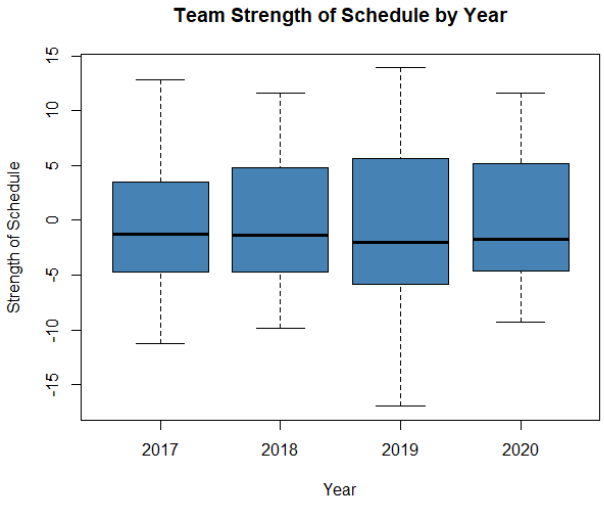

Figure 3: Yearly Side by Side Boxplots for Strength of Schedule.

Looking at the side by side boxplots for strength of schedule in Figure 3, there are no apparent outliers. 2019 had the most spread out data for strength of schedule, but still didn't have any outliers.

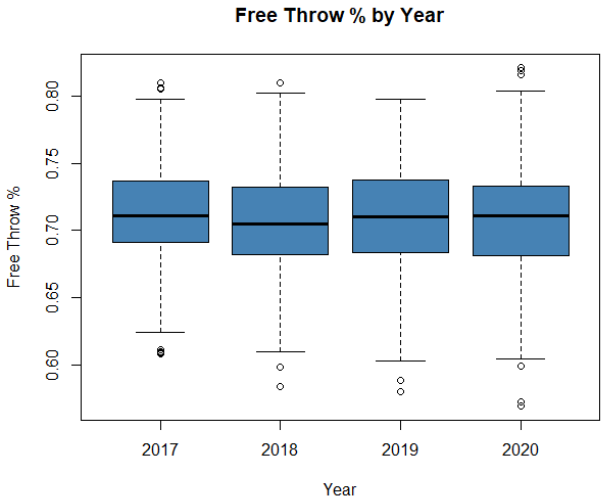

Figure 4: Yearly Side by Side Boxplots for Free Throw Percentage.

In Figure 4 for Free Throw Percentage, there are a total of 18 outliers across the four years of data. All outliers are either above 0.805 or below 0.611. None of these teams attempted less than 110 free throws so they aren't outliers due to a low number of attempts, therefore these outliers have been removed from the dataset.

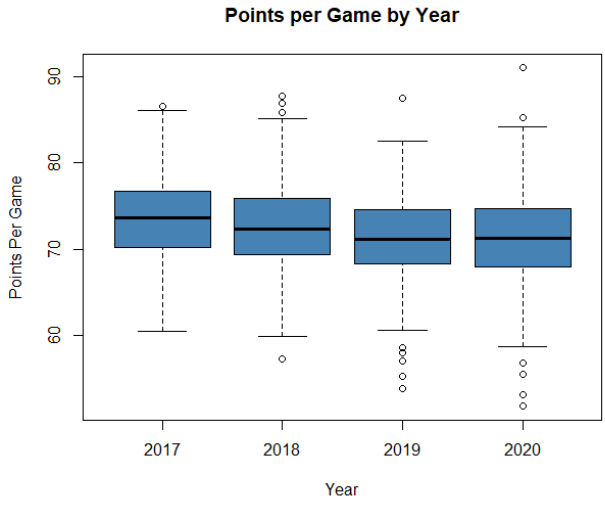

Figure 5: Yearly Side by Side Boxplots for Points per Game.

In the side by side boxplots for points per game (PPG) in Figure 5, there are a total of 18 outliers for the four years. There was only one outlier in 2017 while there were 5 or 6 outliers in each of the other three years. Gonzaga was an outlier for this stat in every year except 2017. All of the outliers were either above 85.1875 PPG or below 58.58065 PPG. However in 2020, 4 of those 6 outliers only played no more than half the number of games in a typical season: Colgate (16), Fordham (14), Maine (9), Chicago State (9). This suggests that these four data points may not be true outliers. Rather than removing them from the data, we calculated what each team's PPG would be for the rest of the games if they were to play a full season using the team's past 3 years of data, and recalculated their PPG for the full year of 2020 to use that value instead. The rest of the outliers played a full season of games and thus were removed from the data.

