Data Analytics: Making its Way in Soccer

Matthew Myers
Capstone Advisor Haley Skipper
University of Washington Tacoma

2019

Abstract

In a data driven world, the ability to put all the information into a predictive tool for competitive environments is becoming ever more valuable. The field of data analytics in sport is becoming a hot topic, yet for some sports, they still have further to go than others. Soccer is a sport that is in a stage of growth for its data analysis uses and capabilities. This study aims to take on the issue of which match statistics are important within a soccer match. Beginning with replicating a previous study on the 2014 FIFA World Cup, match statistics are taken from the 2018 FIFA World Cup to help predict the outcome of a match within the group stage. This study added two more match statistics to the list analyzed in the previous tournament, and found similar results to those from the 2014 study. It was found that in any game of the 2018 World Cup group stage, match statistics such as shots on target, shots from a counter attack, ball possession, aerial advantage, average pass streak, and tackles all could be concluded to have a beneficial effect towards a team's chances of winning. Match statistics that were found to have a harmful effect for teams in any group stage game were offsides, crosses, dispossessed and long passes.

Contents

1 Introduction

The lucrative world of professional sports is always searching for the competitive edge. When the difference of winning and losing a game could be millions of dollars, teams are willing to employ the help of any tool within their arsenal. Due to the current data revolution in our technology driven world, more information is now accessible than ever before. The only issue now is, what can we do with it?

Data analytics was first used to analyze a soccer match in 1950 by Charles Reep, an English accountant. Using pen and paper, Reep went on to annotate over 2,200 matches and to have his findings published in the scientific paper "Skill and Chance in Association Football" by the *Journal of the Royal Statistical Society in 1968*. Reep's findings didn't come without criticism though as he interpreted his results that a more direct style of play is more efficient. [2] As we'll see in more recent studies, that may not be the case currently in soccer.

It wasn't until the "Moneyball" narrative broke out in 2003 through Michael Lewis's book, *Moneyball: The Art of Winning an Unfair Game*, where a revolution for professional sports and data analytics began. Lewis's book explains the use of "Sabermetrics", an empirical analysis in baseball statistics from in-game activity. "Today, every major professional sports team either has an analytics department or an analytics expert on staff" [1]

Soccer is no different, but there is still some progress to be made before it becomes a significant component of the daily, on field tactical operations at a club. As Luke Bornn, Vice President of Strategy and Analytics for the Sacramento Kings puts it, "Baseball is the pioneers of sports analytics. They are 10 years ahead of basketball, which is 10 years ahead of soccer". [4] This large gap in progress of analysis can be credited to the different nuances of the sport, as well as how the statistics were interpreted in the past.

Unlike football, baseball, and basketball, soccer is a game that has a much lower level of scoring. The most common score line is a 1-1 draw [2]. Therefore it is harder to calculate a player's complete effect on the end result in a game using relevant statistics. The other question being which match statistics are relevant, and are they equally relevant for each player? For instance, goals scored and appearances are some of the most common statistics for summarizing a player's career, but is it fair to judge defenders based on their goals?

By using a Sabermetrics style approach, we can look at different match events to determine which play a significant role towards a team's end result each match. Multiple studies have tried to tackle the issue for soccer as to which match statistics matter. In 2012, a study was conducted by Julen Castellano, et al., titled, "The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams", and was published in the *Journal of Human Kinetics*. This study looked at results from the 2002, 2006, and 2010 World Cups and performed an ANOVA test of significance on a handful of team attacking and defending match statistics and how they could predict which teams were successful. [7] In 2015, a similar study was published in the *Journal of Sports Sciences* that aimed to find match statistics from the 2014 World Cup that were significant in predicting the winners.

The study, “Match Statistics Related to Winning in the Group Stage of 2014 Brazil FIFA World Cup" by Hongyou Liu, et al., used a newer form of significance testing called Magnitude Based Inference. [3]

Castellano, et al. only looked at a handful of match events, mostly ones that can either be committed or received. The list of statistics can be seen in (1). Through discriminate analysis of those three world cups, it was found that match statistics that best represented the success of teams were total shots (made and received), shots on target (made and received), and ball possession. They also found that which test statistics that were found to be significant, varied between World Cups as common strategies differed.

Variables studied in the last three soccer World Cups

Variables related to	Variables or match statistics or performance indicators
Attacking play	*Goals scored, Total shots, Shots on target, Shots off target, Ball possession, Off-sides committed, Fouls received, Corners.*
Defence	*Total shots received, Shots on target received, Shots off target received, Off-sides received, Fouls committed, Corners against, Yellow cards, Red cards.*

Figure 1: Test Statistics included in the 2012 study on three soccer World Cups. [7]

The study analyzing the 2014 World Cup took into account twenty-four match statistics from the group stage matches of the tournament. This gave 96 observations from 48 games, while also analyzing 38 of those games separately as close games, matches that end with a goal differential of two or less between the teams. From their results, Hongyou, et al. found that shots on target, shots from counter attack, and tackles all seemed to be significantly beneficial for teams during all games and close games of the group stage. Crosses remained significantly harmful for teams during all games and close games. Other statistics to note that were significantly beneficial for one test group of games, but not both, are shots from inside area and aerial advantage. Yellow cards and red cards were found to be significantly harmful in one of the test groups of games, but decreased in the other. [3]

Our aim is to replicate the 2015 study for the 2018 FIFA World Cup in Russia with the intention of exploring which match statistics previously found as significant indicators hold true for the most recent tournament. As noted earlier, not all common strategies stay true for each tournament every four years. After comparing the results, we will explore whether recent analysis of performance indicators for individual players match some of the match statistics we found to be significant.

2 Methods

This analysis will be based on the 48 group stage matches from the tournament. A group stage is the first portion of matches for the 32 team tournament where teams are broken up into eight groups of four and will play each team in their group once. Due to the group stage matches not being single elimination, these matches end after the allotted 90 minutes give or take stoppage time, which is minutes added at the end of each half determined by the referees. If the result is tied after full time, then the match ends in a draw. Teams receive three points for a win, one for a tie, and zero for a loss during this stage. Top two teams from each group, based on points and tie breakers, advance to the knock out stages.

