Prediction of Ranking Teams in Women's Volleyball

Kyle Vander Meulen
California Baptist University
Email: Kyle.VanderMeulen@calbaptist.edu

Abstract

Volleyball in the United States over the past decade has grown in popularity with participants of all ages both male and female playing at different levels. Especially, volleyball has experienced the most success in terms of interest and participation at the women's collegiate level. The goal of this study is to determine which factors are significant in winning percentage calculation by using multiple linear regression and logit regression. By using the 334 teams during the 2017 season, this study provides the prediction of winning percentages for obtaining positive results in NCAA division I women's volleyball. The results from the multiple linear regression show that winning percentage can be explained not only by my team's performance (Kills = 0.06 [0.001, 0.11], Errors = -0.06 [-0.008, -0.04], Aces= 0.04 [-0.00, 0.07]), but also opponent's performance (Opp Kills = -0.06 [-0.009, -0.04], Opp Errors = 0.12 [0.008, 0.152]. The odds of significant factors are calculated in the logistic model (Kills odds = 1.03 [1.02, 1.04], Errors odds = 0.98 [0.97, 0.99], Opp Kills Odds = 0.97 [0.97, 0.99], Opp Errors odds = 1.03 [1.02, 1.05]). Based off of these results coaches and players at the division I level should not only consider focusing on the improvement of their own performance skills, but should take into account limiting and decreasing opponent's performance.

1. Introduction

The game of volleyball has evolved exponentially over the years in a way in which coaches are more open to embracing analytical practices to improve their team. Very few coaches and players have efficiently improved their performance by identifying the highest correlated skills in regards to winning and allocating the proper amount of time to training. On the other hand, more often then not coaches will focus on analyzing the six fundamental skills which are passing, serving, digging, setting, blocking, and hitting, but the problem with that is the subjectivity bias in the collection of the data. That is why rating systems need to developed and put into place to objectively and accurately make team/player predictions in volleyball [1]. Moreover, there is a deeper issue: there is no widely recognized, accurate, and accepted probabilistic rating system in the game of volleyball. There are general probabilistic rating systems such as the ELO, Glicko, Glicko-2, Stephenson, and Bradley-Terry, but aren't specific enough to accurately predict volleyball games [4]. Furthermore, analysis done on the Spanish Superliga, Turkish Men's and Women's League, Men's Volleyball World Championship, and NCAA Women's Volleyball all have shed light and had some success in being able to filter out the key prediction factors in their leagues [2], [3], [5], and [6]. Pena, Rogriguez-Guerra, Busca, and Serra did an analysis of the Spanish Superliga during the 2010-2011 season. They were able to uncover the results that team category, points obtained in the point-break phase, reception errors, and blocked attacks were significant factors in Spanish Superliga league [2]. Along the same lines we can see that in the study done by Silva, Lacerda, and Joao that in the 2010 Men's Volleyball World Championship that the deciding factors were self-inflicted errors or favoring positive factors for the team being analyzed [6]. Contrastingly, two other scholarly articles took a look at women's volleyball. Firstly, Akarcesme utilized a logistic model to be able to explain Turkish women's volleyball and an efficacy model to examine performance in terms of position [3]. Secondly, Estabrook observed the relationship between NCAA stat categories and the success of women's collegiate volleyball teams. She inspected these box scores in depth by using a OLS model. An aspect of her work that was fascinating was that she compared all three divisions of NCAA women's volleyball. Interestingly, there were differences in the most important factors of all three

divisions. A team's blocking was important for Division I and Division II, but serving was more critical to Division III teams. Hitting Percentage was the most influential stat regardless of division [5]. Unlike this literature which focus solely on one team's statistics, this paper will delve into the impact of opponent's skills.

The first objective of this paper is to answer the following question: Can predictive performance be improved further if we take into account opponent statistics like Opp Kills, Opp Errors, Opp Attacks, and Opp Pct as predictor variables, along with some combination of Kills Errors, and Aces? The new models we created utilizing these statistics outperform previous models in relation to correlation to winPct, mean squared error, and predicting a team's actual winning percentage.

The results of our models can be interpreted as the expected win percentage that a team will achieve based off of their statistics and opponent's statistics during the season. This winPct stat can be helpful in ranking teams as well as assisting a coach in being able to decipher where each team will most likely land at the end of the season.

