Sir Alex Ferguson's Continued Tenure Could Have Only Marginal Improvements over Manchester United's EPL Performance*

Aryaman Sharma

April 16, 2024

Manchester United has been a historical football club. This paper examines Manchester United's performance from 2004-2013 under Sir Alex Ferguson's golden era and compares it with the period from 2014-2023, following his retirement. Using a Bayesian analysis model, this paper attempts to predict potential league points for if Sir Alex would have continued managing (2014-2023). Results predict only a marginal improvement in club's performance. This reasearch delves into recent increase in competition and other factor as a possible explanation.

Table of contents

*Code and data are available at: https://github.com/Ary4m3n/manchester-united.git

1 Introduction

Manchester United, an English football (soccer) team based in Manchester, England, has been one of the richest, most renowned and most supported clubs in the whole world. Founded in 1878, Manchester United is know for its note-worthy history in football. Manchester United's distinct history has been dominated by two long-serving managers, Sir Matthew Busby and Sir Alex Ferguson (*Manchester United* 2024). Sir Matthew Busby was the manager of Manchester United between 1945 and 1969, where he was most known for rebuilding the team after 23 of 44 players died after a plane crash in Munich in 1958.

Sir Alex Ferguson managed the club between 1986 and 2013 where he led the team to an unparalleled spell of dominance in the English Premier League (*Manchester United* 2024). The English Premier League (EPL) since the 1992-1993 season has been the top-tier league in the English football league system (*Premier League* 2024). Under Sir Alex Ferguson, Manchester United won 12 Premier League titles between 1992 and 2013, and dominated the rest of the 19 teams in the league. Sir Alex Ferguson is also well renowned to have nurtured the young talent in Cristiano Ronaldo, arguably now deemed to be one of the greatest players of all time.

Sir Alex Ferguson retired after the 2012-2013 season and that marked the now seen downfall of the club in recent times. Since the 2012-2013 season, Manchester United has seen 8 different managers and a significant drop in the stature of the club around trophies and wins (*Manchester United - Manager History* 2024). For context, the club won 38 trophies under Sir Alex and merely 5 trophies in total since his retirement under 8 different managers (*How Many Trophies Have Manchester United Won?* 2023). Additionally, the club has not won the

English Premier League since Sir Alex's last season, 2012-2013, whereas they won 13 titles when under Sir Alex.

The aim of this paper is twofold, where in the paper will first analyze the differences in performance between 2004-2013 (Sir Alex's era) and 2014-2023 (Post Sir Alex's era). Then, we will use this knowledge to build two Bayesian Models to predict and further analyze the **estimand**: Average League Points in each season for 2004-2013 and 2014-2023. The paper will also illustrate *how much* better Manchester United's position would have been in 2014-2023 if Sir Alex Ferguson would not have retired, helping us contribute to the debate about the greatest manager of all time.

The structure of this paper comprises four sections: Data, Model, Results and Discussion. In the Data section (Section 2), we discuss the data source and the process of measuring and cleaning the datasets. In the Model section (Section 3), we discuss the two Bayesian Models used in the paper, their justifications and how it was constructed. This section will also touch on the process of predicting the league points for 2014-2023 if Sir Alex were managing the club then. In the Results section (Section 4), we delve deeper into the trends observed. Finally, in the Discussion section (Section 5), we discuss the possible factors around contributing to Sir Alex's success, along with limitations and further research.

2 Data

The datasets under examination in this paper were obtained from the R package and API `worldfootballR` (Zivkovic 2022). Employing the open-source `R` programming language (R Core Team 2023), we conducted the cleaning and analysis procedures, leveraging several R libraries and packages such as `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `ggplot2` (Wickham 2016), `knitr` (Xie 2023), `readr` (Wickham, Hester, and Bryan 2023), `dplyr` (Wickham et al. 2023), `styler` (Müller and Walthert 2024) and `arrow` (Richardson et al. 2024). We used `rstanarm` (Goodrich et al. 2022) for the models and `modelsummary` (Arel-Bundock 2022) for the model summary.

As mentioned above, in terms of estimand, the primary focus of this analysis is to assess the impact on league points Sir Alex would have had if he had been the manager between 2014 and 2023 as well. The three key predictors include the Success Rate per season (also interpreted as the Win Ratio or Win Percentage), the total goals scored and the goal difference (i.e. goals conceded deducted from goals scored). These will be talked more in detail about further into the paper.

In (Section 2.1) we provide an overview of the raw data obtained, (Section 2.2) delves deeper into the intricacies of the data cleaning process and (Section 2.3) outlines the measurement aspects for the paper.

