Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 2, 2019
A Regional Conference of the Society for Industrial and Systems Engineering

A Logistic Regression Model for Determining Success Based on Star Player Performance at the 2018 FIFA World Cup

Benton Weeks

Department of Systems Engineering
United States Military Academy, West Point, NY

Corresponding author: benton.weeks@westpoint.edu

Author Note: Cadet Weeks is a senior student at the United States Military Academy at West Point. He is an Engineering Management Major who will commission as a Second Lieutenant in the United States Army in May 2019. A special thanks to Dr. Roger Burk and Mr. Ian Kloo for their assistance and advisory roles in the project.

Abstract: The goal of this project is to determine to what extent star players of a team at the World Cup impact the outcome of a match. The purpose is to gain insight into how teams should prepare and execute their offensive game plan during a match based on significant indicators of success. Research and data for this project focused on eight teams across 33 matches which produced 58 observations representing an individual teams' performance. The team was assessed to either overperform or underperform at the World Cup based upon match results as a method of highlighting deterministic player statistics. Statistics were additionally gathered from star players' club teams in order to compare their performance and assess how World Cup conditions influence differences. This project aims to identify the key variables for predicting match outcomes based on star player performance using a logistic regression model discretely parameterized by wins and losses.

Keywords: Soccer, World Cup, Logistic Regression

1. Introduction

The World Cup is unlike any other sporting event in terms of its size and fandom. International soccer players live in four year cycles with the tournament representing the ultimate prize and achievement of a lifetime. For each nation there is much anticipation and pressure to make the team and represent their country at the highest level. Although it is the highest stage a soccer player may perform on, there are still significant gaps in talent and player quality within every squad. Analysts and fans alike look for the star players to seize the moment and shine as the standard of performance is higher than ever. Regarding the World Cup, however, team composition and playing styles are often very different than most of the players are used to when they compete for their professional clubs. Additionally, the team dynamic of playing for a side which is composed of all players who speak the same language, come from the same country, and share a common goal of winning for their nation presents a unique set of conditions. Therefore, to what extent does the star players' performance for their teams affect the outcome of a match at the World Cup?

My hypothesis states that star player performance will have a direct correlation with match outcomes at the World Cup due to them having the highest proportional level of contribution to success, while the null hypothesis states that it is collective team effort and performance which drives a team to success at the world cup. While there are many unique factors and dynamics to consider when analyzing a World Cup match compared to a normal club match, I hypothesize it is the star players who make the deciding plays necessary to secure a victory. Additionally, the star players' higher level of skill would lead individuals to assume that they would be incorporated within the teams' game plan as much as possible. They are accustomed to the spotlight as they compete on very high levels across leagues and tournaments throughout the world and are additionally looked to by their own teammates to be the cornerstone of their attack out on the pitch. While it is a team sport and a collective effort, the statistics may show how their performance, their contribution or the lack thereof, is indicative of their team's eventual outcome.

1.1 Literature Review

Soccer requires a high level of player coordination, communication, and player chemistry in order to achieve success. Unlike sports such as football and baseball, however, soccer has historically had the fundamental difficulty in attempting to make an objective study of team performance due to a lack of routinely recording quantitative data (Pollard & Reep, 1997).

Where baseball and football can account for specific statistics dating back to the 1800s, soccer has a significantly smaller pool of data to build upon when comparing historical trends. Discrete recordings of goals, assists, and yellow cards vastly underplay the real events of a match and its continuous nature. Additionally, soccer has the problem of player positions and their desired performance objectives, such as a central defensive midfielder whose role it is to repossess the ball for his team compared to a wing who focuses on attacking the goal. Due to this, soccer is highly complex in structure and composition as players have particularly varied roles and subgroups, to include goalkeepers, defenders, attacking midfielders, wings, strikers, etc., with each role requiring a specific set of skills (McHale, 2012). While breaking down positions into generalized areas can be of great use, it is not entirely comprehensive in representing how a player is expected to contribute on the field. As with many sports, soccer has evolved over the years as pace of play increased, talent of players increased, and the desire to win has increased the necessity of winning in the moment. For these reasons, soccer has been a bit slower in adopting statistics as a helpful tool in determining success predominantly due in part to the complex, dynamic nature of the game, compared with the more discrete nature of other sports as aforementioned (Szczepański & McHale, 2016).