Similar to the study of the 2014 World Cup, here we will look to analyze the data from two test groups of group stage matches, all games and close games. Both previous studies analyzed data only from matches that didn't go into extra time due to the fact that a team's tactics may change more often between single elimination and non single elimination games. hence, knock out stage matches would be contained within a different data set and would provide a smaller sample size to model.

2.1 Variables

Twenty-six match statistics were chosen, adding two new ones (key passes and dispossessed) to the original twenty-four from the previous study. Each statistic was recorded from the soccer statistics website *whoscored.com*. These variables can be separated and analyzed in three different groups representing different aspects of the game: variables related to goal scoring (8), variables related to passing and organising (14), and variables related to defending (4). Each variable with the definition as per *whoscored.com* and which group of variables it falls under is as follows:

- **Goal Scoring Variables:** Shot, Shot on Target, Shot Blocked, Shot from Open Play, Shot from Set Piece, Shot from Counter Attack, Shot from Inside Area, Shot from Outside Area
- **Passing and Organising Variables:** Ball Possession (%), Pass, Pass Accuracy (%), Long Pass, Short Pass, Through Ball, Average Pass Streak, Key Pass, Dispossessed, Cross, Dribble, Offside, Corner, Aerial Advantage (%)
- **Defending Variables:** Tackle, Foul, Yellow Card, Red Card
- **Shot:** an attempt to score a goal, made with any (legal) part of the body, either on or off target.
- **Shot on Target:** an attempt to goal which required intervention to stop it going in or resulted in a goal/shot which would go in without being diverted.
- **Shot Blocked:** a shot towards goal that is blocked by a field player on the opposing team.

- **Shot from Open Play:** a shot attempted not stemmed via a dead ball situation.
- **Shot from Set Piece:** a goal attempt that stemmed from a dead ball situation (corner kick, free kick, or throw in).
- **Shot from Counter Attack:** a goal attempt generated from a counter-attack situation. A counter-attack situation is logged when (a) the ball must be turned over in the defensive half; (b) the ball must be quickly manoeuvred (6 sec, 3 passes) into the attacking third and must be under control; (c) the defending team must have four or less defenders in a position to defend the attack and attacking players must match or outnumber the defensive teams players and (d) the ball must be fully under control in the oppositions defensive third.
- **Shot from Inside Area:** a shot attempted within the opponents 18-yard box.
- **Shot from Outside Area:** an attempt on goal taken outside the opponents 18-yard box.
- **Ball Possession (%):** the proportion of total duration when the ball was in play where a team takes over the ball from the opponents without any clear interruption.
- **Pass:** an intentionally played ball from one player to a teammate.
- **Pass Accuracy (%):** the proportion of attempted passes that successfully made it to a teammate.
- **Long Pass:** an attempted pass from 25 yards or more.
- **Short Pass:** a pass attempted from less than 25 yards.
- **Through Ball:** an attempted pass splitting the opposing defensive line (last line of players) to play a teammate in on goal.
- **Average Pass Streak:** the average number of passes attempted in each series of consecutive passes.
- **Key Passes:** a pass leading to a teammate's shot at goal.
- **Cross:** an attempted pass from a wide area to a central attacking area.
- **Dribble:** taking on an opponent and successfully getting past them while retaining the ball.
- **Offside:** being caught in an offside position resulting in a free kick for the opponents.
- **Dispossessed:** being tackled by an opponent without attempting to dribble past them.
- **Corners:** when the ball is played out off of the defending team over their own goal line it results in a goal kick for the attacking team.

- **Aerial Advantage (%):** the proportion of headers won in direct contest with an opponent.
- **Tackle:** an attempt to dispossess an opponent.
- **Foul:** an illegal maneuver by a player that results in a free kick awarded to the other team.
- **Yellow Card:** when a player commits a yellow card infringement for foul(s), hand ball, dangerous play, time wasting, etc. and is shown the yellow card by the referee.
- **Red Card:** a player is shown a red card by the referee for committing a red card infringement or is shown two yellow cards.

As noted by Hongyou and et al., each variable that isn't already given in a percentage or average (Ball Possession, Pass Accuracy, Aerial Advantage, or Average Pass Streak) will be adjusted to per 50% possession of the ball for the team. This is to take into account each team's match statistic's results evenly as a team may have much lower ball possession than their opponents which will negatively affect their variables related to goal scoring or passing. In regards to defending variables, if a team has less ball possession than their opponent, then they are likely to have more results for defending variables. In this case we are able to weight each team's results evenly without having them skewed by Ball Possession.

For the case of variables related to goal scoring or passing, match statistics are adjusted using formula (1) as seen below. V_{ajstd} being the adjusted value of a team's observed match statistic from a specific game, and $V_{original}$ is the original match statistic value for the team from that observed game. From the same game observation, we take BP_{team} as the Ball Possession for that team when adjusting match statistics related to goal scoring or passing and organising. We can then adjust such observations by using formula (1). For example, if a team takes 14 shots during a game, and had a ball possession rate for 60% of the game, we would find the adjusted shots as $V_{ajstd} = (14/.6) \times .5 = 11.667$ shots.

$$V_{ajstd} = (V_{original}/BP_{team}) \times 50\% \tag{1}$$

For the defending variables, the same formula can be applied, except BP_{team} is replaced with $BP_{opposing}$ for Ball Possession of the opponents during the observed game. By using formula (2) we see that if a team had 18 tackles during a game, and their opponents had a ball possession of 40%, then we would find that they had, $V_{ajstd} = (18/0.4) \times 0.5 = 22.5$ adjusted tackles.

$$V_{ajstd} = (V_{original}/BP_{opposing}) \times 50\% \tag{2}$$

We will look to analyze matches from the group stage of the 2018 FIFA World Cup within two test groups, all games and close games. A game is considered "close" if the goal differential between the two teams is two or less. Out of the 48 games during the 2018 tournament, 40 of those were close games, providing 80 observations for all close games where each team in a match is counted as an observation.