The second objective of this paper is the comparison of my two models in relation to which is the most accurate representation of a NCAA women volleyball team's performance. In the first part of this study inferential analysis was performed as we constructed a multiple linear regression function using winPct as the dependent variable. This regression was able to sift out the statistically significant factors in a team's win percentage. Secondly, analysis was executed using a logistic regression with winPct. high as the dependent variable where the winPct.high variable was created in order to discover the odds ratio for each individual team. With the data our logistic regression could estimate the probability (0 to 1) of winning a match in the NCAA women's volleyball division for the 2017-2018 season using the constants and model coefficients. The combination of OLS and logit regression provide a useful means in being able to analyze significant factors and team performance.

The rest of the paper is outlined in this manner. Section 2 outlines an in depth description of our data, variable list, and data source. In Section 3 we clearly highlight our two models and explain our rationale for our model selection. Lastly, in Section 4 we include our thoughts on future work and our conclusion.

2. Data Description

We used both ordinary least squares and logistic regression with sets, kills, errors, total attacks, hitting pct, assists, digs, aces, block solos, block assists, and total blocks, along with the opposing team's statistics over the course of the 2017-2018 season. Below is a list of all statistics we considered:

Sets (S) The amount of games a team played over the course of the season

Kills Attacks by a player that directly leads to a point

Errors An attack that directly results in a point for the opposing team

Total Attacks The sum of a team's attacks over the season

Hitting Percentage (Pct) Hitting % = (Kills – Errors) / Total Attacks

Assists When a player passes, sets, or digs the ball to a teammate who attacks the ball for a kill

Digs A player passes that ball that has been attacked by the opposition

Aces A serve that directly results in a point

Solo Blocks A single player blocks the ball into the opponent's court leading to a point

Block Assists Awarded when multiple players are involved in blocking the ball into the opponent's court leading to a point

Total Blocks (TB) The sum of a team's blocks over the season

Opponent Kills Attacks by an opposing player that directly leads to a point

Opponent Errors An attack by the opposition that directly results in a point for the other team

Opponent Attacks The sum of an opposing team's attacks over the season

Opponent Hitting Pct Opponent Hitting % = (Kills – Errors) / Total Attacks

We used data over the full course of the 2017-2018 NCAA Division I Women's Volleyball season. The data was obtained from the NCAA statistics database and was compiled into an Excel spreadsheet. The analysis was done in Rstudio 3.5.1 by using these specific packages: corrplot, leaps, and mosaic. We desired to get an accurate depiction of the most important factors in Division I by including the full scope of the league. All teams that were recognized as

Division I were included in our analysis. The total observation count is 334 women's volleyball teams. The only issue with the current data was adding an additional calculation variable winPct as Wins and Losses were not an accurate depiction of a team's overall winning percentage. Creating this dependent variable winPct was necessary for the logistic model prediction calculation of the team's forecasted win percentage for the season. Other than that issue, the data was collected by a reliable source in the NCAA and was easily available for analysis.

We performed a variety of different analytical tests to distinguish what variables should be included in our two models. First off, to gain a better sense of which of our variables have an effect on a team's winning percentage we performed a correlation plot.

The correlations between winning percentage and all the other independent variables taken into consideration are depicted in Figure 1.

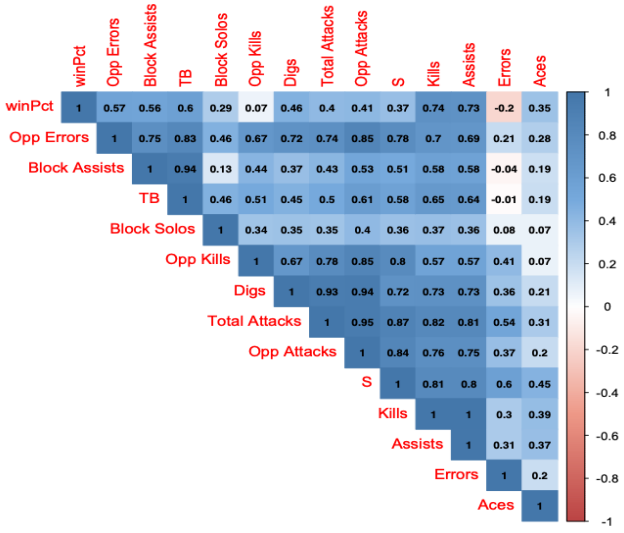

Figure 1: Correlation between winning percentage and Opp. Errors, Block Assists, TB, Block Solos, Opp. Kills, Digs, Total Attacks, Opp. Attacks, S, Kills, Assists, Errors, and Aces. Our scale's correlation range is from -1 to 1 utilizing the darkness of color to indicate a stronger positive or negative correlation.