The most significant addition to the field of soccer analytics in creating deterministic models may be credited to Hughes and Franks on the 1990 World Cup where they conducted an analysis comparing successful and unsuccessful teams (Hughes & Franks, 2005). Their findings revolved around the differences between the two sides in converting possession into shots on goal, with the successful teams having the better ratios. However, Hughes and Franks found that there were no significant differences between the successful and unsuccessful teams' patterns of play leading to shots. Although their study could be viewed as too simplistic, it catalyzed a field which has since exponentially grown as technology and available resources have flooded the industry. As the game has evolved, so have the strategies of clubs and coaches aiming to gain an edge over the opponent. Coaches are known to make more subjective calls during a match based off the current play of their players, their health, and what substitutions may be necessary to increase their sides probability of winning. However, they may be unable to recall events reliably or have a perfect answer to every situation and are therefore increasingly turning to match analysis as a way of optimizing the training process of their players and teams (Castellano, Casamichana, Lago, 2012).

There have been many studies within club leagues and international competitions which share similar trends and areas of focus. The art of passing, for example, has found a strong niche within the soccer analytics community as experts realize that it is the most fundamental aspect of the sport. One study on the 2014 World Cup in Brazil focused entirely on long distance passes and their rates of possession loss (Mattos dos Reis, 2017). Another study on the same event found that the matches that ended in a win were characterized by a significantly higher number of medium passes and a significantly higher percentage completion rate for long passes (Król, 2017). While passing is crucial to success, it is not the only variable in determining success. Passing within soccer is rather the foundation upon which the execution of a multitude of other skills is built upon. For this reason, studies have considered many other variables within their individual models in pursuit of determining success to include total shots, shots on goal, crosses, crosses against, ball possession, venue, which part of the field passes are made in, etc. As the field becomes as competitive as ever, soccer fans around the world are eager to see their teams race to find that extra inch of advantage over their opponent which might bring home a victory.

This project considers many of the same variables which others have focused on and applies them in a specific manner. By focusing on the performance of the star players within these contexts it will be possible to conclude to what extent they have an impact on the outcome of a FIFA (International Federation of Association Football) World Cup match.

2. Data Collection and Variable Selection

2.1 Selection of Teams

In order to collect representative data for a logistic regression model in determining winning or losing a match, it is important to select the right teams to be reviewed. I chose eight teams of focus based on their initial FIFA international ranking before and after the tournament. Four teams that reached the final four, France, Croatia, Belgium, and England, represent the successful teams while Germany, Poland, Peru, and Costa Rica represent four teams that had high hopes and rankings entering the tournament and vastly underperformed (Table 1). The teams were deliberately chosen not only based on rankings but also to make sure there were relatively similar rankings across the board when the tournament began to the most recent rankings following its conclusion. If two teams had similar rankings before the tournament but performed differently, that ended up being the focus of my research. My model does not only look at their performance. To construct a comprehensive logistic model, I had to use comparative statistics from their opponent as well and to see if there was something that one side did that was effective compared to the other. This was also useful in establishing the discrete output of the model as a win or loss in a binomial fashion. Ties were accredited as losses due to the focus of the model in predicting winning. The teams chosen to represent comparative statistics against one another appear along the same line of Table 1.

