DSO 578 Fundamentals of Sports Performance Analytics
Group Project Report
Professor Lorena Martin
November 29, 2022.

Would the Ace and Block be the Leading Indicators to Increase the Propensity of Winning in the AVP Women's Beach Volleyball?

Yuting Jiao

Yi-Ching (Millie) Lin

Shih-Ting (Vicky) Liu

Xinyi Li

Keying Ji

1. Abstract

The purpose of this study was to explore the relationship between performance indicators and the propensity of winning. A total of 5 years and 1174 records of players' performances were utilized to conduct analysis. The performance indicators analyzed were age, height, attack, kill, total error, ace, serve error, block, dig, and average hitting percentage. Exploratory Data Analysis and Logistic Regression were used to compare the indicators and determine which one(s) contributed significantly to winning in matches. The result shows that total aces and total blocks are the leading indicators to increase the propensity of winning in the AVP women's beach volleyball. This report aims to help young athletes who want to enter AVP and start their professional journey, by giving recommendations and practicable advice on how they should set their practice goals and what to improve.

Keywords: Beach Volleyball, performance indicators, statistical analysis

2. Introduction and background

Volleyball is one of the most popular sports in the world. Beach volleyball (BV)—usually played, as its name implies, on a sand court with two players per team—was introduced in California in 1930. The first official beach volleyball tournament was held in 1948 at Will Rogers State Beach, in Santa Monica, California. It is a team sport in which two teams oppose one another divided by a net. The sequence of actions in beach volleyball is: serve, serve reception, set, attack, block, and dig (Giatsis and Zahariadis, 2008). The objective of the game is to send the ball over the net and ground it on the opponent's side of the court.

The AVP, The Association of Volleyball Professionals is the biggest and longest-running professional beach volleyball tour in the United States where hundreds of professional players are competing for the championship.

Beyond the reason that all our teammates are super interested in playing or watching beach volleyball games, we choose BV for reason that we feel honored connected to this sport since Southern California wins the 2022 NC beach volleyball championship and many of our USC Alumni as well as professional BV players, for example, Kelly Claes, Maddison McKibbin and his brother Riley McKibbin have shown their excellence in AVP games. College BV Players are likely to start playing professionally immediately after graduating, and AVP will be their first landmark.

Therefore, based on the data source we can find, for this project, we focused on analyzing the leading performance indicators at AVP tour games from 2018 to 2022, such as Age and Height of players, Attacks, Kills, Errors, and Blocks. Looked for any relationship between those indicators and the propensity of winning.

Research in Volleyball examined team performance (Baacke, 1988; Byra and Scott, 1982; Rose, 1983). George & Panagiotis (2008) analyzed the technical skills and key performance indicators: serve, attack, block and dig, proving certain skills contributed significantly to winning matches. However, regarding the player's gender and the different rules of FIVB (Fédération Internationale de Volleyball) and AVP, we need to develop our own hypothesis and there is a big chance we will get a different analysis outcome.

Thus, based on the support of previous research and literature as above, in this study, we set the null hypothesis as total aces and total blocks do not have an impact on the propensity of winning in the AVP women's beach volleyball and the alternative hypothesis as total aces and total blocks are the leading indicators to increase the propensity of winning in the AVP women's beach volleyball.

We hope that through this study, we can find a good pattern to indicate if total aces and total blocks are the most important factors in order to win an AVP game, thus we can report to and empower our USC BV athletes to get purposeful preparation and targeted training if they want to achieve sustained competitive excellence and well-being in their professional journey.

3. Data Source and Data Manipulation

The origin datasets contain men's and women's beach volleyball match statistics from the AVP tours from the year 2002 to 2022. The dataset includes players' physical measurements (e.g., Age, Weight, and Height), players' performances in each match (e.g., Attacks, Errors, and Kills) and matches information (e.g., Tournament, Bracket, Match Number).

Before diving into the data and running the analysis directly, we first conducted data cleaning using Python to ensure the data quality is good for further analysis. First, we decided to only include the recent five-year data (2018-2022) since data from many years ago might not be able to provide significant insights due to the fast-changing environment in the sports industry. We also only kept match data whose gender is a woman and removed unimportant fields. Secondly, we separated the statistics of players in the winning team and losing team for the same match into two rows, then added one label column where 1 is winning and 0 is losing. Lastly, we removed rows that contain missing values.

