Spring 2019

Soccer Analytics & its future

Sunil Srinivas Sukumar

Moneyball or Moneyfall? Current State of Analytics in Soccer and What Future Holds

Creative Component presented to the faculty of
Management Information Systems

Iowa State University

In partial fulfillment of the requirements for the
degree of Master of Science

Program of Study Committee:
James A Davis
Russel N Laczniak

By

Sunil Srinivas Sukumar

Table of Contents

Contents

Acknowledgement

I would like to acknowledge everyone who played a major role in my academic accomplishments to date. I would like to specially thank my committee Dr. Russell N. Laczniak and Dr. James A. Davis for their continued support throughout this research. I would also like to thank Dr. Anthony M. Townsend for his patient advice and guidance from the beginning of this research study. I would also like to thank Dr. Abhay Mishra for his guidance on how to write a research paper. Finally, I would like to thank my parents and sister, who supported me with complete love and understanding from the very beginning. In addition, I would like to thank my friends, the department faculty and staff members like Debbie Johnson for making my time at Iowa State University a memorable experience. Thank you all for your unwavering support.

Abstract

Researchers in the field of sports analytics have studied the effects of analytics on sports and specifically how it could be leveraged to constantly improve the game and make it beneficial to everyone involved. While this has been successful to an extent in sports like baseball, basketball, etc., it has not been immensely successful in soccer. Hence, this research aims to study the current state of analytics in soccer, how the analytical models are being used to predict the outcome of soccer games and suggest solutions to make the existing prediction models better so that it would benefit various groups like sport analysts, managers of soccer clubs, coaches, etc. Later, the paper will also discuss the future and implications in brief.

Introduction

Analytics in the field of sports has always existed, with or without us knowing it. Sometimes, coaches make unbelievably nonsensical decisions and the same coaches, at other times make decisions that put most brilliant strategic minds to shame. This leads to a lot of questions about what goes through in their minds when they decide on a substitution or reacting to a situation at a point in a game. There are hundreds of questions like these that people have pondered over for years together until the advent of sports analytics.

What is sports analytics? One of the best definitions of the term is as quoted by Masoud Nikravesh in his article, "Sports analytics is the process that identifies and acquires the knowledge and insight about potential players' performances based on the use of a variety of data sources such as game data and individual player performance data. These advanced and sophisticated type of analytics should be able to extract valuable actionable insights for the coaches and managers to utilize (1)." In addition, recent advancements in data collection and management technology has broadened the scope of sports analytics significantly.

To be specific, soccer analytics is the art of creating meaningful insights and decision that can be acted upon using soccer related data. The data can be anything ranging from how many goals a team has scored to multiple factors like, how much distance a single player in a team has covered, or how many passes he has played and how many out of those were misplaced, how many out of those created chances for the team to score etc. In soccer, both predictive and descriptive analytics is used. While predictive analytics predicts the possibility of an outcome, descriptive analytics analyzes the data in hand to come up with suggestions to increase the possibility even further. Hence, without good data, the analytics is almost as good as nothing.

Analytics have many on-field applications in a sports environment, including managing both individual and group performance. Coaches can use data to optimize exercise programs for their players and develop nutrition plans to maximize fitness. Analytics are also commonly used in developing tactics and team strategies. With thousands of games worth of data to study, analysts can look for patterns across a broad sample size regarding formation, counter strategies and other key variables [(2)].

With growing access to unimaginable volumes of data & technology, a good number of teams across various sports have begun employing analytics to their benefit. Different techniques of analytics, mainly predictive analytics has taken center stage recently, especially when it comes to predicting performance of players, teams and managing teams based on the analytics. "Since most professional sports teams function as businesses, they are always seeking ways to improve sales and reduce expenses across their organization. Some sports analysts specifically focus on issues regarding the marketing and sale of sports tickets and team merchandise. Modern marketing and fan outreach efforts also rely heavily on analytics to predict their consumer base and identify opportunities to increase brand engagement.[(2)]"

Out of all these applications, as one would expect, analytics works the best in business and management to help teams find new players, manage their physical conditions rather than trying to predict the outcome of games based on historic data and the current data available before a game. A company, named 21st Club, even assists clubs scout for young and exciting talents and assist the teams to buy them for reasonable prices in this growing age of inflated transfer market. Improved player recruitment is the most popular application of analytics in soccer, and to an extent, successful as well. On an interview, the co-founder of the company, Blake Wooster, says, "We're all about trying to help teams understand their identity, their strategy and what they're trying to achieve. We might be dealing with a club, let's say in the bottom half of the Premier League, and their aspiration is to be a club in the top 10 or the top six. In a very consultative way, we'll apply an analytics and research team to go away and find out what a top 10 club in the Premier League looks like. You work out what the gaps are between where they are today and where they want to be and try to put in strategies to help them close the gap. Those strategies could be anything from player recruitment to changing the head coach or changing the style of play with the existing head coach."