Table 1. Teams Selected for Logistic Regression Model

Top Four Teams	Pre-WC FIFA Ranking	Post-WC FIFA Ranking	Underperforming Teams	Pre-WC FIFA Ranking	Post-WC FIFA Ranking
France	7	1	Poland	8	18
Croatia	20	4	Costa Rica	23	32
Belgium	3	2	Germany	1	15
England	12	6	Peru	11	20

2.2 Selection of Star Players

Following team selection, the selection of the star players along with defining what constitutes a star player became the subsequent task. A combined method of utilizing the most recent FIFA player rankings before the tournament, personal knowledge of the players, and an account of their actual contribution within the tournament filtered the selection. Regarding assessment of a players' contribution, a team would often have two star caliber players at the same position, but one was removed due to a lack of playing time, injury, or another significant reason. Other scenarios arose, for example, where an excellent player from that nation did not make the team based upon the managers discretion or a star player did not contribute enough playing time for it to be significant data. Additionally, I limited the number of star players per team to three as a means of collecting accurate data for comparing the difference between the star players and the team performance overall. Although many teams had more than three high ranking players, there were always three who stood out and would fit the model for what it is designed to produce. Some teams did not have a total of three star players but rather only one or two. This was deliberately done as a means of keeping the integrity of what a star player is along with avoiding corruption of deterministic star player statistics. This varied from country to country as some of the less competitive nations who competed against the selected teams had no star players at all. A key component of this pool of players is that they are all either midfielders or forwards. This was primarily due to the statistics analyzed within the study focusing predominantly on the offensive side of the sport. While defenders and goalkeepers are very important, the focus of the model is on certain statistics which are more indicative of scoring and attacking play. The model uses many metrics of these star players during the World Cup along with data from their most recent club competition year as a comparison between their output, performance, and utilization from one environment and playing style to another.

2.3 Data Collection

The data is drawn from two main sources. The first source was the official 2018 FIFA World Cup website where are found many different types of statistics for every match played at the tournament (FIFA, 2018). Box scores, individual player statistics, and game passing matrices were available for use. Many of the described variables are products of the passing matrices collected. These matrices display every pass which occurred throughout the game to include the passing and receiving player. The size of the player is relative to their degree within the matrix along with every directional arrow representing a pass as displayed in Figure 1. Additionally, the star players are highlighted in yellow for reference. The second source was an online database of all major soccer club league statistics (WhoScored.com, 2019). This was helpful in determining specifically how their play differed during and throughout the World Cup along with analyzing if their performance variance became a determinant of success. In total, there were 33 observed matches which produced 58 observations of one side's play during that match. The missing observations are predominantly due to a lack of star players on that side for the exhibition. It was important to exclude these observations as the presence of a star player often impacts how the rest of the team performs as far as playing style and ball distribution are concerned. Including such observations would alter how the impact of a star player is measured within the model and a success may be attributed to many other factors such as strategic formations which expose a team altogether. One limitation of this project is the lack of goalkeeping and defending players as additional contributors to the outcome of a match.

2.4 Variables

There is a total of 19 variables analyzed per observation within the dataset. Each observation has a value for each variable with some specifications to clarify. In order to not inflate percentage data, a few rules were emplaced to dissuade such manipulation. For long, medium, and short passing totals there must be at least four, six, and eight attempts made in total, respectively, in order for the observation to be included. If the threshold is not met, the value becomes zero as the player or players collectively did not contribute enough statistically to be significant. A passing matrix was produced for every observation which was utilized to calculate the average degree, betweenness, closeness, and eigenvector value of each players’ performance. These variables represent the exploratory aspect of the model compared to many variables, such as “Star Assists per Minute,” which are more readily considered to be predictors of success.

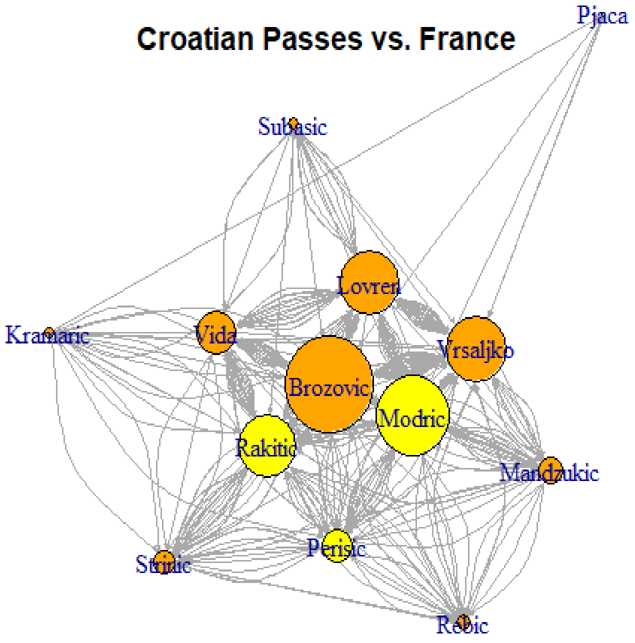


Figure 1. Passing Matrix of Croatian Players in the World Cup Final Against France