Note the correlations for all of our factors have a moderate to high correlation, except for two variables. The two variables that have no impact on win percentage are as follows: an opposing team's number of kills and the total amount of errors committed by our team. The highest positively correlated factors are a team's number of kills and assists which can explain a majority of a team's ability to win. Despite these factors having a great influence there are other variables to consider in being able to grasp the full picture of a team's ability to win. All variables were considered in this correlation because of the fact that there is no defined win percentage calculation in the sport of volleyball.

Moreover, additional analysis needed to be made in order to determine the best model to accurately predict a team's winning percentage. In order to do so we used the regsubsets function to perform model selection using the criteria BIC.

Illustrated below in Figure 2 is a table of models showing which variables are in each model. The models are ordered by the specified model selection statistic.

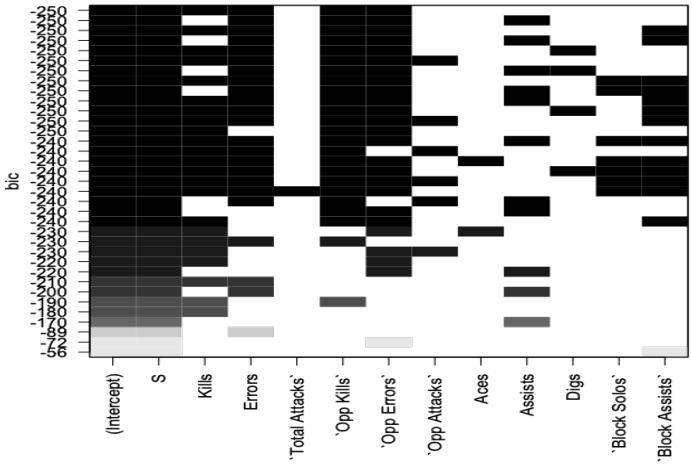

Figure 2: Table of various models organized in terms of the model selection criteria BIC.

We chose to include all of the predictor variables to be included in our multiple linear regression model as well as in our logistic model. To justify our reasoning that all variables should be

included is two discovered points. First, as mentioned above there is no well-known and defined winning percentage calculation in the game of volleyball. Secondly, our model selection plot highlights the fact that for a large majority of our models the BIC value stayed the same. So there is no difference whether you include or exclude a variable(s) within your model.

Furthermore, we went on to set up our linear model based on the previous analytical tests results. We included all variables of interest within our linear model to get a full scope of the predicted win percentage. So our results indicate the significant factors in relation to win percentage.

Below in Figure 3 is the summary of our linear model and the significant factors.

	Coefficients	2.5%	97.5%	SE
(Intercept)	35.09***	{27.39	42.79}	3.91
S	-0.29+	{-0.59	0.00}	0.15
Kills	0.06**	{0.02	0.11}	0.02
Errors	-0.06***	{-0.08	-0.04}	0.01
Total Attacks	-0.01	{-0.02	0.00}	0.00
Opp Kills	-0.06***	{-0.09	-0.04}	0.01
Opp Errors	0.12***	{0.08	0.15}	0.02
Opp Attacks	0.00	{-0.03	0.02}	0.01
Aces	0.04+	{0.00	0.07}	0.02
Assists	0.03	{-0.02	0.07}	0.02
Digs	0.01	{-0.01	0.04}	0.01
Block Solos	-0.02	{-0.07	0.03}	0.03
Block Assists	-0.01	{-0.03	0.00}	0.01

Figure 3: Results from the linear model that highlights the 95% confidence interval, standard error, and coefficients for each individual variable. The coefficients highlighted in red are the significant factors in relation to Win Percentage.