4. Variable Identification

After getting a cleaned dataset, we created 10 numerical variables shown below that we will use for our analysis:

- mean_age: The average age of two players on the same team.
- mean_hgt: The average height of two players on the same team.
- sum_tot_attacks: The total attacking swings over the net of two players on the same team.

- sum_tot_kills: The total point-ending attacks of two players in the same team.
- sum_tot_errors: The total mistakes of two players in the same team.
- sum_tot_aces: The total point ending serves of two players in the same team.
- sum_tot_serve_errors: The total mistakes made on the serve of two players in the same team.
- sum_tot_blocks: The total point-ending blocks of two players in the same team.
- sum_tot_digs: The total successfully defending an attack of two players in the same team. (i.e. running down an opponent's shot and passing to your partner)
- mean_tot_hitpct: The average hitting percentage of two players in the same team. It signifies the player's effectiveness at scoring points. It is calculated by (Kills – Errors) / Attacks.

The table below shows the summary of variables:

Variable Name	Min	1st Qu	Median	Mean	3rd Qu	Max	Stdev
mean_age	16.74	25.86	28.61	28.69	30.60	39.03	3.59
mean_hgt	66.00	71.00	72.00	71.74	73.00	75.00	1.58
sum_tot_attacks	20.00	46.00	53.00	55.29	64.00	102.00	13.45
sum_tot_kills	7.00	24.00	28.00	28.68	33.00	55.00	5.50
sum_tot_errors	0.00	4.00	6.00	6.68	8.00	21.00	3.22
sum_tot_aces	0.00	1.00	3.00	2.84	4.00	12.00	2.04
sum_tot_serve_errors	0.00	3.00	4.00	4.59	6.00	17.00	2.44
sum_tot_blocks	0.00	1.00	200	2.24	3.00	10.00	1.67
sum_tot_digs	2.00	12.00	17.00	17.22	21.00	43.00	6.63
mean_tot_hitpct	-0.0555	0.3220	0.4143	0.4099	0.4965	0.8725	0.1315

Table 1: Variable Summary

5. Exploratory Data Analysis

After preparing the data, we then conducted the exploratory data analysis to more understand our dataset. As the figure shown below, our data mostly came from 2018 and 2019. The data contained both the winner's bracket and the contender's bracket. The top 3 tournaments were in Manhattan Beach, Chicago, and Hermosa Beach.