2.1 Raw Data

As mentioned above, the raw data for this paper was collected from the R package and API `worldfootballR` (Zivkovic 2022). Team statistics per season for the English Premier League were obtained from the package. We obtained data per season, that implies that there were 20 files of raw data for 20 seasons we wish to analyze for 20 different teams in league each season. The dataset contained a lot of statistics that were not all relevant to us. Relevant statistics included the Squad Name, the finishing Rank, the number of Wins, Draws, Losses, Goals For, Goals Against, Goal Difference and the League Points. Other statistics included the Average Attendance per season, Expected Goals etc.

The next section (Section 2.2) will show the data cleaning process, further outlining the structure of the refined dataset employed in our analysis.

2.2 Cleaned Data

The indicators or statistics mentioned in (Section 2.1) were extracted from each raw data file and combined together to form two data tables for seasons between 2004-2013 and 2014-2023. Table 1 shows the cleaned data table for the seasons between 2004 and 2013 under Sir Alex. Table 4 in the appendix shows the cleaned data table for the seasons between 2014 and 2023 for reference.

Table 1: Cleaned Data showing Premier League statistics for Manchester United (2004-2013)

Year	Rank	Wins	Draws	Losses	Win Percentage (%)	Goals Scored	Goals Conceded	Goal Difference	League Points
2004	3	23	6	9	60.53	64	35	29	75
2005	3	22	11	5	57.89	58	26	32	77
2006	2	25	8	5	65.79	72	34	38	83
2007	1	28	5	5	73.68	83	27	56	89
2008	1	27	6	5	71.05	80	22	58	87
2009	1	28	6	4	73.68	68	24	44	90
2010	2	27	4	7	71.05	86	28	58	85
2011	1	23	11	4	60.53	78	37	41	80
2012	2	28	5	5	73.68	89	33	56	89
2013	1	28	5	5	73.68	86	43	43	89

There are 10 columns in total in the League Statistics data table:

1. The `Year` column refers to the Premier League season. Generally, a season starts in August and ends in May of the next year. Hence, seasons are generally represented as,

for instance, 2003-2004. However, here for simplicity, the `Year` column refers to the year the season ends. So, the season 2003-2004 will be represented as 2004.

2. The `Rank` column points to the position Manchester United finished in, in the particular season. The Rank is going to between 1 and 20 because there are 20 teams in the league each year.
3. The `Wins` column contains the total number of wins by Manchester United in the particular season. There can be at maximum 38 wins because there are 38 games in total.
4. The `Draws` column contains the total number of draws by Manchester United in the particular season. There can be at maximum 38 draws because there are 38 games in total.
5. The `Losses` column contains the total number of losses by Manchester United in the particular season. There can be at maximum 38 losses because there are 38 games in total. Additionally, in total there are going to be 38 wins, draws and losses.
6. The `Win Percentage (%)` column, also referred to as the *Success Percentage* is calculated as the total number of wins divided by the total number of games, i.e. 38. It is shown as a percentage for simplicity of understanding. It is understood that higher the Win Percentage, the better the season was, and the better the league points will be.
7. The `Goals Scored` column outlines the number of goals scored by the team in the particular season. It is understood that this measure is a good indicator of how well the offense played in a particular season. The higher the goals scored, the better the offense played.
8. The `Goals Conceded` column outlines the number of goals conceded by the team in the particular season. It is understood that this measure is a good indicator of how well the defense played in a particular season. The lower the goals conceded, the better the defense played.
9. The `Goal Difference` column outlines the difference between the number of goals scored and goals conceded. It is understood that this measure is a good indicator of how well the team played as a whole. This value can be negative as well, implying that the team did bad in the season.
10. The `League Points` column indicates to the total number of points scored that season. This indicator is essential in determining where the teams end up in the league. A team receives 3 points for a win, 1 point for a draw and 0 points for a loss.

We will be using the *Win* or *Success Percentage*, *Goals Scored* and *Goal Difference* as our predictors to model for the *League Points* ahead, which will be explained in detail in (Section 3). Now, in (Section 2.3), we will delve into the measurement aspects of the data presented in the tables shown. Understanding this process is crucial for drawing meaningful insights from the data analysis results.

2.3 Measurement

According to the Premier League (*Statistics Explained* 2024), all the official performance data is collected and reported by Opta, a part of Stats Perform (*Stats Perform* 2024). All the data is collected by a team of three people which cover each match. These three people include two highly trained analysts who go through video-based collection system to gather data, and a quality control analyst who can rewind the video feed frame-by-frame to make sure the data collected is correct (*Statistics Explained* 2024). Additionally, the data collected is then subject to an exhaustive post-match check to ensure accuracy. This comprehensive process ensures that the data collected is complete and highly accurate. As the data analyzed in this paper is *not* collected and reported by the teams themselves, it is highly unlikely that a collection bias would exist. However, some might argue that some teams might be able to pressure the officials into tipping the statistics over in their favor, but this might be highly unlikely as well.