In reference to the results we can see the negative significant factors as such, the amount of sets a team played over the course of the season, errors committed, and opponent kills. The more sets a team played over the course of a season the more fatigued the team as a whole and individual players, which in turn decreased a team's overall win percentage. The total errors a team commits obviously decrease a team's win percentage. Lastly, the number of opponent kills will result in opponent's points which will decrease your team's odds of winning. Something interesting to see was the opposite effect of kills and errors, which in turn will cancel each other out. Also, aces were positively significant in increasing a team's win percentage and the most positive significant factor was the amount of errors the opposing team committed. This factor being the highest positive coefficient was unexpected.

Below in Figure 4 is the summary of our logit model and the significant factors.

	Coefficients	2.5%	97.5%	SE
(Intercept)	0.03	{0.00	5.57}	2.812
S	0.77*	{0.60	0.97}	0.122
Kills	1.03*	{1.00	1.07}	0.016
Errors	0.98*	{0.97	1.00}	0.007
Total Attacks	1.00	{0.99	1.00}	0.003
Opp Kills	0.99	{0.97	1.00}	0.009
Opp Errors	1.05**	{1.02	1.08}	0.014
Opp Attacks	1.00	{0.98	1.02}	0.008
Aces	1.02	{0.99	1.04}	0.012
Assists	1.00	{0.97	1.03}	0.015
Digs	1.00	{0.98	1.02}	0.009
Block Solos	0.99	{0.96	1.02}	0.016
Block Assists	0.99	{0.98	1.00}	0.006

Figure 4: Results from the logit model that highlights the 95% confidence interval, standard error, and coefficients for each individual variable. The coefficients highlighted in red are the significant factors in relation to Win Percentage.

From the table we can see four significant predictors of win percentage. Once again we see the amount of sets played was negatively significant. For every set played it decreases the odds of

win percentage by 23%. Kills had a 3% positive effect and Errors had 2% negative effect on a team's win percentage. Lastly, once again opponent errors were significant by increasing the log-odds likelihood of win percentage by 5%. Logistically, our model identified those key predictive factors that reaffirmed what our linear model revealed through its analysis.

3. Results

Corresponding to the data analysis we were able to filter out the key significant factors from our beginning variables of interest list. The most significant variable that was found through the analysis was the total count of the errors an opponent committed. The visualization is highlighted below.

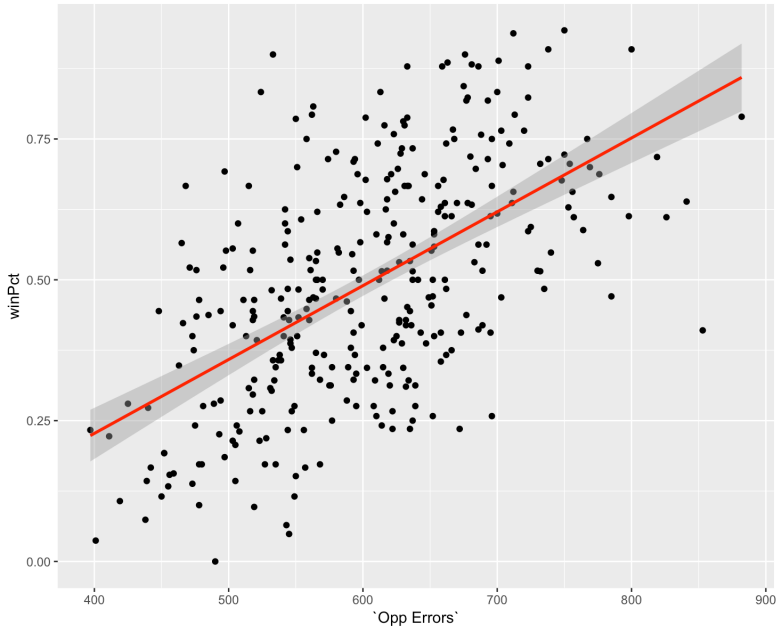

Figure 5: Scatterplot with a regression line that includes the relationship between win percentage and the amount of errors an opponent team committed.

The plot highlights the strong positive correlation between an accumulation of opponent's errors and the winning percentage. It can be concluded that taking into account opponent's individual stats is key in being able to combat and limit their performance, which in turn will increase your team's winning percentage.

Our goal from the onset was being able to create and define a winning percentage calculation to utilize in volleyball. We were able to do so in respect to all of significant analysis and results that were completed. To reaffirm our models even more we computed predicted winning percentages for the top 10 women volleyball teams of the 2017 season.