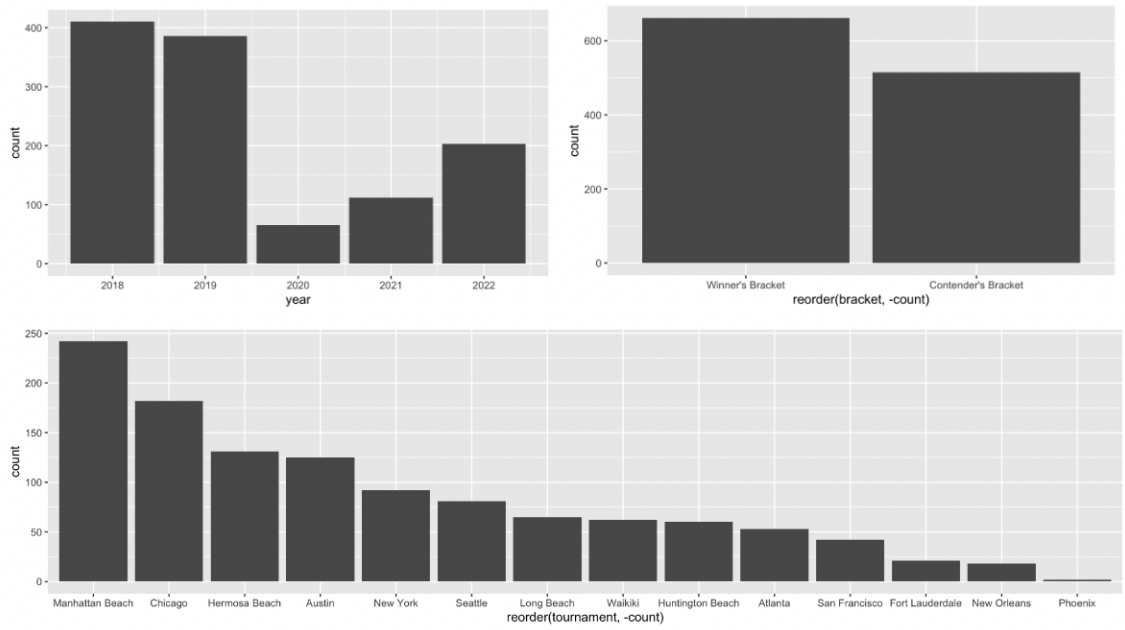


Figure 1: Data distribution by year, bracket, and tournament

Below were the distributions of each feature we would use in the model. We could tell that most of the players were 25-30 years old with a height of 7”0 to 7”2. The average sum of attacks and digs was 55.3 and 17.2. The average mean hit percentage was 41.0%. Regarding the errors, the average sum of errors and sum of serve errors were 6.7 and 4.6. The average kills, aces, and blocks at point ending were 28.7, 2.8, and 2.2.

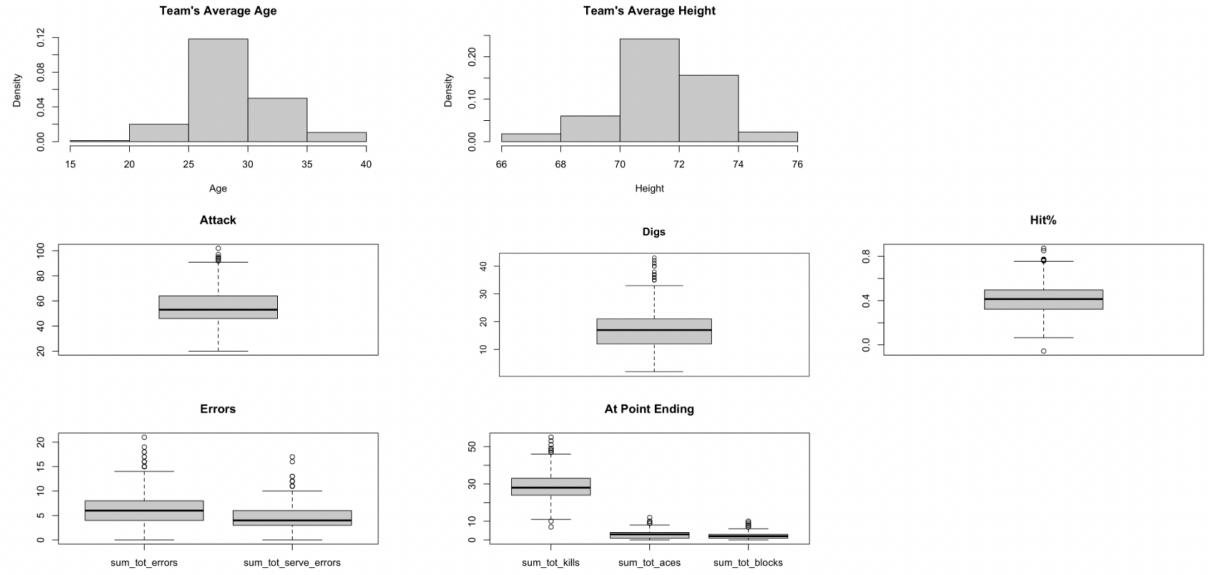


Figure 2: Data distribution for all features

Since we would like to investigate the leading indicators for winning the game, we also looked into the distribution of the features breaking down by whether the team wins or not. From the graph below, we noticed that the number of kills, blocks, aces, errors, and the percentage of the hit were differently distributed in the winning and losing teams.