The `Win Percentage` or also referred to as the `Success Percentage` was calculated manually by taking taking a ratio of the wins in a season and the total number of games in a season (i.e. 38). As mentioned above, as the wins were reported by Opta (*Statistics Explained* 2024) and not by individual teams, hence once again the issues about any biases caused can be avoided.

There might be a concern when considering the fact that this paper analyses data for the past 20 years. There might have been a difference in data reporting back in 2004 when compared to right now. It is essential to note that, this difference would have a bigger impact on data for a minute to minute match data analysis. However, in this report we use data about each season. As this data includes most the number of points scored, wins, draws, losses and goals, we work under the assumption that this data should not be affected to a large extent. However, at the same time it is important to still acknowledge that data collection in the past could have impacted the data being analyzed and modeled with.

In (Section 3) we will describe the models used in this paper and provide with justification and hypothesis of what we expect from the models.

3 Model

The goal of our modelling strategy is twofold. Firstly, we want to analyze to what effect the *Win* or *Success Percentage*, *Goals Scored* and *Goal Difference* have on the end-of-season league points during two different time periods, i.e. 2004-2013 and 2014-2023. Secondly, we want to use these results to predict the end-of-season league points for if Sir Alex would have continued his managership after 2013 to between 2014 and 2023. In this paper, we will make 2 models, one for modelling Sir Alex's final 10 years of his career at Manchester United in the English Premier League, and the one for modelling the next 10 years at Manchester United which saw a total of 8 different managers. We study two models in this paper so that we can effectively compare how Sir Alex fared when compared to 8 other managers at running Manchester United.

Here we briefly describe the Bayesian analysis model used to investigate and study the multiple linear regression model of end-of-season league points for two time periods, 2004-2013 and 2014-2023. Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the end-of-season league points obtained by Manchester United. Then β_1 is the win or success percentage for the season, γ_1 is the number of goals scored in the season, and δ_1 is the goal difference for the season. The two models in this paper were run with auto scaling priors values for which have been obtained by running `prior_summary`. Model 1 for the time period 2004-2013 is represented as:

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i + \delta_i && (2)\\
\alpha &\sim \text{Normal}(0, 15) && (3)\\
\beta &\sim \text{Normal}(0, 2.13) && (4)\\
\gamma &\sim \text{Normal}(0, 1.30) && (5)\\
\delta &\sim \text{Normal}(0, 1.25) && (6)\\
\sigma &\sim \text{Exponential}(0.18) && (7)
\end{aligned}
$$

Model 2 for the time period 2014-2023 is represented as:

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{8}$$
$$\mu_i = \alpha + \beta_i + \gamma_i + \delta_i \tag{9}$$
$$\alpha \sim \text{Normal}(0, 16) \tag{10}$$
$$\beta \sim \text{Normal}(0, 2.36) \tag{11}$$
$$\gamma \sim \text{Normal}(0, 2.27) \tag{12}$$
$$\delta \sim \text{Normal}(0, 1.43) \tag{13}$$
$$\sigma \sim \text{Exponential}(0.15) \tag{14}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (Goodrich et al. 2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the end-of-season league points and the win percentage, the goals scored and the total goal difference per season. As we are modelling for the end-of-season league points, we found the predictors of the win percentage, the total goals scored in a season and the total goal difference per season to be appropriate. The win or success percentage communicates to us how *successful* the club was that season. As mentioned above, we calculate this indicator by taking a ratio of the wins with the total number of games. We expect that as the win percentage goes up, so should the league points.

The total goals in a season illustrates to us how well the offense of the team performed and how successful they were at scoring goals. We expect that a higher season goal scored count should imply a higher amount of league points. Wins and losses in the league are highly dependent on the goal difference. A team gets a win by simply scoring more goals than their opponent. The goal difference is an essential indicator which helps us model for the total end-of-season league points. We expect that a higher and positive goal difference will result in higher total league points. It is important to note that the goal difference can be a negative value as well, where in for that season the goals conceded would be great than the goals scored (we can infer a lot about how the league points should be).