2.2 Logistic Regression Analysis

A logistic regression model was used by taking the value of each match statistic (adjusted as needed) being the independent variable, to build a model for predicting the match outcome as the dependent variable. As stated in [3], we must use a logistic regression as our dependent variable has thresholds. Teams can do no better than a win, or worse than a loss. We then have the success rate of a team (dependent variable), as 1.0 for a win, and 0.0 for a loss. This allows us to take into account observations from matches that ended as draw, with the success rate as 0.5. All statistical analysis and logistic regression was done using the statistics software **gretl**. [11] As shown in figure (2), we see that the log coefficient for shots on target in all games for 2018 is 0.276345.

```
Model 1: Ordered Logit, using observations 1-96
Dependent variable: Result
Standard errors based on Hessian

            coefficient   std. error     z     p-value
  ---------------------------------------------------------
  SOT         0.276345     0.0998295    2.768   0.0056  ***

  cut1        0.668985     0.426955     1.567   0.1171
  cut2        1.48914      0.448057     3.324   0.0009  ***
```

Figure 2: Shots of Target (SOT) logistic regression run on all games of the group stage for the 2018 tournament.

Once the logistic coefficient is found, a formula for the line of best fit can be found as formula (3). [8] Where p is the proportion of a team's success rate, x is the independent variable (match statistic), b_0 is the constant, and b_1 is our log coefficient. By [10] we know that formula (3) gives us the Odds Ratio. This is important as in order to build Magnitude Based Inferences (MBI) for the effect of each test statistic, we need to calculate the effect a two standard deviation increase has on the Odds Ratio for the team's success rate. [3]

$$\frac{p}{1-p} = e^{(b_0+b_1x)} \tag{3}$$

From logistic regression equation 3, we can also apply natural log (ln) to both sides to give us the equation in terms of log-odds. By using formula (4), we are able to find how a change in the odds ratio or the observed test statistic has an effect on a team's success rate. [9] This will be needed to help calculate the change in Odds Ratio and after a two standard deviation increase in a match statistic.

$$ln(\frac{p}{1-p}) = b_0 + b_1x \tag{4}$$

To see how a change will effect the Odds, we set up the equation (4) as $ln(Odds) = b_0+b_1x$. As our aim is to calculate the effect a two standard deviation increase on a match statistic has on the success rate of teams, we will apply this change to a confidence interval of the

Odds Ratio. From there we can then conclude the chances a change would have a significant effect.

2.3 Confidence Intervals

Using **gretl** with our data for the logistic regression analysis, we can also find confidence intervals representing the relationships between our match statistics and the results for teams during the group stage. As per the study aimed at the 2014 World Cup, we will use 90% confidence intervals in our analysis. For example, looking at figure (3) we find the 90% confidence interval of shots on target (SOT) calculated for all group stage games within the 2018 FIFA World Cup.

VARIABLE	COEFFICIENT	90% CONFIDENCE INTERVAL	
SOT	0.276345	0.112141	0.440550
cut1	0.668985	-0.0332936	1.37126
cut2	1.48914	0.752152	2.22613

Figure 3: Shots of Target (SOT) 90% confidence interval calculated for all games for the group stage of the 2018 World Cup.

To adjust these confidence intervals for a two standard deviation increase, we can multiply the values of the upper and lower bounds by the calculated two standard deviations for each match statistic. By [3] we will also use "1" as two standard deviations for a yellow card or red card. That is to find the effect that an additional card would have on the end result for a team. By applying this calculation into our equation for the Odds Ratio (formula 3) and disregarding the constant (b_0) by the calculations done for the 2014 tournament, we then have our adjusted Odds Ratio as seen in equation (5). Where OR_a is the Odds Ratio adjusted at the given value of the bound ($bound$) for the match statistic's confidence interval after a two standard deviation ($2SD$) increase. Note by using the log coefficient for a match statistic in place of a bound, we can then calculate the mean of the the adjusted Odds Ratios.

$$OR_a = e^{(2SD \times bound)} \tag{5}$$

2.4 Magnitude Based Inferences

The magnitude based inference (MBI) test was developed in 2006 in response to criticism against the common null hypothesis test and is often used in clinical trials, or other tests related to finding a potentially harmful or beneficial effect a testing variable may have. Instead of testing the null hypothesis to a specific p-value, a magnitude based inference analyzes the change in the odds of a success after a two standard deviation change for the independent variable. After adjusting the odds ratio for this increase, an interval can be created to demonstrate the changes within the probability of success from the two standard deviation increase. This interval of the change in the odds ratio provides us with our Magnitude Based

Inference. [5]

In regards to determining the smallest worthwhile change, a 10% increase in the chance of a team winning their match will be considered. This is the same as the study from the 2014 World Cup. With a magnitude based inference, you do not simply accept or reject a result based on meeting a certain p-value like a null hypothesis test. Instead it is inspected at what proportion of each interval is located in either the harmful, trivial, or beneficial areas determined by the thresholds of the smallest worthwhile change.

Looking at figure (4) we have examples for different resulting MBI intervals and how to provide a clinical decision based on where the interval is located. In the example, the smallest worthwhile change is determined as a $\pm 5kg$ change in the end result being analyzed, with the vertical center line representing no change.