This, to an extent, can be compared to the 2011, award-winning Hollywood movie, that highlighted the importance of sports analytics to the entire world, "Moneyball". The movie shows how analytics is used in baseball to analyze players and come up with a strategy to buy the based on what the team needs and how each player can contribute to the team to a winning cause. Soccer, the most widely followed sport in the world has also started adopting predictive analytics for the aforementioned reasons, however, the problem of soccer lies elsewhere. It has not been nearly as successful as other sports in predicting the performance of players or predicting the outcome of games. For example, the world cup 2014 predictions: there were different predictions, but almost every single model had Brazil as their world cup winner with Germany finishing runners-up. However, while Germany went out in the semi-finals, Brazil did not even reach semi-finals. France, who were predicted to finish fourth won the world cup, while, Croatia, who were predicted to go out in the round of 16 reached the finals.

The failure in these models leads one to think about the reasons and how could the models be made better. It could partly be because soccer is dynamic and has a multitude of factors that are difficult to measure that contribute to the outcome of a game, which builds up to the research questions in this paper. For example, in the semifinal game of the 2014 World cup, Brazil's best players, Neymar (who was out with an injury) and Thiago Silva (who had received a red card the previous game) were out and the models did not take these factors into account while predicting the results.

In order to understand why analytics does not have as much impact on soccer as it has on other sports, the existing models must first be analyzed and understood. Then, the variables that contribute the most to these models must be discovered and then analyzed. Hence, this paper aims at diving deep into all the factors that are used in a model to predict results and suggest ways to make the prediction models better (from approximately 70% to more than 90%). Since the current models are very complex, this paper aims to take out a few important factors, attempt to find out the significance of these factors in predicting the outcome of a game with the help of some data and research. Moreover, this paper will attempt to provide some insights on how soccer analytics will move in the future based on other researches and facts available.

Research Questions

As already discussed, this paper intends to address the gaps that previous researches have failed to answer or failed to provide further explanations. Hence, the following questions would guide the research paper:

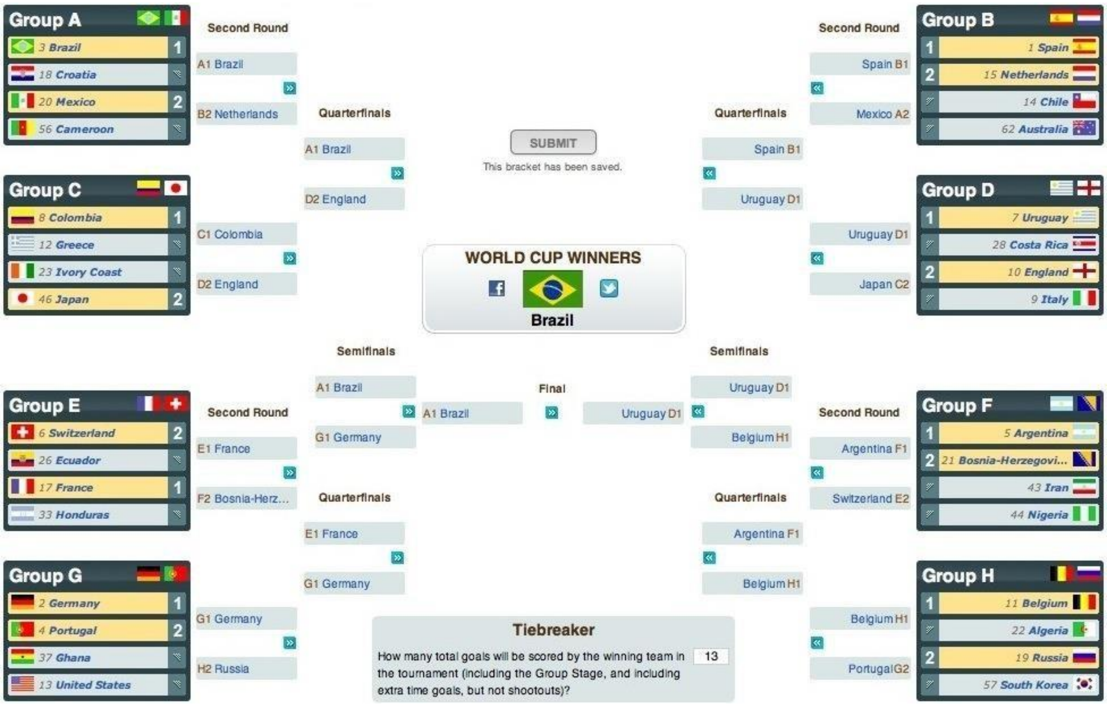
- RQ1. Why is predictive analytics not as successful in soccer as compared to other sports?
- RQ2. What are some factors that really influence the outcomes of a soccer game?