Table 2. Logistic Regression Model Variables

Variable	Unit	Description
Star Passing Long	passing %	The Star players’ long pass completion percentage
Star Passing Medium	passing %	The Star players’ medium passes completion percentage
Star Passing Short	passing %	The Star players’ short passes completion percentage
Star Passing Total	passing %	The Star players’ total passing completion percentage
Star Passing Differential	passing %	The percent difference between the passing completion percentages of the two sides’ star players
Team Passing Long	passing %	The team’s long pass completion percentage
Team Passing Medium	passing %	The team’s medium passes completion percentage
Team Passing Short	passing %	The team’s short passes completion percentage
Team Passing Total	passing %	The team’s total passes completion percentage
Team Passing Differential	passing %	The percent difference between the two teams’ collective passing completion percentages
Average Star Degree	dimensionless	The average Star degree within their passing matrix
Average Star Betweenness	dimensionless	The average Star betweenness within their passing matrix
Average Star Closeness	dimensionless	The average Star closeness within their passing matrix
Average Star Eigenvector Value	dimensionless	The average Star eigenvector value within their passing matrix
Star Shots per Minute	shots/minute	The Star players’ average shots per minute of playtime
Star Club Assists Differential	assists/minute	The difference between the Star players’ average assists per minute in the World Cup and their club play
Star Club Shots per Minute Differential	shots/minute	The difference between the Star players’ average shots per minute in the World Cup and their club play
Star Club Passing Differential	passing %	The difference between the Star players’ average passing completion percentage in the World Cup and in their club play
Non-Star Shooting	shot %	The rest of the team’s players percentage of the number of shots taken

3. Logistic Regression Model

The final logistic regression model was found by conducting three iterations of breaking the data into parts and using the resulting significant variables as a test to predict winning on the remaining observations. The 75/25 rule was adhered to by randomly separating the data into one set of 44 observations, which generated the model, and another set of the remaining 14 observations, which became the test subject. After every iteration, the two most prominent and significant variables became evident.

$$P(Win) = 0.45702 + 2.4980(Star\ Passing\ Long) - 0.0223(Star\ Degree\ Average) + \varepsilon \quad (1)$$

The probability of a team winning based on the passing percentage of long passes along with the average degree of the star players within their sides' passing matrix became the two most significant determinants of success. It is interesting to point out how the star degree average has a negative coefficient indicating that star players should not be too heavily relied upon within the context of distributing the ball to teammates.

4. Results

4.1 Model Results

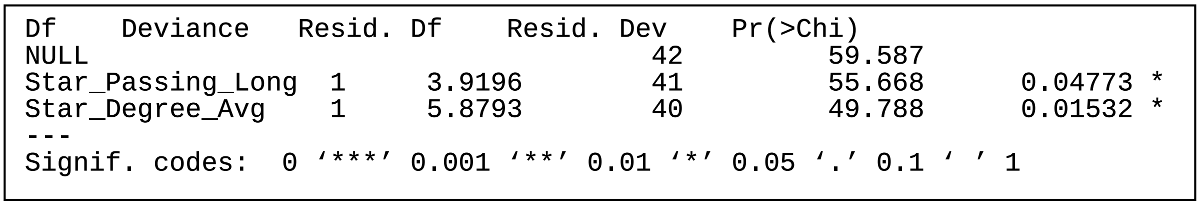
The significance of the variables is evident in their small p-values indicating they are relevant and effective determiners of success as seen in Table 3. Conclusively, the model may be confident in relying upon these variables as accurate determinants of success.