Actual Ranking	Top 10	Fitted Values	Win Pct.
1	Penn St.	98.03	94.29
13	Wichita St.	95.25	87.9
8	Western Ky.	92.72	88.57
7	Nebraska	92.23	88.9
9	Stanford	90.58	88.24
2	Florida	89.75	93.75
37	UMES	88.93	75.00
10	Colorado St.	85.92	87.87
6	Texas	85.73	90.00
12	North Texas	85.21	87.87

Figure 6: This table depicts the top 10 teams in terms of our predicted winning percentage. In column 1 these are the actual rankings of where the team ended up in relation to their winning percentage. Column two mentions the college team names, while column three highlights our

predicted winning percentages from number 1 to number 10. The last column shows the actual winning percentages the teams finished up with at the end of the season.

As you can see we were accurate in being able to predict the top team in Penn State. All of the other fitted values were fairly close in relation to where a team finished up in the ranking. The one outlier in this top ten that needs an explanation of why they underperformed in a huge way is UMES. We had them finishing up 7th in the top 10 teams, but they actually finished 37th. Something to note is that we can never fully be able to predict the exact winning percentage of every team, but with the variables we included in our models we were able to predict a team's winning percentage at a fairly high rate.

Future Work and Conclusion

We noted after our analysis that significant factors for a team are dependent upon league, gender, and age. With this knowledge in mind our future work will revolve around whether we stick with Division I Women's Volleyball. One idea for future work is to use the widely known Data Volley program to analyze a smaller scope of teams and use self-data collection. We would be able to obtain more in depth results and statistics which potentially could have a better correlation to win percentage. Unlike our analysis which was a broader analysis of team's volleyball statistics, future work might take into account a team's sideout percentage, player position, and rotation. Sideout percentage is essential for a team that is attempting to get out of serve receive by scoring a point and getting the ball back to serve. Player position may be significant in the grand scheme of increasing a team's winning percentage. An outside hitter may be more influential and valuable than a libero, but they are both needed for a team to win. This information could be used in recruiting processes as coaches may expend more effort towards a great outside hitter over an opposite hitter. Lastly, analyzing a team's ability to produce quality stats may potentially be different for the six rotations. Most teams have that one rotation that is a struggle to get out of and these results additionally may be used to focus more time on that specific rotation.

The results of our models show that winning percentage can be explained by not only by my team's performance (Kills, Errors, Ace) but also opponent's performance (Opp Kills, Opp Errors). We believe that teams tend to overlook working on reducing the opposing team's statistics and focus solely on self-improvement. Based on our results coaches and players in NCAA Division I Women's Volleyball should allocate more practice time to honing these significant skills. With that being done teams can be most efficient over the course of the season in getting better and maximizing their odds of winning volleyball games.

References

[1] C. Bagley and B. Ware, "Bump, Set, Spike: Using Analytics to Rate Volleyball Teams and Players," March 2017. http://www.sloansportsconference.com/content/bump-set-spike-using-analytics-rate-volleyball-teams-players/, Accessed 03-27-2019

[2] J. Pena, J. Rodriguez-Guerra, B. Bernat, and N. Serra, "Which Skills and Factors Better Predict Winning and Losing in High-Level Men's Volleyball?" September 2013. https://journals.lww.com/nsca-jscr/fulltext/2013/09000/Which_Skills_and_Factors_Better_Predict_Winning.17.aspx, Accessed 03-27-2019.

[3] C. Akarcesme, "Is it Possible to Estimate Match Result in Volleyball: A New Prediction Model," May 2017. https://www.researchgate.net/publication/327545035_Is_it_Possible_to_Estimate_Match_Result_in_Volleyball_A_new_Prediction_Model , Accessed 03-27-2019.

[4] M. Glickman, "A Comparison of Rating Systems for Competitive Women's Beach Volleyball," March 2014. http://glicko.net/research/volleyball-FINAL.pdf, Accessed 03-27-2019.

[5] N. Estabrook, "The Relationship between NCAA Volleyball Statistics and Team Performance in Women's Intercollegiate Volleyball," August 1996. https://digitalcommons.brockport.edu/cgi/viewcontent.cgi?article=1058&context=pes_theses
Accessed 03-27-2019.

[6] M. Silva, D. Lacerda, and P. Joao, "Game-Related Volleyball Skills that Influence Victory," July 2014. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120451/, Accessed 03-27-2019.