While these research questions might seem generic and broad, the research model aims to dive deep into each of the factors and try to come up with a solution for each of the questions that could be explored under these main questions. The focus of this research and research questions could further be narrowed down based on some literature review on the topic.

Literature Review

As discussed, the purpose of this study is to delve in depth to find out the reasons for analytics being less successful in soccer and identify factors to help improve the existing models. This could be made possible only by understanding what previous papers have attempted in this area and try to identify gaps in the area of study. An excerpt from the article written for a webpage, Fansided, says: "Soccer, then, has come increasingly to seem like the final frontier of the analytics movement. If soccer falls, there will be no worlds left to conquer. The analysts may sit, and weep. There are good reasons for this: soccer is low-scoring, fluid, recordable actions are relatively scarce, there seems occasionally to be a direct correlation in certain positions, mostly center-back, between being good and not doing *anything* (in the words of the great Paolo Maldini, "If I have to make a tackle then I have already made a mistake") and, above all, stats other than goals have only really been recorded for the past 25 years. There are also bad reasons for this, like that some people blame good analysts for bad analysis, and other people are afraid to admit the shortcomings of their own, non-statistical expertise." This just highlights how much is still left unconquered by analytics in soccer. This also underlines the fact that soccer is not just a numbers sport like other sports, but, is a dynamic sport with various factors affecting a game at any given point in time. This also shows how difficult of a sport soccer is to turn into statistics. As someone said, it isn't a series of individual events like football or baseball, and there isn't as high a volume of shots, assists, turnovers and so on as there is in basketball and hockey.

The papers written in this area try to study predictions that have been made by different models and compare it with actual results and then they try to study the models. For example, the world cup 2014 and the model used to predict the results. A paper written on Umbel.com, in 2014 shows how bad the predictions were, especially that of Goldman Sachs, and how it was not even 60% accurate. Figure 1.0 shows exactly how the prediction models failed to predict the outcome of games. The article says, "Nate Silver, the celebrated statistician behind the FiveThirtyEight blog, had pegged Brazil as the favorites to win the cup and slated their chances of success against Germany at 65% (3)".



Group A	
3 Brazil	1
18 Croatia	
20 Mexico	2
56 Cameroon	

Group B	
1	1 Spain
2	15 Netherlands
	14 Chile
	62 Australia

Group C	
8 Colombia	1
12 Greece	
23 Ivory Coast	
46 Japan	2

Group D	
1	7 Uruguay
	28 Costa Rica
2	10 England
	9 Italy

Group E	
6 Switzerland	2
26 Ecuador	
17 France	1
33 Honduras	

Group F	
1	5 Argentina
2	21 Bosnia-Herzegovi...
	43 Iran
	44 Nigeria

Group G	
2 Germany	1
4 Portugal	2
37 Ghana	
13 United States	

Group H	
1	11 Belgium
	22 Algeria
2	19 Russia
	57 South Korea

How many total goals will be scored by the winning team in the tournament (including the Group Stage, and including extra time goals, but not shootouts)? 13

Figure 1.0: Models predicting Brazil as the champions of WC 2014 while Uruguay was supposed to reach finals

While Brazil went down in the semifinals of the tournament, the other predicted finalists, Uruguay did not even reach semifinals. The article also said, "As you may remember, Goldman Sachs economists crunched data on 14,000 past matches to arrive at a 67-page report that confidently calculated the outcomes of all matches, published May 30. The bank changed 50% of the teams after the first stage once it became clear that reality wasn’t quite matching the historical dataset. Looking at those original, pre-tournament predictions, only 37.6% of the group stage matches ended the way Goldman Sachs forecast. That is, it correctly calculated one of three possible outcomes ― win, lose or draw ― just over a third of the time, hardly better than predicting it at random.” This shows historical dataset is not a great factor in deciding the outcome of world cup games and it needs much more.