Table 3. Logistic Regression Model Variables and Significance

COEFFICIENTS:

	Estimate	Std. Error	z value	Pr(>|z|)
(INTERCEPT)	0.45702	0.74947	0.61	0.542
STAR_PASSING_LONG	2.49799	1.04446	2.392	0.0168*
STAR_DEGREE_AVG	-0.02228	0.01071	-2.08	0.0375*

Additionally, a Chi-Squared test was run as a means of analyzing if the two aforementioned variables are related to one another in a significant manner. Since the p-values of both variables are again small and therefore significant, we may be confident in our model at accurately predicting success. This "goodness of fit" statistic measures how well the observed data fits with the distribution that is expected if the variables are independent.



```
Df    Deviance   Resid. Df    Resid. Dev    Pr(>Chi)
NULL                            42         59.587
Star_Passing_Long  1     3.9196        41         55.668      0.04773 *
Star_Degree_Avg    1     5.8793        40         49.788      0.01532 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2. Chi-Squared Test of Model

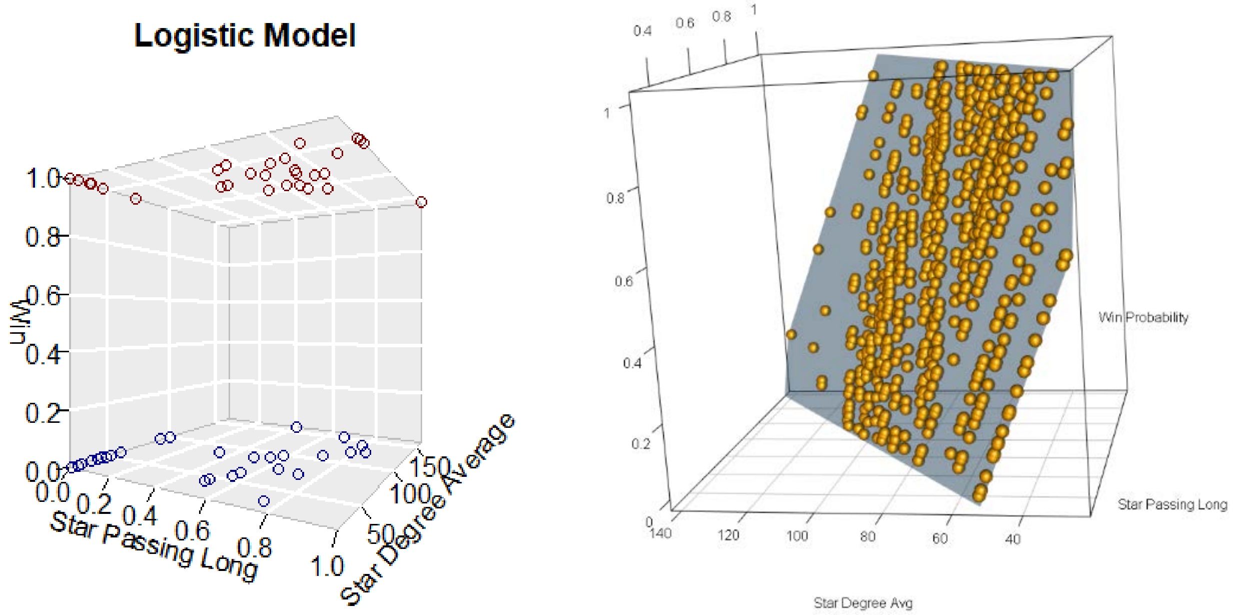


Figure 3. Data Observations Plotted (left) Along with a 3D Spatial Representation of the Model (right)

After vetting the model, it was plotted in order to have a better visual representation of its output in predicting success. In Figure 3 the blue and red circles represent the observation outcomes and the actual observation outcomes of a win or loss along with their position with respect to the other two variables. The cluster of data points along the bottom of the “Star Passing Long” variable minimum output validates the importance of star players leveraging their skill in completing the more difficult passes. The second plot displays every possible combination of the two variables from the data and their resultant probability of success. The plane is the model in a three-dimensional space rather than a traditional S-curve shape. It is a plane because it has only two predictive variables rather than a more complex surface.

4.2 Testing Model

The model was then compared to the remaining data set to test for accuracy in predicting success. Figure 4 below displays the graphical outputs from the three randomized iterations where the x-axis is the false positive rate and the y-axis is the true positive rate. The AUROC (Area Under a Receiver Operating Characteristic) number signifies the success rate of predicting matches at the World Cup. While some are high, such as the third iteration, and all of which are above 50%, the large gap in the success rates raises concern. Ideally the ROC curve should steeply increase and level out as a way of signifying a good and accurate model. The small data set size is likely the root cause of the large variation seen below.

The three randomized test sets when used in conjunction with the model produced a misclassification rate of 28.57%, 50%, and 28.57% along with a concordance value (the percentage of pairs, whose scores of actual positives are greater than the scores of actual negatives) of 57.78%, 55.10%, and 75.51% respectively. A perfect model has a concordance value of 100% but anything over 50% means that it has some level of validity in successful predicting outcomes.

5. Conclusion and Discussion