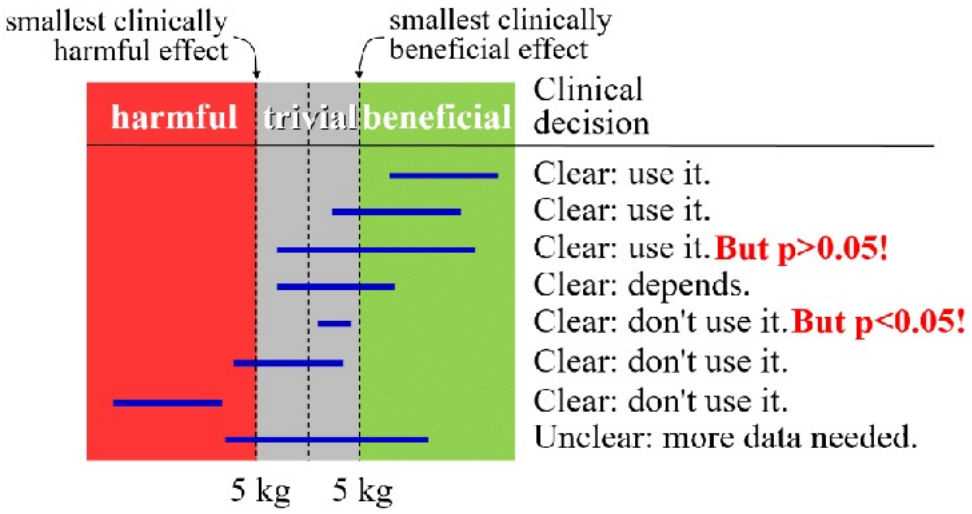

Figure 4: Examples for how to interpret a Magnitude Based Inference interval.

It is then from a quantitative analysis of the MBI interval proportions falling under the categories of beneficial, trivial, and harmful, that a we can can give qualitative statements of significance. Therefore we can dictate how likely it is for a match statistic to play a significant role towards the final result for a team in any given group stage match. Based on the calculated proportion of the interval, and using the list below, we are able to give how likely that effect is. For example, if an interval has 0% past the lower worthwhile change threshold into the **harmful** category, 15% within **trivial**, and 85% as **beneficial**, then we can conclude that the independent variable is *likely* beneficial.

- *<0.5%* - most unlikely
- *0.5-5%* - very unlikely
- *5-25%* - unlikely
- *25-75%* - possibly
- *75-95%* - likely
- *95-99.5%* - very likely
- *>99.5%* - most likely

To get our final, calculated MBI intervals, we need the Odds Adjusted equation (5) from the previous section. Our new interval needs to show the difference of probability of winning by a two standard deviation change in the value of a match statistic. "The two-SD (standard deviation) increase stands for the change in a statistic from a typical low value (-SD) to a typical high value (+SD)". [3] In order to see how this change affects the Odds Ratio, we'll begin by looking at the log-odds equation (4) where $ln(Odds) = b_0 + b_1x$. If $x = x_0$, $Odds_0 = \frac{P_0}{1-P_0}$, and $P_0 = 50\%$, we know that $ln(Odds_0) = ln(1) = 0 \rightarrow 0 = b_0 + b_1x_0$. Since we can find the Odds Adjusted mean, we know that we need to calculate a two standard deviation change where $\Delta x =$ one standard deviation. The following formalizes two different Odds for one deviation in either direction.

$$\begin{gathered} x_1 = x_0 + \Delta x \\ \rightarrow ln(Odds_1) = b_0 + b_1(x_0 + \Delta x) \\ \rightarrow ln(Odds_1) = b_0 + b_1x_0 + b_1\Delta x \\ \rightarrow ln(Odds_1) = b_1\Delta x \end{gathered}$$

As we are looking to include the probability difference for a standard deviation change in either direction, we need to formalize a similar change as the one above, but for $x_2 = x_0 - \Delta x$. This will give us a new Odds Ratio as $Odds_2$, and with the same process as before we find $ln(Odds_2) = -b_1\Delta x$.

Now to find $Odds_1 + Odds_2$ we see that $ln(Odds_1) + ln(Odds_2) = ln(Odds_1 \times Odds_2) \rightarrow b_1\Delta x - b_1\Delta x = 0$. Therefore we can conclude that $Odds_1 \times Odds_2 = 1$. With $Odds_1 = \frac{P_1}{1-P_1}$ and $Odds_2 = \frac{P_2}{1-P_2}$, we know that $\frac{P_1}{1-P_1} \times \frac{P_2}{1-P_2} = 1$. Through more algebra we are able to find the relationship between P_1 and P_2.

$$\begin{gathered} \frac{P_1}{1-P_1} \times \frac{P_2}{1-P_2} = 1 \\ \rightarrow \frac{P_1}{1-P_1} = \frac{1-P_2}{P_2} \\ \rightarrow P_1P_2 = (1-P_1)(1-P_2) \\ \rightarrow P_1P_2 = 1 - P_2 - P_1 + P_1P_2 \\ \rightarrow P_2 + P_1 = 1 \end{gathered}$$

To find the $OddsAdjusted$ (OR_a) in terms of $Odds_1$ and $Odds_2$, we simply have $OR_a = \frac{Odds_1}{Odds_2}$. From there we'd like to be able to find $P_1 - P_2$, or the difference of probability of winning after a two standard deviation change for the test statistic. Below we will look to calculate that by using our Odds Adjusted values.

$$\begin{gathered} OR_a = \frac{Odds_1}{Odds_2} = \frac{Odds_1}{(1/Odds_1)} = (Odds_1)^2 \\ \rightarrow (Odds_1)^2 = (\frac{P_1}{1-P_1})^2 = OR_a \\ \rightarrow \sqrt{OR_a} = (\frac{P_1}{1-P_1}) \\ \rightarrow \sqrt{OR_a}(1 - P_1) = P_1 \\ \rightarrow \sqrt{OR_a} - \sqrt{OR_a}(P_1) = P_1 \\ \rightarrow \sqrt{OR_a} = P_1 + \sqrt{OR_a}(P_1) \\ \rightarrow \sqrt{OR_a} = P_1(1 + \sqrt{OR_a}) \\ \rightarrow P_1 = \frac{\sqrt{OR_a}}{\sqrt{OR_a}+1} \end{gathered}$$