Worldsoccertalk, in 2012 said exactly what the problem is when it comes to analytics in soccer: “Not only this, but the two sports are, by nature, worlds apart. Football is much more fluid; each player’s performance is dependent on the play of others. A striker can’t score unless he is provided with service from supporting players. A goalkeeper can’t keep a clean sheet without the help of his defense in front of him. Baseball, on the other hand, is a more structured game. Each play follows the same basic format and results in players being either credited or debited [(4)]”.

Another article in Unprofession (2017), tried to find out how the current models are setup and why they don’t work well, “The most recent prediction model (released on January 21, 2017) was created by Nate Silver and Five Thirty Eight, and it draws heavily on Expected Goals—how many goals a team is expected to score in a game based on the quality and angle of the shots they took (and a few other factors). In 2009, Silver came up with SPI (Soccer Power Index), which assigns each team an offensive and defensive rating—how many goals they’d be expected to score and concede against an average squad. SPI updates itself independent of results, as teams can win and still have their ratings lowered if they play poorly. Silver’s

2017 model is an updated version of his 2009 model. It updates the thinking surrounding Expected Goals by considering more match events and the circumstances surrounding goals and stats. It takes into account four specific metrics when ranking a team: goals, adjusted goals, shot-based expected goals, and non-shot expected goals [(5)]."

With the help of these research and evidences, I was able to find key themes and use it as a base for this research & assist me frame my research propositions and hypotheses, which are as follows:

1. Soccer analytics is currently circled mostly around goal scoring and not a lot of other factors.
2. There are less quantifiable events in soccer as compared to other sports.
3. Soccer is more qualitative than quantitative, hence, being good sometimes cannot even be measured.

Research Model & Hypotheses

The research studies in this area are mostly concentrated on either acknowledging the fact that the current models are not good or trying to build small, real time models that follow prescriptive analytics by looking at historical data to observe and draw out patterns. This research intends to find out the reasons as to why the models fail and figure out the most important variables that are required for a model to be successful. Hence, the propositions and the hypotheses below would help us to further narrow down on the research questions discussed earlier, which are a mix of qualitative and quantitative studies.

Hypotheses for Research Question 1:

- H1a: The chances of winning a game remains unaffected even if a team scores a lot of goals
- H1b: The chances of winning a game at home remains unaffected even if a team scores a lot of goals

Propositions for Research Question 2:

- Quality, current form of the teams are important factors in the analytics models
- Player fitness, possession, playing sequence & style also contribute to the models

Here, the first two hypotheses intend to answer the first research question. The first hypothesis tries to understand if scoring goals alone is a great influence in predicting the outcome of a game, which is how the current predictive models are setup. The second hypothesis is framed to show home advantage, when acted upon as a moderating variable, has an effect on scoring goals, showing that, a variable when acted upon by another moderating variable has a greater effect, however, would not be the best model.

With the help of the next two propositions, I intend to answer the second research question. Some factors like quality of a team, which is not easily measurable, and the current form, which has traditionally been overlooked are included in the first proposition to prove that, when these factors are given more weightage, the predictive model could become a little more accurate. The final proposition adds a moderating variable in team chemistry, which again is not easily quantifiable, when acted upon as a moderating variable over the other variables, further improves the performance of the model. With the help of these propositions, this research also tries to provide some recommendations for future models.

Research Methodology and Variables

For this research, I intend to work with the data and feeds already collected. I intend to use the data from the English Premier league’s last three seasons. This data was collected from sources, like Opta Sports, premierleague.com while the majority will be used from the data already collected by Opta Sports. More information about the data sources can be found in Appendix 1.

This research intends to find the success of current models based on qualitative research and come up with new factors and use it on existing models to test the validity of those models. Hence, the major goal would be to find the relationship between some variables and the final outcome of the games, which, here, would be the independent variable. To find this, I consider multiple variables such as below for both quantitative and the qualitative study:

- Number of goals scored,
- Quality of team,
- Player form,
- Team chemistry & Morale,
- Team form,
- Playing conditions,

- Opposition’s quality & form,

- Historical results of the game,

- Number of player features/attributes not limited to nationality, position, style, etc.

Some of these variables will be used to find their significance in predicting the outcome of the games. Most of the models currently being used are regression models to predict the outcome of games. Hence, multiple regression models would be used to test the significance of these variables at a smaller scale with the data mentioned previously. Some of the variables are unquantifiable, such as chemistry and quality of the teams, and, these variables will be used for research to explain solutions to the second research question.