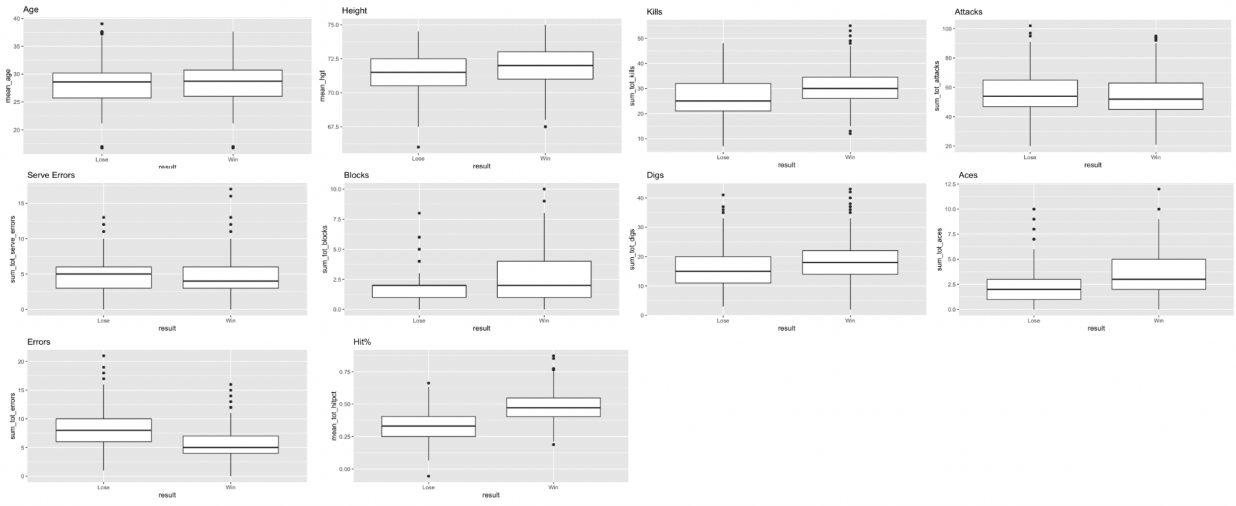
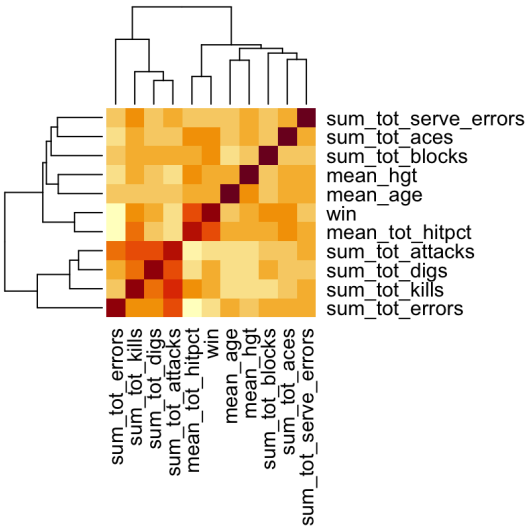


Figure 3: Data distribution for all features broken down by Win

To further validate the pattern from the boxplot above, we ran the correlation analysis to see what feature had a higher correlation to the result of the game. According to the heatmap of the correlation, we realized the hit percentage played a prominent role in affecting whether the team would win or lose. The linear correlation between win/lose and average hit percentage was 57.2%, followed by the sum of aces, which was 30.1%. And the correlation between win/loss and the sum of errors was -40.2%. Therefore, the result of the game had a relatively strong positive correlation with the total number of aces and average hitting percentage while it had a relatively strong negative correlation with the total number of errors.



Feature Name	Correlation to Win
mean_age	0.0354696
mean_hgt	0.1258047
sum_tot_attacks	-0.07855409
sum_tot_kills	0.2944179
sum_tot_errors	**-0.4017174**
sum_tot_aces	**0.3007221**
sum_tot_serve_errors	-0.01759939
sum_tot_blocks	0.2868204
sum_tot_digs	0.1869515
mean_tot_hitpct	**0.5723484**

(Left)Figure 4: Correlation Heatmap | (right) Table 2: Correlation to Win

6. Logistic Regression

After data cleaning and exploratory data analysis, we had the cleaned dataset ready for our modeling. Since our targeted variable is if the team wins the game or not, we decided to use logistic regression to find the leading variables that could affect the outcome. Given that Women's AVP is teamwork-based, we used average age, average height, total attacks, total kills, total errors, total aces, total serve errors, total blocks, total digs and average hit% of a team as our predictor variables.