Again we are looking to be able to calculate the change in probability of a success ($P_1 - P_2$) from our adjusted Odds Ratio. We can then find P_2 as $P_2 = 1 - P_1$. After substituting in $\frac{\sqrt{OR_a}}{\sqrt{OR_a}+1}$ for P_1 and $\frac{\sqrt{OR_a}+1}{\sqrt{OR_a}+1}$ for 1, we have $P_2 = \frac{\sqrt{OR_a}+1}{\sqrt{OR_a}+1} - \frac{\sqrt{OR_a}}{\sqrt{OR_a}+1} = \frac{1}{1+\sqrt{OR_a}}$. It is from there that we can conclude the difference in the probability of winning from a two standard deviation change in the value of a match statistic can be written as formula (6). By performing this calculation for each adjusted Odds Ratio and bounds, we then have our MBI interval as a proportion, only to multiply the result by 100 to get it as a total percentage.

$$P_1 - P_2 = \frac{\sqrt{OR_a} - 1}{1 + \sqrt{OR_a}} \tag{6}$$

For example, shots on target (SOT) during all games of the group stage for the 2018 tournament had two standard deviations of about 4.502 with the lower bound of the 90% confidence interval at 0.112. The OR_a is equal to $e^{4.502 \times 0.112}$ from formula (5). This gives $OR_a = 1.657$. From there we can apply it to formula (6) such that the MBI lower bound for SOT in all games is $100 \times \frac{\sqrt{1.657}-1}{1+\sqrt{1.657}} = 12.555$.

3 Results

Relationships between each match statistic with a two standard deviation increase and the end result of the match can be seen in figure (5) with proportion rates within categories for each of the intervals found in table (1). Just as in [3], a 10% change in a team's likelihood of winning will be considered as the threshold for the smallest worthwhile change. We will consider a match statistic as significant if a majority of the proportion of the interval is past one threshold of the smallest worthwhile change, and does not pass both thresholds.

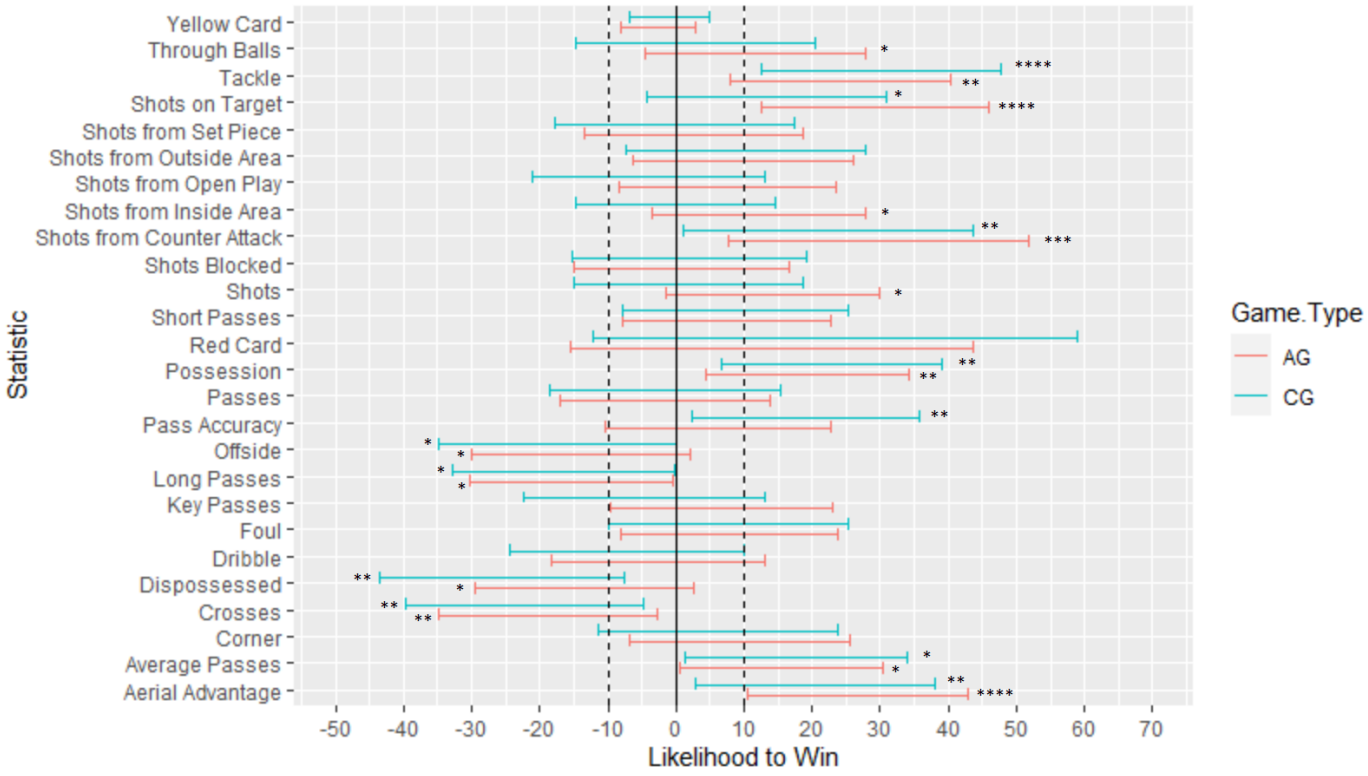

Figure 5: Results from a Magnitude Based Inference analysis of 26 match statistics and their effect on a team's end result for group stage games in the 2018 FIFA World Cup. AG = All Games, CG = Close Games. Asterisks indicate the likelihood for the significance of a match statistic having an effect on the end result of a game. * = possible, ** = likely, *** = very likely, **** = most likely.

At first glance of figure (5), there seems to be two outliers. Red card is the only match statistic that can be ruled as inconclusive. While yellow card is the only match statistic that remains completely trivial. It doesn't pass the thresholds for both harmful or beneficial effects in either all games or close games.