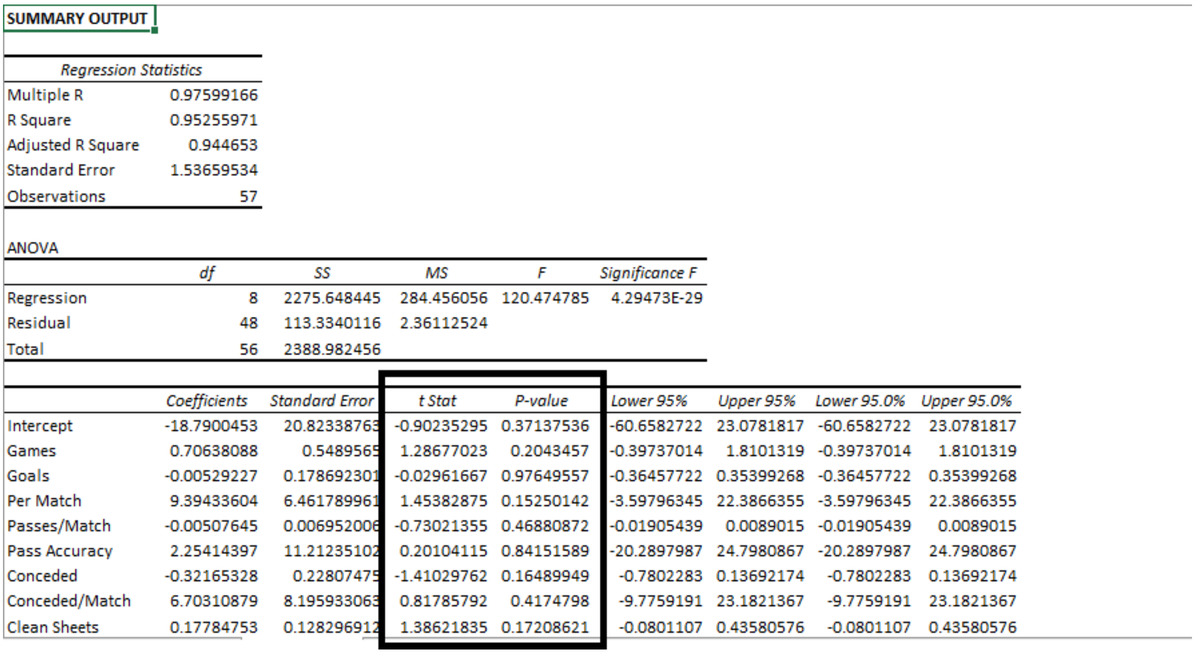
Expected Results & Future

To understand more about how the models in soccer analytics work, I collected data of all the teams from the English Premier League for the last three seasons. This was done to ensure that there weren't any aberrations in the results. While the English league has a very fast tempo and more physical, and hence, is very difficult to quantify than other leagues that are not so physical and more straightforward.

Hypotheses for Research Question 1:

- H1a = The chances of winning a game remains unaffected even if a team scores a lot of goals

  I fail to reject this hypothesis because the number of goals scored, which is an independent variable in the regression analysis, is not significant in explaining the dependent variable, which is winning games (the p value is not less than 0.05). This is based on a regression analysis shown as below:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.97599166
R Square	0.95255971
Adjusted R Square	0.944653
Standard Error	1.53659534
Observations	57