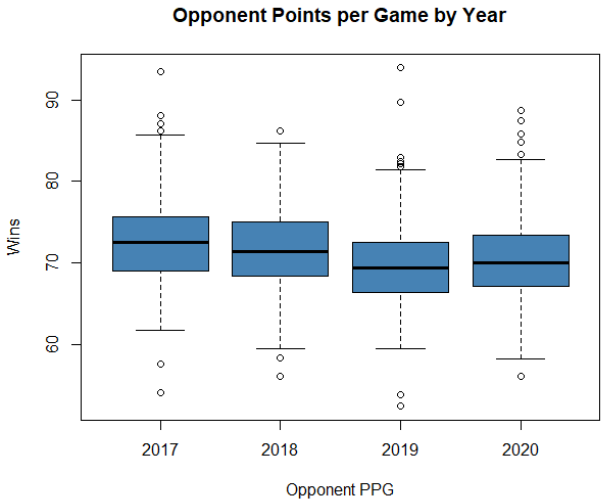

Figure 6: Yearly Side by Side Boxplots for Opponent Points per Game.

The variable with the most outliers is Opponent Points per Game (Opp PPG) with a total of 25 as seen in Figure 6. 2019 has the most outliers for this stat with 10. Both Virginia and Chicago State are an outlier in 3 of the 4 years. All of these outliers have allowed either more than 81.76923 or less than 58.32432 points per game. Just like with Points per Game in Figure 5, we run into some issues with the outliers in 2020. Two outliers, Chicago State and Howard, played a total of 9 and 5 games, respectively, for that year. Now since Chicago State is also an outlier in two of the other three years of data, we can believe that Chicago State is also an outlier in 2020. However with Howard, we believe they are only an outlier because they only played 5 games, about one sixth of the total games in a typical season. Therefore we won't remove Howard from the data, rather we will recalculate their expected Opp PPG using data from the past three years for that school. The rest of the outliers are removed from the data.

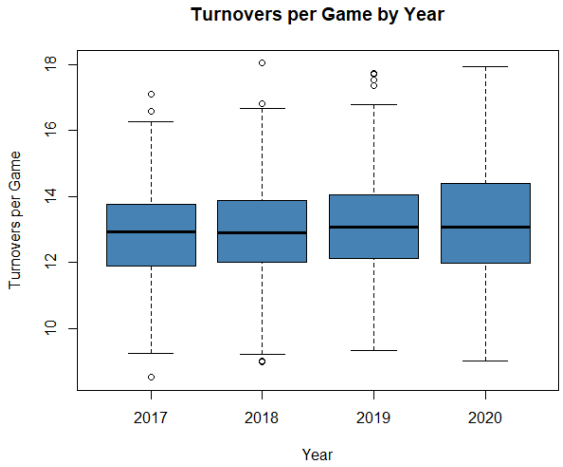

Figure 7: Yearly Side by Side Boxplots for Turnovers per Game.

In Figure 7, we can see there are 12 outliers from 2017 to 2019, and none in 2020. Only three of these outliers were on the low end of the spread. These teams have averaged more than 16.58085 or less than 9.027027 turnovers per game. All these schools have played around the typical number of games in a season, all at least 29 games, so they are all believed to be true outliers and are removed from the data.

Now that all the outliers have been identified and removed from the dataset, we can move on to the data modeling. But first we need to split the data into training and testing datasets. The models will be constructed using the training dataset, and then tested for accuracy using the testing dataset. The data has been split randomly with 75% of it into the training and 25% into the testing set. The variables that will be used in the models and their abbreviations can be found in Table 1 below.

Abbreviation	Variable
Wins	Season Wins
SoS	Strength of Schedule
FT%	Free Throw Percentage
PPG	Points Per Game
Opp PPG	Allowed Points Per Game
TOV pG	Turnovers Per Game
Games	Season Games
TmPts	Team's Total Points
OppPts	Total Points Allowed

Table 1: Variable Legend

Our modeling process will begin with using R's lm() function to create the full linear model (Model 1) that uses every variable to predict the total number of season wins.