Looking at all 48 games from the group stages, we see that shots on target, shots from a counter attack, aerial advantage, possession, and tackles had a clearly significant effect towards improving a team's chance of winning the match. By table (1) for all games, shots

on target and aerial advantage can be written as *most likely* ($> 99.5\%$) beneficial. We can also conclude that shots from counter attacks, possession, and tackles are *likely* ($75 - 95\%$) beneficial. Other notable match statistics that could have a potential benefit for teams in all games are shots, shots from inside the area, through balls, and average pass streak. These match statistics can be considered as *possibly* ($25 - 75\%$) beneficial contributors.

Regarding potentially harmful effects in all games, crosses seems to be the only clear, significant match statistic with the 77% of the interval qualifying as harmful and a *likely* indicator. Other notables are long passes, dispossessed, and offsides, where each can be stated as *possibly* a harmful indicator.

The 2018 tournament resulted in 40 close games during the group stage. This provides us with a sample size of 80 observations (16 less than all games) for that test group. For beneficial effects, there are five match statistics that seem to be clearly significant in close games. Tackles were found to be most likely beneficial for close games. With shots from counter attack, possession, pass accuracy, and aerial advantage all considered as *likely* beneficial. It can also be said that shots on target and average pass streak (average passes) have a *possible* benefiting effect on a team's result in close games.

For potentially harmful indicators in close games, there are two match statistics that resulted in having a clearly significant effect. Both dispossessed and crosses are *likely* harmful for a team's likelihood to win in those competitive games. We also found that during close games in 2018, long passes and offside have a *possible*, harmful effect for a team's chances of success.

MBI Proportions of Effects						
Match Statistics	**(AG)** Harmful Effect	**(AG)** Trivial Effect	**(AG)** Beneficial Effect	**(CG)** Harmful Effect	**(CG)** Trivial Effect	**(CG)** Beneficial Effect
Shots	0.0	0.363	0.637 *	0.15	0.592	0.258
Shots on Target	0.0	0.0	1.0 ****	0.0	0.406	0.594 *
Shots Blocked	0.156	0.635	0.209	0.152	0.58	0.268
Shots from Open Play	0.0	0.578	0.422	0.327	0.587	0.086
Shots from Set Pieces	0.105	0.626	0.269	0.219	0.571	0.210
Shots from Counter Attacks	0.0	0.053	0.947 **	0.0	0.209	0.791 **
Shots from Inside the Area	0.0	0.429	0.571 *	0.160	0.685	0.155
Shots from Outside the Area	0.0	0.504	0.496	0.0	0.492	0.508
Ball Possession	0.0	0.185	0.815 **	0.0	0.1	0.9 **
Passes	0.224	0.648	0.128	0.249	0.592	0.159
Pass Accuracy	0.010	0.604	0.386	0.0	0.229	0.771 **
Long Passes	0.679 *	0.321	0.0	0.703 *	0.297	0.0
Short Passes	0.0	0.584	0.416	0.0	0.539	0.461
Through Balls	0.0	0.45	0.55 *	0.133	0.567	0.3
Average Pass Streak	0.0	0.318	0.682 *	0.0	0.264	0.736 *
Key Passes	0.0	0.6	0.4	0.348	0.568	0.084
Dispossessed	0.606 *	0.394	0.0	0.936 **	0.064	0.0
Crosses	0.772 **	0.228	0.0	0.853 **	0.147	0.0
Dribbles	0.263	0.636	0.101	0.421	0.579	0.0
Offsides	0.622 *	0.378	0.0	0.709 *	0.291	0.0
Corners	0.0	0.522	0.478	0.037	0.57	0.393
Aerial Advantage	0.0	0.0	1.0 ****	0.0	0.205	0.795 **
Tackles	0.0	0.06	0.94 **	0.0	0.0	1.0 ****
Fouls	0.0	0.567	0.433	0.0	0.564	0.436
Yellow Card	0.0	1.0	0.0	0.0	1.0	0.0
Red Card	0.568	0.338	0.094	0.689	0.281	0.03

Table 1: Here are the proportions for each Magnitude Based Inference Interval of the match statistics that lies within zones declared as either having harmful, trivial, or beneficial effects. AG = All Games and CG = Close Games. Chances of significance being marked as * - possibly, ** - likely, *** - very likely, and **** most likely.

4 Discussion

When it comes to the red cards being found as inconclusive, this is most likely due to the limited data for this event during the group stage. There were only three total red cards during the whole group stage of the tournament. For yellow cards, treating 1 as two standard deviations most likely altered the results. When looking at the data for yellow cards within the group stage, calculating two standard deviations actually gives us 2.746. In order to hold the methods the same as in [3], two standard deviations for yellow and red cards was kept at 1 for these results. For future tests, using the calculated two standard deviations for yellow and red cards within the group stage may provide better results.

Since Red Cards are such a rare occurrence, it could be best to analyze it with matches that ended with red cards only of a much larger sample size, but the issue being there are even more variables that could affect the results. The time at which the ejection was issued during the match could determine how much of an impact it plays on the final score. With a Red Card issued to a team during the first ten minutes being more significant than the last ten minutes, as well as the scoreline at that point in the game. Another factor could be potential psychological effects it could have on players, as well as tactical changes both teams could make. As the team with less players will likely have to play a more defensive game, while the team with the player advantage has more space to work with.

During the 2018 FIFA World Cup, there were some clear changes in significance for match statistics between all games and close games of the group stage. From all games to close games, there was a drop of significance for previously significant beneficial match statistics such as through balls, shots on target (SOT), shots form inside area, shots from counter attack, and aerial advantage. With SOT having the biggest drop, falling from *most likely* to *possibly* with a 41% decrease in the proportion past the +10% chance of success threshold. Aerial advantage went from *most likely* a significant indicator to *likely*. Shots from counter attacks did drop from a 94% to 79% past the beneficial threshold, but stayed significant as a *likely* benefiting statistic. Both shots from inside area and through balls fell from being a *possible* benefit, to trivial in close games.