We first split the dataset into two parts: the training set and the test set. We used the training set to train the model and used the test set for evaluation. 70% of the dataset belonged to the training set.

The model summary is as below:

```
Call:
glm(formula = win ~ mean_age + mean_hgt + sum_tot_attacks + sum_tot_kills +
    sum_tot_errors + sum_tot_aces + sum_tot_serve_errors + sum_tot_blocks +
    sum_tot_digs + mean_tot_hitpct, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0607  -0.4492   0.1341   0.5211   3.3564

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.41331    5.25410  -0.650 0.515920
mean_age              0.03147    0.03133   1.005 0.315138
mean_hgt             -0.01524    0.07039  -0.217 0.828553
sum_tot_attacks      -0.16339    0.02901  -5.633 1.77e-08 ***
sum_tot_kills         0.22780    0.05988   3.804 0.000142 ***
sum_tot_errors       -0.14370    0.06363  -2.258 0.023925 *
sum_tot_aces          0.38524    0.05834   6.603 4.02e-11 ***
sum_tot_serve_errors -0.17506    0.04498  -3.892 9.95e-05 ***
sum_tot_blocks        0.50890    0.07104   7.163 7.88e-13 ***
sum_tot_digs          0.20435    0.02510   8.142 3.90e-16 ***
mean_tot_hitpct       6.15621    2.85208   2.158 0.030889 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1131.28  on 822  degrees of freedom
Residual deviance:  576.16  on 812  degrees of freedom
AIC: 598.16

Number of Fisher Scoring iterations: 6
```

Figure 5: Logistic Regression Model 1

From the result, average age and average height are not statistically significant. We then removed these two variables for better model performance.

```
Call:
glm(formula = win ~ sum_tot_attacks + sum_tot_kills + sum_tot_errors +
    sum_tot_aces + sum_tot_serve_errors + sum_tot_blocks + sum_tot_digs +
    mean_tot_hitpct, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0589  -0.4436   0.1339   0.5162   3.3258

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.54168    1.22600  -2.889 0.003867 **
sum_tot_attacks      -0.16585    0.02878  -5.763 8.25e-09 ***
sum_tot_kills         0.23157    0.05963   3.884 0.000103 ***
sum_tot_errors       -0.13990    0.06332  -2.209 0.027151 *
sum_tot_aces          0.38687    0.05788   6.684 2.33e-11 ***
sum_tot_serve_errors -0.17741    0.04453  -3.984 6.77e-05 ***
sum_tot_blocks        0.49830    0.06981   7.138 9.46e-13 ***
sum_tot_digs          0.20395    0.02508   8.131 4.26e-16 ***
mean_tot_hitpct       6.10536    2.85524   2.138 0.032492 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1131.28  on 822  degrees of freedom
Residual deviance:  577.18  on 814  degrees of freedom
AIC: 595.18

Number of Fisher Scoring iterations: 6
```

Figure 6: Logistic Regression Model 2

Based on Model 2, all variables are statistically significant now using a significance level of 95%.

To evaluate the performance of our model, we created predictions for the test set. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions is summarized with count values and broken down by each class. The confusion matrix for the test set is shown below:

Actual/Predict	**Win**	**Lose**
Win	124	34
Lose	23	172

Table 3: Confusion Matrix

We computed the accuracy for the model, which is 0.836. It meant that our model achieved a good performance. It can predict the outcome with an accuracy greater than 80%.

Then, we plotted the ROC-AUC curve for visualization. AUC - ROC curve is a performance measurement for classification problems at various threshold settings. The Higher the AUC, the better the model is at distinguishing between two classes. AUC is 0.8334 for our model, which is significantly high.