ANOVA

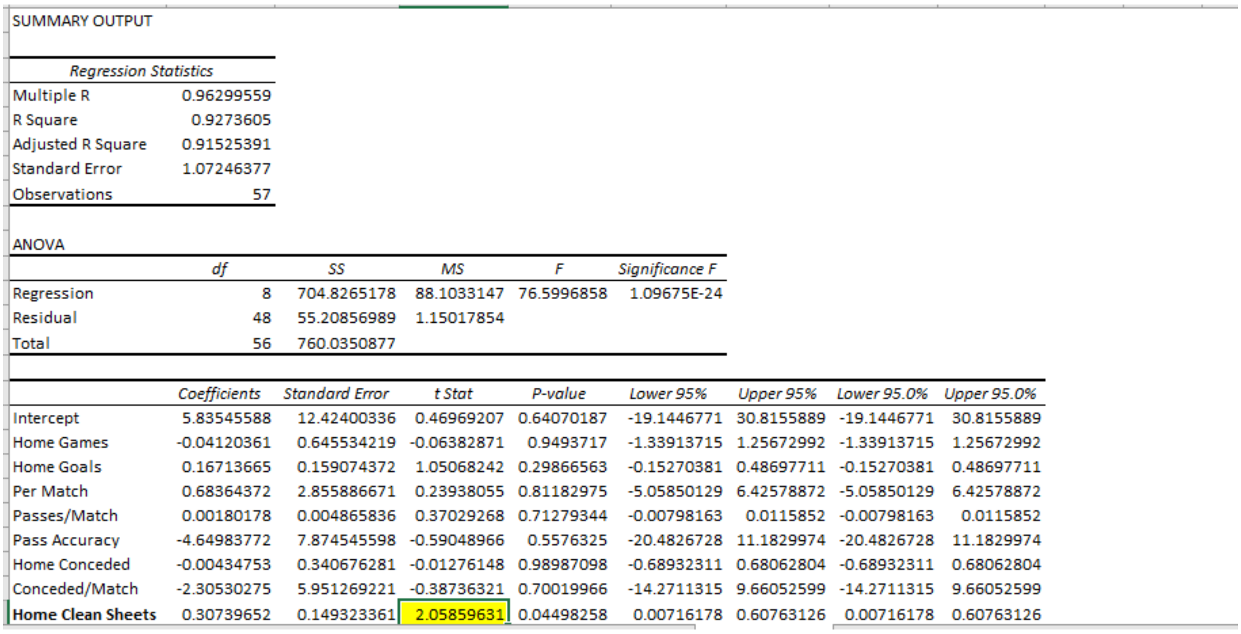
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	2275.648445	284.456056	120.474785	4.29473E-29
Residual	48	113.3340116	2.36112524		
Total	56	2388.982456			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-18.7900453	20.82338763	-0.90235295	0.37137536	-60.6582722	23.0781817	-60.6582722	23.0781817
Games	0.70638088	0.5489565	1.28677023	0.2043457	-0.39737014	1.8101319	-0.39737014	1.8101319
Goals	-0.00529227	0.178692301	-0.02961667	0.97649557	-0.36457722	0.35399268	-0.36457722	0.35399268
Per Match	9.39433604	6.461789961	1.45382875	0.15250142	-3.59796345	22.3866355	-3.59796345	22.3866355
Passes/Match	-0.00507645	0.006952006	-0.73021355	0.46880872	-0.01905439	0.0089015	-0.01905439	0.0089015
Pass Accuracy	2.25414397	11.21235102	0.20104115	0.84151589	-20.2897987	24.7980867	-20.2897987	24.7980867
Conceded	-0.32165328	0.22807475	-1.41029762	0.16489949	-0.7802283	0.13692174	-0.7802283	0.13692174
Conceded/Match	6.70310879	8.195933063	0.81785792	0.4174798	-9.7759191	23.1821367	-9.7759191	23.1821367
Clean Sheets	0.17784753	0.128296912	1.38621835	0.17208621	-0.0801107	0.43580576	-0.0801107	0.43580576

Even though a regression analysis shows a very high R squared value, a deeper look at the t-stat and the P-values show that none of the variables listed on the table is actually significant in explaining the y-variable, which is the number of games won by a team. In fact, the number of goals scored has the least absolute t-stat value. Hence, we can say that scoring more goals does not necessarily win you games.

H1b = The chances of winning a game at home remains unaffected even if a team scores a lot of goals

I fail to reject this hypothesis because the number of goals scored by teams playing at home, which is an independent variable in the regression analysis, is not significant in explaining the dependent variable, which is winning games at home (the p value is not less than 0.05). This is based on a regression analysis shown as below:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.96299559
R Square	0.9273605
Adjusted R Square	0.91525391
Standard Error	1.07246377
Observations	57

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	704.8265178	88.1033147	76.5996858	1.09675E-24
Residual	48	55.20856989	1.15017854		
Total	56	760.0350877			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.83545588	12.42400336	0.46969207	0.64070187	-19.1446771	30.8155889	-19.1446771	30.8155889
Home Games	-0.04120361	0.645534219	-0.06382871	0.9493717	-1.33913715	1.25672992	-1.33913715	1.25672992
Home Goals	0.16713665	0.159074372	1.05068242	0.29866563	-0.15270381	0.48697711	-0.15270381	0.48697711
Per Match	0.68364372	2.855886671	0.23938055	0.81182975	-5.05850129	6.42578872	-5.05850129	6.42578872
Passes/Match	0.00180178	0.004865836	0.37029268	0.71279344	-0.00798163	0.0115852	-0.00798163	0.0115852
Pass Accuracy	-4.64983772	7.874545598	-0.59048966	0.5576325	-20.4826728	11.1829974	-20.4826728	11.1829974
Home Conceded	-0.00434753	0.340676281	-0.01276148	0.98987098	-0.68932311	0.68062804	-0.68932311	0.68062804
Conceded/Match	-2.30530275	5.951269221	-0.38736321	0.70019966	-14.2711315	9.66052599	-14.2711315	9.66052599
Home Clean Sheets	0.30739652	0.149323361	2.05859631	0.04498258	0.00716178	0.60763126	0.00716178	0.60763126

This regression model shows that home goals does not play a big role in explaining the dependent variable, which is the games won. However, more clean sheets at home could translate to winning games, as evident from the model. The high R squared value could mean that there is multicollinearity in the data, however, with more variables available in the future, this issue could be avoided.