There were some increases for significantly beneficial match statistics from all games to close games. Pass accuracy and tackles were the two match statistics that saw a positive increase in their effects for such games. While pass accuracy was unclear, or *possibly* trivial in all games, it became *likely* beneficial in close games. Tackles went from *likely* beneficial in all games, to *most likely* with 100% of the interval past the beneficial threshold.

Close games saw only increases in significance for harmful match statistics in all games. Dispossessed had the greatest increase going from *possibly* harmful to *likely* a harmful effect, with 93% of the interval past the harmful threshold. All of long passes, crosses, and offsides had slight increases in the harmful proportion of their respective intervals, but all held their respective significance label. With the long passes and offsides remaining *possibly* harmful, and crosses as *likely*.

Looking back to the 2014 study results displayed in figure (6), we find similar results to those of this study for Shots, Shots from Counter Attack, Through Ball, and Average Pass Streak. All held the same level of significance for all games and close games from the 2014 and 2018 tournaments.

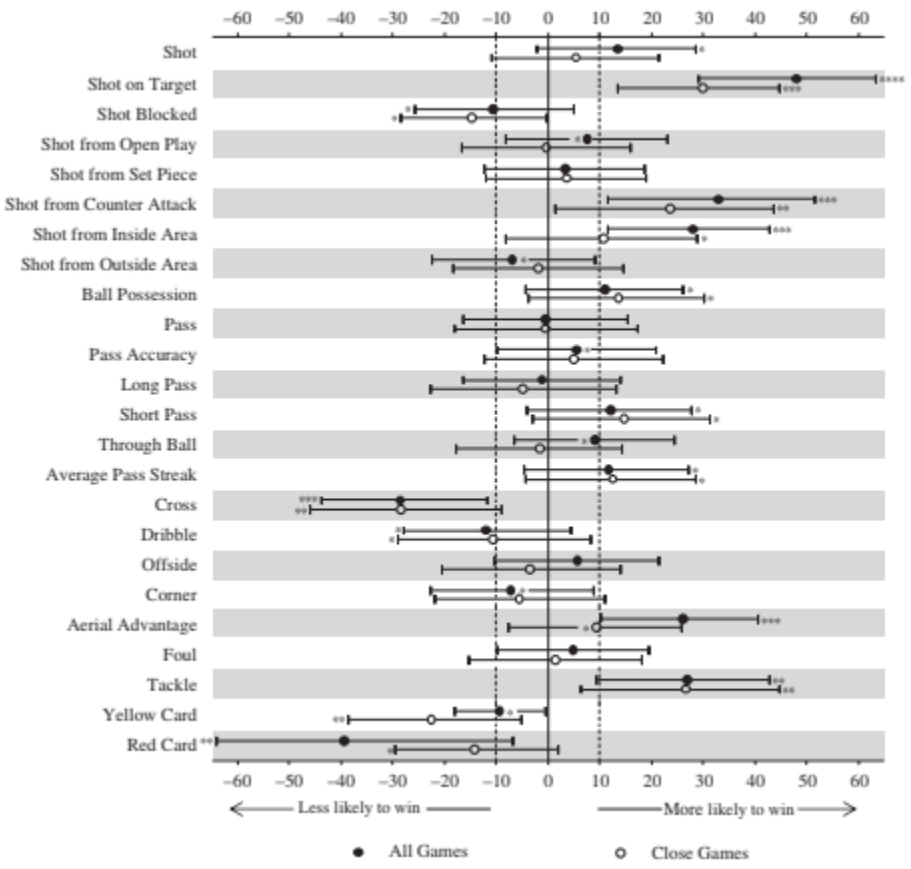

Figure 6: Results from a Magnitude Based Inference analysis of 26 match statistics and their effect on a team's end result for group stage games in the 2018 FIFA World Cup. AG = All Games, CG = Close Games. Asterisks indicate the likelihood for the significance of a match statistic having an effect on the end result of a game. * = possible, ** = likely, *** = very likely, **** = most likely.

Shots from inside area, ball possession, short passes, dribbles, red cards, shots on target, and yellow cards all had a drop in significance from 2014 to 2018. In 2014, shots from inside area was found to be *very likely* beneficial in all games and *possibly* a benefit in close games. For 2018, shots from inside area was only *possibly* beneficial in all games and trivial in close games. Short passes were recorded as *possibly* beneficial for all games and close games in 2014, yet in 2018 became trivial for all games and close games. A similar instance occurred with dribbles and Yellow Cards. Dribbles were found to be *possibly* harmful in all games and close games, while yellow Cards were *possibly* harmful in all games and *likely* in close games for 2014. Yet both became trivial in 2018 for all games and close games. Red cards were seen as significantly harmful in 2014, but became inconclusive in 2018. Shots on target

stayed *most likely* beneficial in all games form 2014 and 2018, but saw a large drop from *very likely* in 2014, to *possibly* beneficial for 2018 close games.

Findings from the 2018 World Cup showed some increases in significance for pass accuracy, long passes, offsides, ball possession, shots on target and tackles. Both long passes and offsides posted trivial results in 2014 for all games and close games. However, in 2018 they both resulted in *possibly* harmful for a team in all games as well as close games. The accuracy of a team's passing was another that was seen as trivial in 2014, but saw an increase in significance in 2018 for close games only, where it *likely* has a beneficial effect. In 2014, tackles resulted as a *likely* benefit for a team in all games and close games. Similar results for all games were found in 2018, but in close games it increased to a *most likely* beneficial effect. For ball possession, the 2018 tournament saw an increase to *likely* beneficial for all games and close games of the group stage.