The hypothesis of star player performance having a significant impact on match results appears to be true, to an extent, based on the findings. Although the validity of the model might be called into question due to a small sample size, I believe the model has some utility even though there are some apparent gaps in its application when applied to this data set. For the two variables of significance, star player long passes completion percentage and their average degree within the passing context, they

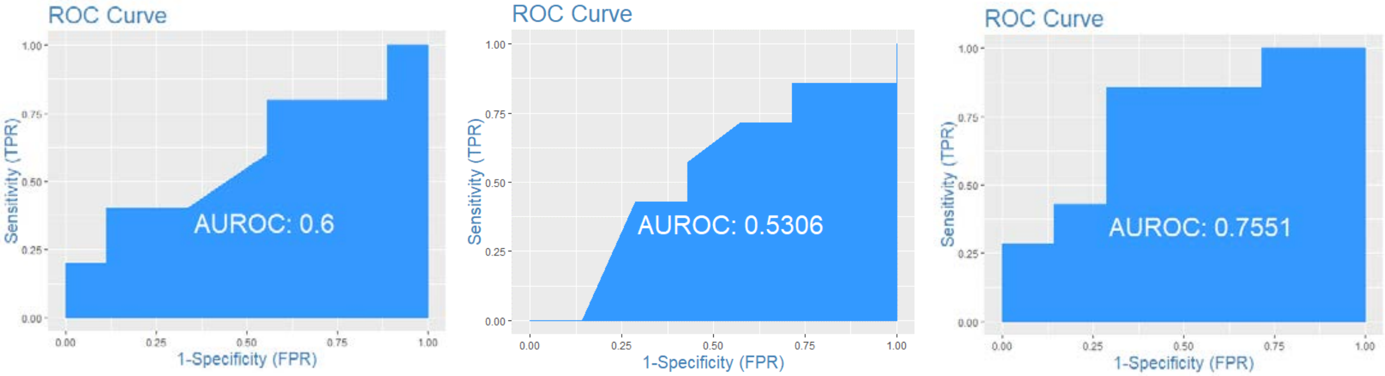


Figure 4. ROC (Receiver Operating Characteristic) Plots for Predicting Match Success on Randomized Data Test Sets

intuitively make sense. Many studies have found long pass completion percentages as strong determinants of success in part to their risky nature. If a long pass is not completed then it is likely a turnover, giving the opposition an extra possession to capitalize on and therefore an additional chance to score. The degree of the star players is the more interesting of the two as it advocates for the star players to not become too large of an influence within the attacking game plan. Their perceived higher level of skill would lead people to believe that they should attempt to do as much as possible. However, within the context of the World Cup it is possible that the opposition focuses heavily on star players, putting them under pressure and causing negative results. It is important to highlight the distinction that the degree only refers to star player presence within their coordinated passing scheme and other variables, such as who should take the most shots or assists generated, may still be significant. The fact that these variables were more significant predictors in determining success than "Star Shots per Minute," for example, also brings validity to the model as it is often one of the strongest determinants of winning a game if you are taking possessions and turning them into scoring opportunities.

It is also interesting that the comparison of the star players with their club performance was not significant in determining success as they are often called upon to do more at the World Cup due to their elevated capabilities compared to their teammates. Many players even play for the same club and then oppose one another at the World Cup. These players play for top clubs around the world where they are constantly performing alongside the best in the world. This causes them to feel as if they don't have to do too much in pursuit of a victory as they may rely on their teammates whereas in the World Cup their entire nation is looking to them to make the decisive play to secure a victory.

The methodology of focusing on the star players is something which I believe may be easily translated not only to a bigger context such as multiple World Cups over time, but also in multiple soccer markets around the world. Although the dynamic of a club match is different than the World Cup the same principle of star player performance still holds true. Perhaps their performance is more relevant in that environment than in the World Cup which would conclude that the Word Cup is a more team-oriented game than club competition.