Propositions for Research Question 2:

- Quality, current form of the teams are important factors in the analytics models
- Player fitness, possession, playing sequence & style also contribute to the models

Based on current data and analytics in soccer, we could infer that some of the factors mentioned above could play a bigger role in predicting the outcome of games. Some variables like player attributes, playing style of a team, possession style and goal scoring opportunities are already being collected by some companies. These factors, when added into the models would definitely make the models more successful.

Discussion & Conclusion

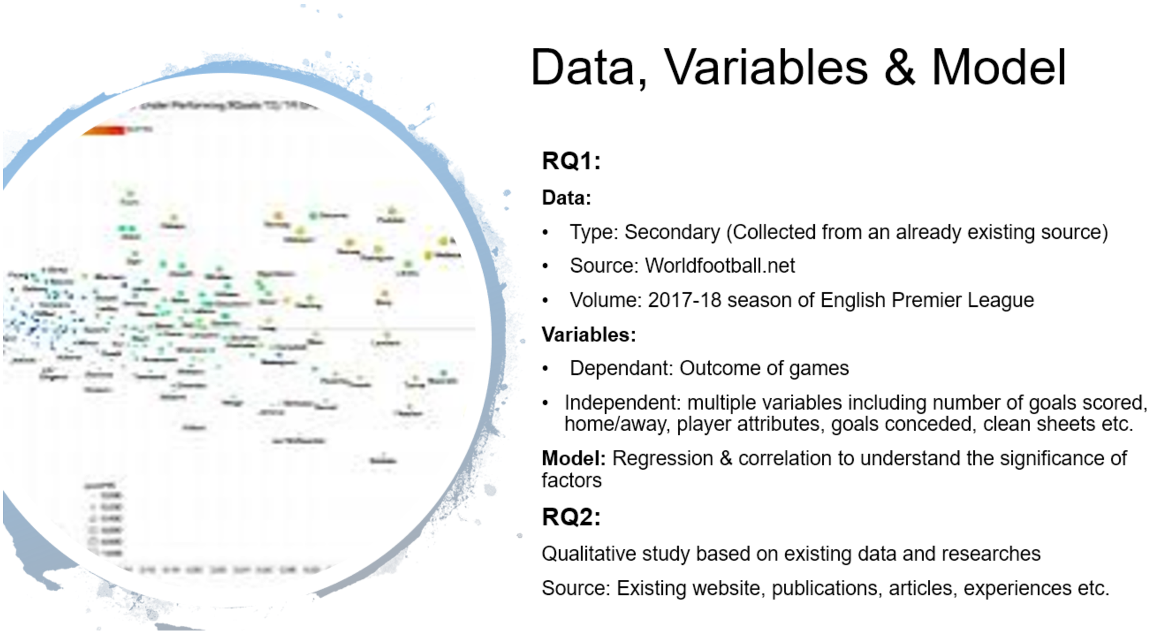
In today's fast evolving world of analytics, soccer analytics is still considered to be at a rather primitive level as compared to other sports. It has not been successful in evolving as well. As discussed in the research model, this research intends to contribute to the already existing predictive models used in the industry by sports analysts. Data collection, especially in soccer has increased multifold off late to capture various measures that have never been captured before. For example, Adidas have tied up with the Major League Soccer (MLS), in the US to have the players' vitals and neurological reactions monitored with the help of a chip that will be inserted onto the players' accessories like socks, jerseys etc. With the advent of such technologies, I believe would lead to the development of new variables that have never been used in soccer analytics before. Also, organizations like Opta Sports, is constantly trying to improve their prediction models by trying to introduce and analyze new variables such as expected goals, expected assists, sequences and possessions. For example, they assign points to each pass that could potentially be an assist based on the type of pass and then, with the help of other factors such as the player who plays the pass, the player who receives the pass, location from which the shot is tried, and other factors combined, a goal scoring opportunity is derived. Furthermore, they also consider the progress made, directness of goals scored when trying to assess the style and chemistry of a team, which goes hand in hand with this research by trying to quantify different unquantifiable measures. To conclude, predictive analytics models in soccer can be made better with the help of descriptive models that help understand the data well. As already stated in this paper, without good & meaningful data, analytics is as good as nothing.