Two match statistics that were added to the study of the 2018 FIFA World Cup, key passes and dispossessed. Key passes resulted as a trivial effect in both all games and close games. Dispossessed on the other hand was only *possibly* harmful during all games, while becoming *likely* harmful in close games. The results for key passes could be tied with the significance for each shot statistic since a key pass is any pass leading to a teammate's shot. Be that the shot metrics weren't as significant in 2018, it could be that key passes may have held more significance in 2014. With the increase in significance for dispossessed in close games, it seems to go with an idea for games with a tight scoreline. If one or two moments could decide the match or tip it in favor of one team, then mistakes become even more costly. Looking at the statistics that dropped as well as gained significance from all games to close games, you see an emphasis on mistakes from a team. The only type of shots that remained *likely* significant or greater in 2018, were shots from counter attacks. This is while pass accuracy became *likely* beneficial and Tackles increased to a *most likely* benefit. Teams that were more successful with their passing, attempted to dispossess their opponent (tackle) more, while creating more shots from counter attacks, and being dispossessed less, had a better chance of winning.

Looking at the results from the past two World Cups, a surprising negative factor may be crosses. This is an event that had an average of 15.4 occurrences per team for a match within the group stage. It is a popular option for a team's attack, in the hopes of catching the defense out of position or a player open in front of goal. When seeing how beneficial an aerial advantage can be, this begins to make more sense. Crosses can be the biggest gamble when it comes to creating a goal scoring opportunity since it is often a long pass played in the air (sometimes on the ground) where defenders and attackers battle or race to the ball hoping to win it first. All the defender needs to do is get it away from their own goal, while the attacker needs to redirect it with an attempt on goal. This, as well as long passes often result in what is referred to as a "50/50" ball since either side has players that could potentially win possession of the ball. Just because long passes and crosses can be harmful to a team's result, that doesn't mean they can't be of use. If a team is able to use them to find a player open in space, that "50/50" chance becomes significant in allowing the player to make a dangerous play. That being said, with long passes seeing an increase in the potential harm they could play on a team's chance of winning in the group stage for 2018, it seems

that a long ball and cross strategy did not go well for teams.

Going forward, results found of which match statistics can play a significant role in a team's chances of winning a game can help lead towards an improved capability of identifying key performance indicators for individual players. Data analysts within the sport are finding themselves in an opportunity to be creative in developing new match statistics to develop for players. **Statsbomb**, a soccer analytics company has developed "player radars" in recent years to provide a visual representation of player performances as well as newer match statistics recorded of those players. These player radars have also been designed to be specific for a player's position. [12]

By using logistic regression, this allows us to potentially build a model using multiple significant match statistics, to attempt to create an even better prediction of end results to matches. During the **Statsbomb** soccer analytics conference of 2019, Ryan Beal and Prof. Gopal Ramchurn presented their idea of *Valuing Player Influence Within Teams*, by quantifying the strength of on field relationships with their centrality within the network of the team and each pairs key performance indicator values of the two players. Through this they created a model for giving a best predicted lineup for a team. [13]

5 Conclusion

Sabermetrics with baseball is the founding father of sports analytics and has paved the way for other sports to invest resources into the field. From there data analysis of games began to change how top level professional organizations looked at their performance indicators for athletes. Now soccer, along with other sports, are changing how they look at their respective sport with this emerging tool. Studies such as this one, looking at key performance indicators in match statistics, allow teams to determine what style and statistics in their game play are most important for their tactics. Coupled with analysis of a team's opponent's game play can provide them with the strategy best suited for disrupting the other team's game plan. With an idea of which match statistics are important, teams can also have a better idea of how much of an impact each player has on the end result of their games. From the 2018 FIFA World Cup group stage games, it is noted that match statistics such as shots on target, shots from a counter attack, ball possession, aerial advantage, average pass streak, and tackles all could be concluded to have a beneficial effect towards a team's chances of winning the game. While offsides, crosses, dispossessed and long passes had negative effects on a team's probability of winning.

References

[1] Leigh Steinberg *Changing the Game: The Rise of Sports Analytics.* Forbes, 2015

[2] Anderson, Christopher and Sally, David *The Numbers Game: Why Everything You Know About Soccer is Wrong* Penguin, 2013

[3] Liu, Hongyou and Gomez, Miguel-Angel and Lago-Penas, Carlos and Sampaio, Jaime *Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup* Journal of Sports Sciences, Volume 33 Number 12 Pages 1205 - 1213, 2015

[4] Rosemary Pennington
G.O.A.L. Celebrating The Statistics Of The Beautiful Game Stats + Stories, June 14 2018, Podcast
`https://soundcloud.com/statsandstories/sns-058-g-o-o-a-a-a-l-l-l-l-celebrating-the-statistics-of-the-beautiful-game`

[5] Hopkins, William G. and Marshall, Stephen W. and Batterham, Alan M. and Hanin, Juri *Progressive Statistics for Studies in Sports Medicine and Exercise Science* Medicine and Science In Sports and Exercise, Volume 42 Issue 1 Pages 3 - 12, 2009

[6] https://statsbomb.com/2018/08/new-data-new-statsbomb-radars/

[7] Castellano, Julen, David Casamichana, and Carlos Lago.
The use of match statistics that discriminate between successful and unsuccessful soccer teams Journal of human kinetics 31 (2012): 137-147.

[8] *What Is Logistic Regression?*
`https://www.statisticssolutions.com/what-is-logistic-regression/`

[9] Stephanie *Log Odds: Definiton and Worked Statistics Problems*
`https://www.statisticshowto.datasciencecentral.com/log-odds/`

[10] Magdalena Szumilas *Explaining odds ratios* Journal of the Canadian academy of child and adolescent psychiatry 19.3 (2010): 227.

[11] *gretl*
`http://gretl.sourceforge.net/`

[12] Ted Knutson *New Data, New Statsbomb Radars*
`https://statsbomb.com/2018/08/new-data-new-statsbomb-radars/`

[13] Beal, Ryan, and Ramchurn, Gopal. *Valuing Player Influence Within Teams* StatsBomb Innovation in Football
`https://youtu.be/USiRTCiPpqw`