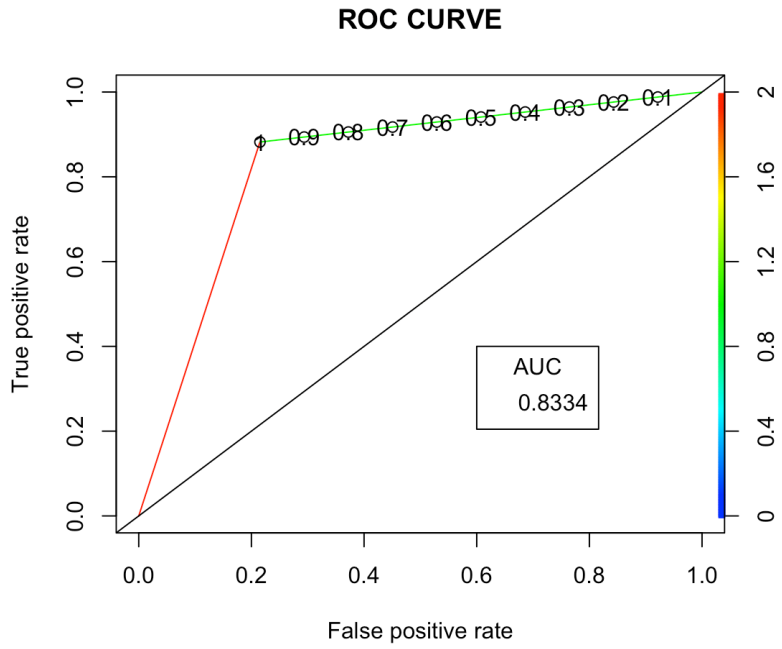


Figure 7: ROC-AUC Curve

After validating our model through the test set that it can perform well in predicting the outcome, we used all the data to obtain our final model. The result is as below:

```
Call:
glm(formula = win ~ sum_tot_attacks + sum_tot_kills + sum_tot_errors +
    sum_tot_aces + sum_tot_serve_errors + sum_tot_blocks + sum_tot_digs +
    mean_tot_hitpct, family = "binomial", data = bv)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0395  -0.4217   0.1366   0.5096   3.3039

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.53582    1.02476  -3.450  0.00056 ***
sum_tot_attacks      -0.17365    0.02393  -7.258 3.94e-13 ***
sum_tot_kills         0.26796    0.04913   5.454 4.93e-08 ***
sum_tot_errors       -0.16103    0.05343  -3.014  0.00258 **
sum_tot_aces          0.40894    0.04862   8.411  < 2e-16 ***
sum_tot_serve_errors -0.17891    0.03812  -4.693 2.69e-06 ***
sum_tot_blocks        0.49736    0.05893   8.439  < 2e-16 ***
sum_tot_digs          0.20451    0.02065   9.902  < 2e-16 ***
mean_tot_hitpct       4.67601    2.34123   1.997  0.04580 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1616.8  on 1175  degrees of freedom
Residual deviance:  817.1  on 1167  degrees of freedom
AIC: 835.1

Number of Fisher Scoring iterations: 6
```

Figure 8: Logistic Regression Final Model

The final logistic regression equation is:
log ((probability of win)/(1-probability of win)) = -3.53582 - 0.17365*total_attacks + 0.26796*total_kills - 0.16103*total_errors + 0.40894*total_aces - 0.17891*total_serve*errors + 0.49736*total_blocks + 0.20451*total_digs + 4.67601*average_hitpct

Based on the absolute coefficient value, we found that total aces and total blocks play the most important roles in leading to the success of a game. Although the average hit percent has the highest coefficient value, we should convert it to a percentage base before making a comparison. After conversion, the coefficient value of it becomes 0.04676.

Therefore, we can reject the null hypothesis and conclude that improving total aces and total blocks will bring higher chances of winning. One unit increase in total aces is associated with an increase of 50.5% in the odds of winning. And one unit increase in total blocks is associated with an increase of 64.4% in the odds of winning.