References

- Masoud Nikravesh. Moneyball: Sports Analytics in Soccer to Predict Performance and Outcomes. Sports Analytics, Experfy, 03 MAY 2016. https://www.experfy.com/blog/moneyball-some-insights-to-soccer-analytics
- Kennemer J. The Best – and Worst – Predictions of the FIFA World Cup. Sports, Umbel, 11 JUL 2014. https://www.umbel.com/blog/sports/world-cup/
- Ben Weich. Why Moneyball Will Not Work in Soccer. Leagues-EPL, Worldsoccertalk, 06 JUN 2012. http://worldsoccertalk.com/2012/06/06/why-moneyball-will-not-work-in-soccer/
- Cornelius Arndt, Ulf Brefeld (2016). Predicting the performance of soccer players. Retrieved from: **https://doi.org/10.1002/sam.11321**
- Cassimally KA Soccer's Big Data Revolution. Scitable by Nature Education, 21 July 2012. http://www.nature.com/scitable/blog/labcoat-life/soccers_big_data_revolution
- Araújo D, Davids K, Hristovski R. The ecological dynamics of decision making in sport. Psychol Sport Exerc. 2006;7(6):653–676. doi: 10.1016/j.psychsport.2006.07.002
- Neil Paine. What Analytics Can Teach Us About the Beautiful Game. Worldcup, FiveThirtyEight, JUN. 12, 2014. https://fivethirtyeight.com/features/what-analytics-can-teach-us-about-the-beautiful-game/
- Andrea Missinato. The world of Soccer as we know it is about to be hit by Big Data tsunami. Big Data and Analytics, Spindox, 19 OCT 2017. https://www.spindox.it/en/blog/soccer-and-big-data
- Jack Pitt-Brooke. Inside the world of football analytics and how professional number crunchers are giving clubs a competitive advantage. Sport, Independent. 7 SEP 2017. https://www.independent.co.uk/sport/football/premier-league/transfer-window-football-betting-analytics-moneyball-a7934181.html

- Atanu Biswas. Sport and the arrival of big data analytics: In World Cup soccer too, statistics is now the third eye. Sports, World, Times of India. 14 JULY 2018. https://blogs.timesofindia.indiatimes.com/toi-edit-page/sport-and-the-arrival-of-big-data-analytics-in-world-cup-soccer-too-statistics-is-now-the-third-eye/
- Mikeie Reiland. Why Aren't Soccer Analytics a Bigger Deal? Unprofession, 5 FEB 2017. https://unprofession.com/why-arent-soccer-analytics-a-bigger-deal-706670ab8685
- Nick Kolakowski. Goldman Sachs World Cup Analytics Show Limits of Big Data. Big Data, World Cup, 17 JULY 2018. https://insights.dice.com/2018/07/17/goldman-sachs-world-cup-analytics-limits-big-data/
- Jure Rejec. How Big Data is Changing the World of Soccer. Smart Data Collective, 18 JAN 2016. https://www.smartdatacollective.com/how-big-data-changing-world-soccer/
- Shant Hovsepian. Five Things Soccer Analytics Teaches Us About Data Lakes. Forbes Community Voice, 20 JULY 2018. https://www.forbes.com/sites/forbestechcouncil/2018/07/20/five-things-soccer-analytics-teaches-us-about-data-lakes/#2aa5b3683ee7
- What Affects the Outcome of Football Games?, Gamblingsites, https://www.gamblingsites.com/football-betting/strategy/what-affects-outcome/
- David Sheehan. Github, 2018. https://dashee87.github.io/football/python/predicting-football-results-with-statistical-modelling/
- Rhonda Magel, Yana Melnykov. Examining Influential Factors and Predicting Outcomes in European Soccer Games, International Journal of Sports Science, p-ISSN: 2169-8759, e-ISSN: 2169-8791, 2014; 4(3): 91-96 doi:10.5923/j.sports.20140403.03

Appendix



Data, Variables & Model

RQ1:

Data:

- Type: Secondary (Collected from an already existing source)
- Source: Worldfootball.net
- Volume: 2017-18 season of English Premier League

Variables:

- Dependant: Outcome of games
- Independent: multiple variables including number of goals scored, home/away, player attributes, goals conceded, clean sheets etc.

Model: Regression & correlation to understand the significance of factors

RQ2:

Qualitative study based on existing data and researches

Source: Existing website, publications, articles, experiences etc.