Model 1: Full Linear Model

$$Wins = 11.49 + 0.16 * SoS + 0.80 * FT\% + 0.50 * PPG - 0.45 * Opp\ PPG - 0.07 * TOV\ pG$$

From there we will use backwards elimination to create a reduced model. Backwards elimination is the process of removing the variable with the highest p-value, i.e. the most insignificant variable, one by one until every variable in the model is significant. Using this process for our model, 'FT%' had the highest p-value (0.8677) in the full model, meaning it should be the first variable removed. Creating a model with all predictors except 'FT%' yields a model with only 3 of 4 predictors being significant. 'TOV pG' has the highest p-value (0.5323) in this model, and thus it gets removed from the model. A new reduced model with predictors 'SoS', 'PPG', and 'Opp PPG' is created. In this model, all three pedictors are significant having p-values less than 0.001. This model (Model 2) is our reduced linear model that we will use and test to predict the number of wins.

Model 2: Reduced Linear Model

$$Wins = 10.99 + 0.16 * SoS + 0.51 * PPG - 0.45 * Opp\ PPG$$

Next we will create another model, a generalized additive model (GAM). This type of model has one or more predictors in the model as a function of the predictor, denoted as s(). We will start with a full GAM model where every predictor is a function. The effective degrees of freedom (edf) for the predictors are used to determine if they should enter the model linearly rather than as a function. In this full GAM model, the edf for 'SoS', 'FT%', and 'Opp PPG' all have edf value of 1, meaning they can enter the model linearly. Next a GAM model is created with those three predictors listed above are linear and the other two variables, 'PPG' and 'TOV pG', are a function. From here we use backwards elimination again to remove insignificant predictors one by one until every predictor is significant. 'FT%' is the first predictor removed having a p-value of 0.824. Next predictor to go is the function of 'TOV pG' having a p-value of 0.34. This leaves us with a model (Model 3) where every predictor is significant and with 'SoS' and 'Opp PPG' are linear and 'PPG' is a function.

Model 3: Generalized Additive Model

$$Wins = 48.23 + 0.16 * SoS + s(PPG) - 0.46 * Opp\ PPG$$

where s(PPG) is a function of PPG

The last model we will construct is a Pythagorean Theorem model. How this process works is we will use the formula (2) below on the training data with different values for 'x' to find the exponent that produces the smallest average absolute error, as defined in formula (3).

$$Win\ Pct = \frac{TmPts^x}{TmPts^x + OppPts^x} \quad (2)$$

$$Abs\ Error = abs(acutal\ win\% - predicted\ win\%) \quad (3)$$

First a vector of exponents ranging from 0.5 to 20 in increments of 0.5 was created and used in formula (2) for the training dataset. The exponent that produced the smallest average absolute error was 5 (0.1107). To break it down farther and try to find a more accurate error, we repeated the process with a vector of exponents ranging from 4.5 to 5.5 in increments of 0.1. The exponent that had the smallest absolute error was once again 5. Therefore we will use a Pythagorean Theorem model (Model 4) with an exponent of 5 to predict the number of season wins.

Model 4: Pythagorean Theorem Model

$$Wins = Games * \frac{TmPts^5}{TmPts^5 + OppPts^5}$$

These four models will be tested to find the 'best' model for predicting the number of wins.

EVALUATION

Anomaly detection was conducted to deal with any outliers in the data, such as removing them, before creating a model to predict the number of wins a college basketball team will predict. As we saw, there were a number of outliers discovered.

The data was divided into a training a testing dataset. The models were created using the training data, and then the models get tested on the testing dataset. When we calculated the projected win amount for the testing dataset, it will be compared to the actual win amount for that set. The adjusted R^2 values of these models will be calculated as well as the Mean Squared Error for the models on the testing dataset.

Model	Adj R	MSE
Model 1	0.2794	29.5838
Model 2	0.2806	29.4909
Model 3	0.283	29.3778
Model 4	0.5127	20.103

Table 2: Evaluation Metrics for the four models.