7. Conclusion and Recommendation

Based on our exploratory data analysis, correlation analysis, and logistic regression model, here are the conclusions we draw:

- The result of the game has a relatively strong positive correlation with the total number of aces and average hitting percentage
- The result of the game has a relatively strong negative correlation with the total number of errors
- Age and height do not significantly affect the propensity of Winning in the AVP Women's Beach Volleyball
- Aces and blocks significantly affect the propensity of Winning in the AVP Women's Beach Volleyball

Based on our conclusions, here are some of the recommendations we can make for athletes. Since aces and blocks are the important indicators to determine the result of the competitions, for victory, athletes must focus on exercising those areas and pay attention to get maximized group performance results on these two factors. Then, the aim question transits to which factors will contribute to good aces and blocks. According to João, P. V., Medeiros, A., Ortigão, H., Lee, M., & Mota, M. P. (2021), high jump performance and high velocity will enhance the blocking performance. Therefore, we suggest that athletes can do some bodyweight exercises as well as Plyometric Training to improve jumping ability and speed. For aces, Denardi, R.A., Clavijo, F.A.R., Oliveira, T.A.C, Silva, S.L., Travassos, B. & Corrêa, U.C. (2017) mentions that defenders' position would have a big impact on the place of server finalization. In other words, the direction of the serve ball is essential for aces. Thus, except for regular strength and power training, it is important to repeat serving practice for gaining experience from other athletes' positions as well as analyzing the opponents' weaknesses.

8. Future Improvements

Our group analyzes ten variables from the dataset. The statistical analysis clearly shows the strong relationship between the probability of winning and two variables, total aces, and total blocks. It indicates these two influence factors that impact the winning probability; however, for future deeper analysis, our group could collect more data and more diverse factors to prove and emphasize the result.

There are also some limitations to our modeling. First of all, we only included data from official champion games such as AVP. This is to say, our data is limited in terms of variety. In the future, we could explore more data on volleyball games to come up with more general ideas. Secondly, we used linear regression and found significant variables that could affect the outcome of games. Next, we could explore more advanced models to predict the results based on these variables more precisely.

Reference

Baacke, H. (1988). Study on the time structure of international championships. Volleyball, FIVB Official Magazine. Jan/Feb, 34-37.

Byra, M., and Scott, A. (1982). A method for recording team statistics in volleyball. Volleyball Technical Journal, 7, 39-44.

Denardi, R.A., Clavijo, F.A.R., Oliveira, T.A.C, Silva, S.L., Travassos, B. & Corrêa, U.C. (2017). The influence of defender's positional gap on the aces in the sport of volleyball. Journal of Human Sport and Exercise, 12(2), 286-293. doi:10.14198/jhse.2017.122.05

George, G., & Panagiotis, Z. (2008). Statistical Analysis of Men's FIVB Beach Volleyball Team Performance. *International Journal of Performance Analysis in Sport*, *8*(1), 31–43. https://doi.org/10.1080/24748668.2008.11868420

João, P. V., Medeiros, A., Ortigão, H., Lee, M., & Mota, M. P. (2021). Global Position Analysis during Official Elite Female Beach Volleyball Competition: A Pilot Study. *Applied Sciences*, *11*(20), 9382. https://doi.org/10.3390/app11209382

Kelly Claes. (n.d.). AVP Beach Volleyball. Retrieved November 28, 2022, from https://avp.com/player/kelly-claes/

Logistic Regression in R Programming. (2020, June 1). GeeksforGeeks. https://www.geeksforgeeks.org/logistic-regression-in-r-programming/

Maddison McKibbin. (n.d.). AVP Beach Volleyball. Retrieved November 28, 2022, from https://avp.com/player/maddison-mckibbin/

Riley McKibbin. (n.d.). AVP Beach Volleyball. Retrieved November 28, 2022, from https://avp.com/player/riley-mckibbin/

Rose, R. (1983). Statistical analysis at the 1983 men's NCAA national championship. Volleyball Technical Journal, 7, 15-17.

Southern California wins 2022 NC beach volleyball championship | NCAA.com. (n.d.). Www.ncaa.com. Retrieved November 28, 2022, from https://www.ncaa.com/live-updates/beach-volleyball/nc/southern-california-wins-2022-nc-beach-volleyball-championship#:~:text=USC%20takes%20down%20Florida%20State

V, A. (2022, October 12). *Beach Volleyball Match Data and Statistics*. GitHub. https://github.com/BigTimeStats/beach-volleyball