For future work, one might consider increasing the sample size to include all matches from the world cup rather than focusing on four teams who, in theory, should show apparent differences in their star players performance based on outcome. An increased data size, even across multiple world cups, would likely produce more conclusive and reliable results. There are also many more variables to explore and a model which considers defenders and their performance might prove to be more robust and accurate in determining success than focusing on the other two-thirds of the team. An individual or organization which possesses more resources to include increasingly detailed or interesting variables might also consider leveraging their capabilities to generate a more comprehensive study. Modern technology has allowed the ability for many match analysis factors, such as player distance covered and time of possession within specific parts of the field, to add additional dimensions to the evaluation of player performance. Based upon the findings, however, a manager should focus on improving the long passing ability of their players along with hesitating to incorporate their star players too heavily within their offensive game plan.

6. References

Castellano, J., Casamichana, D., & Lago, C. (2012). The Use of Match Statistics that Discriminate Between Successful and

Unsuccessful Soccer Teams. *Journal of Human Kinetics*, *31*, 139–147. Retrieved from https://usmalibrary.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=73996933&site=eds-live&scope=site.

FIFA. (n.d.). 2018 FIFA World Cup Russia™. Retrieved from https://www.fifa.com/worldcup/.

Hughes, M., & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences*, *23*(5), 509–514. Retrieved from https://usmalibrary.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=17385149&site=eds-live&scope=site.

Król, M., Konefał, M., Chmura, p., Andrzejewski, M., Zając, T., & Chmura, J. (2017). Pass Completion Rate and Match Outcome at the World Cup in Brazil in 2014. *Polish Journal of Sport & Tourism*, 24(1), 30–34. Retrieved from https://usmalibrary.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=123483308&site=eds-live&scope=site.

Lago-Peñas, C., Lago-Ballesteros, J., & Rey, E. (2011). Differences in performance indicators between winning and losing teams in the UEFA Champions League. *Journal of Human Kinetics*, 27(1), 135-146. doi: https://doi.org/10.2478/v10078-011-0011-3.

Mattos dos Reis, M. A., do Amaral Vasconcellos, F. V., & Bezerra de Almeida, M. (2017). Analysis of the Effectiveness of Long Distance Passes in 2014 Brazil FIFA World Cup. / Análise da Eficácia dos passes de longa distância na Copa do Mundo FIFA Brasil 2014. *Brazilian Journal of Kineanthropometry & Human Performance*, 19(6), 676–685. Retrieved from https://usmalibrary.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=127926661&site=eds-live&scope=site.

McHale, I., Scarf, P., & Folker, D. (2012). On the Development of a Soccer Player Performance Rating System for the English Premier League. *Interfaces*, 42(4), 339-351. Retrieved from http://www.jstor.org/stable/23254864.

Pollard, R., & Reep, C. (1997). Measuring the Effectiveness of Playing Strategies at Soccer. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 46(4), 541-550. Retrieved from http://www.jstor.org/stable/2988603.

Szczepański, Ł., & McHale, I. (2016). Beyond completion *rate: evaluating the passing ability of footballers. Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2), 513–533. https://doi.org/10.1111/rssa.12115.

WhoScored.com. (n.d.). Retrieved from https://www.whoscored.com/Statistics.