As seen in Table 2, the full linear model (Model 1), reduced linear model (Model 2), and generalized additive model

(Model 3) all have about the same adjusted R^2 values and mean squared errors. Therefore, between these three models, one isn't necessarily better than another. However Model 4, the Pythagorean Theorem model, has a much greater adjusted R^2 value and a much lower mean squared error. This leads us to believe that this model is significantly better than the other three models. Unfortunately, the adjusted R^2 value for this model is only 0.5127, suggesting that even this model isn't great for predicting the total number of wins for a team's season. The error rate for this is about 11%, meaning the model is off by an average of around 3 wins for a season.

DISCUSSION

Originally, this study was going to use the same data as well as other data to determine the best college basketball teams of the past five years. However, after further evaluation and discussion with others, it was decided that determining the best college basketball teams wasn't the best way to look at the data. It would be difficult as there's a lot of data and information that goes into that conclusion as well as it is subjective to determine what data is important and should be used in that case. Therefore, the study uses most of the same data, but has been shifted to attempting to predict the number of wins a college basketball team will get in a season. This way the data deemed important will be determined by statistics.

The Pythagorean Theorem model in sports was originally used in professional baseball to predict the number of wins for a team using the total runs scored and allowed for that same season. As we saw in this study, the Pythagorean Theorem can also be used in college basketball and to use the total points scored and allowed in a season to predict the total number of wins for the next season, and it was shown that this was the best model. However, with an error rate of 11% and an adjusted R^2 value of 0.5127, it's not the best at predicting. This just further proves how hard it is to predict what's going to happen in the sports world.

The research and modeling doesn't have to stop here though. There are a number of other variables out there that were not looked at in this study that could also be used in models to predict the number of wins. Variables such as assists, rebounds, blocks, and many more. You could also look at the distribution of players on a team, like how many freshmen, sophomores, etc are on the team and/or the minutes that they play. There are also other 'Win' variables (conference wins, home wins, away wins) that can be used in models either as dependent variables or predictors. There is still a lot more research that can be done that could possibly yield a better model for predicting the number of season wins.

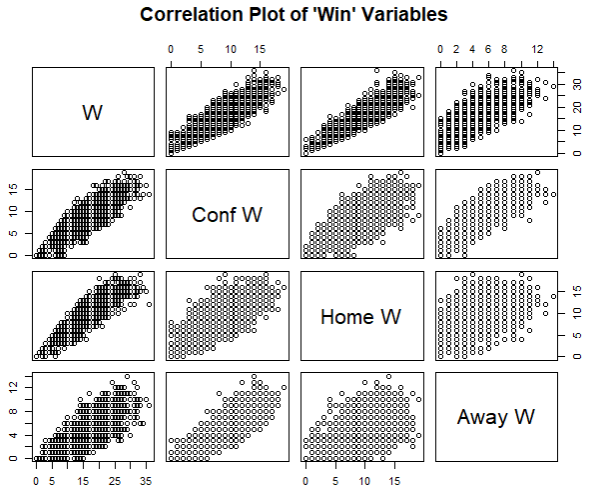

Figure 8: Correlation plot of the different 'Win' variables.

In Figure 8, it can be inferred that away wins ('Away W') aren't as strongly correlated with total wins ('W') as conference wins ('Conf W') and home wins ('Home W'). Concluding that away wins aren't as important when it comes to total wins for the season.

CONCLUSION

This study gathered college basketball stats from the past five years, 2017-2021, and performed an in-depth outlier analysis to identify any anomalies that may influence the data in a negative way. A number of outliers were found and were either removed or altered slightly in order to make the results of the study more accurate.

Four different models were constructed to attempt to predict the number of wins for a team's season using the previous year's stats. These models were then tested and evaluation metrics were calculated to compare the different models. It was found that the Pythagorean Theorem Model, Model 4, produced the most accurate predictions compared to the other three models, with an error rate of about 11%.

Model 4: Pythagorean Theorem Model

$$Wins = Games * \frac{TmPts^5}{TmPts^5 + OppPts^5}$$

More research and work can be done to create more models, using either the same data or new data, to attempt to find a new model that better predicts the number of wins for a team's season with a lower error rate.

REFERENCES

[1] Wayne L. Winston. 2009. *Mathletics.* Princeton University Press, Princeton, NJ. .3-10