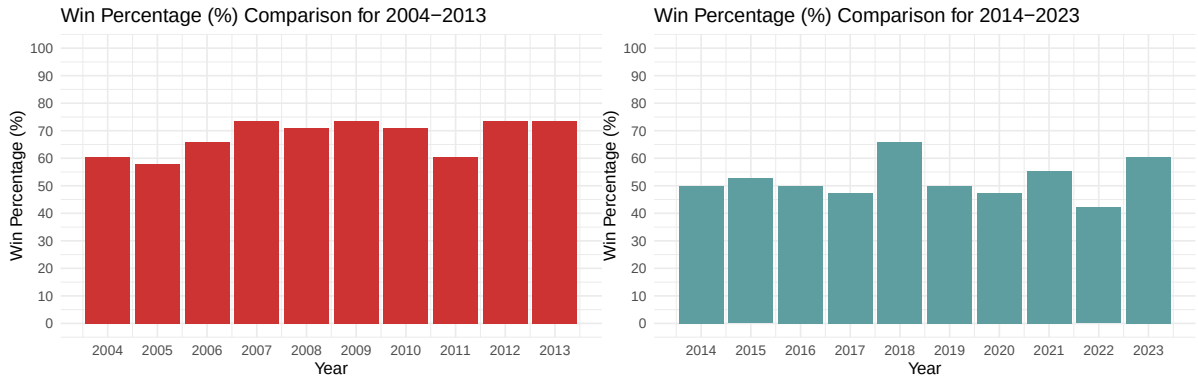
4 Results

The results section divided into two parts. Section 4.1 outlines the comparison between predictors such as *Win* or *Success Percentage*, *Goals Scored* and *Goal Difference* for the two eras under study: Sir Alex Ferguson's (between 2004 and 2013) and the other 8 managers' (between 2014-2023). As a prelude to studying the results of the model, this will allow us to visualize the differences in performance between both the eras. Section 4.2 communicates the

model summary, predicts how Sir Alex would have fared in the years 2014-2023, and discusses a comparison between the increase in league points scored over the subsequent years when compared with the other 8 managers.

4.1 Comparison of Predictors: Ferguson's Era vs. Subsequent Periods

Figure 1 shows us a side-by-side comparison of the Win or Success Percentage for when Manchester United was managed by Sir Alex Ferguson (Figure 1a) and when it was managed by the subsequent 8 managers (Figure 1b). We can clearly see in Figure 1a that the Win Percentage ranges from 60%-75% approximately, whereas, on the other hand in Figure 1b, the Win Percentage ranges from 40%-65% approximately. There is a very clear difference observed between performances in each of these eras. Remarkably, during Sir Alex's era (2004-2013), the win percentage was above 70% for 6 out of 10 years. On the other hand, during the subsequent era (2014-2023), the win percentage was below or equal to 55% for 8 out of 10 years. It is believed that this indicator has a direct impact on the league points scored for that season. From Figure 1, we can certainly infer a more successful managership by Sir Alex.

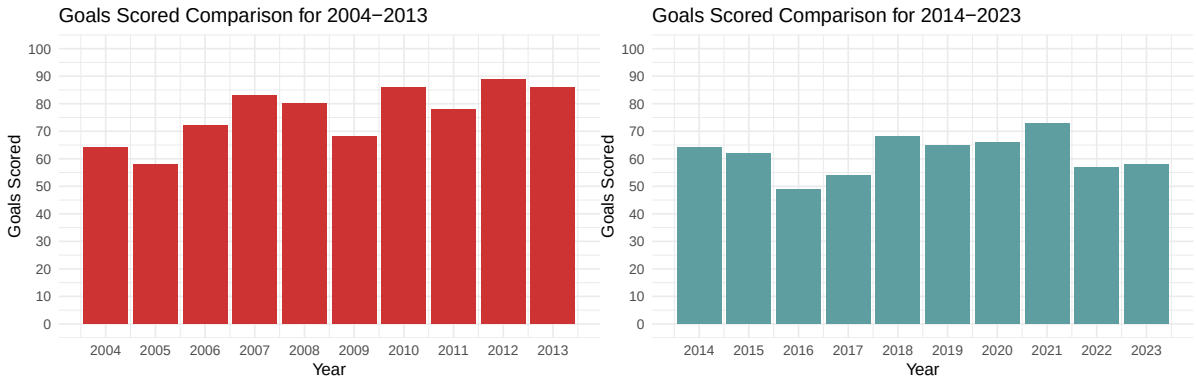


(a) Win Percentage (%) Under Sir Alex (2004-2013)(b) Win Percentage (%) Under Subsequent 8 Managers (2014-2023)

Figure 1: Comparing the Win Percentage (%) for eras between 2004-2013 and 2014-2023

Figure 2 shows us a side-by-side comparison of the Goals Scored for when Manchester United was managed by Sir Alex Ferguson (Figure 2a) and when it was managed by the subsequent 8 managers (Figure 2b). We observe that for Sir Alex's era the goals scored range from 60-90 goals, most of which are above 80 goals. On the other hand, for the subsequent period of time, the goals scored are lesser than 65 in approximately 8 out of 10 years. This general gap of 15 goals on average tells us how the offense was much better during Sir Alex's era. As mentioned in Section 3, we expect a positive relationship between the goals scored and the league points which will be explored in Section 4.2.

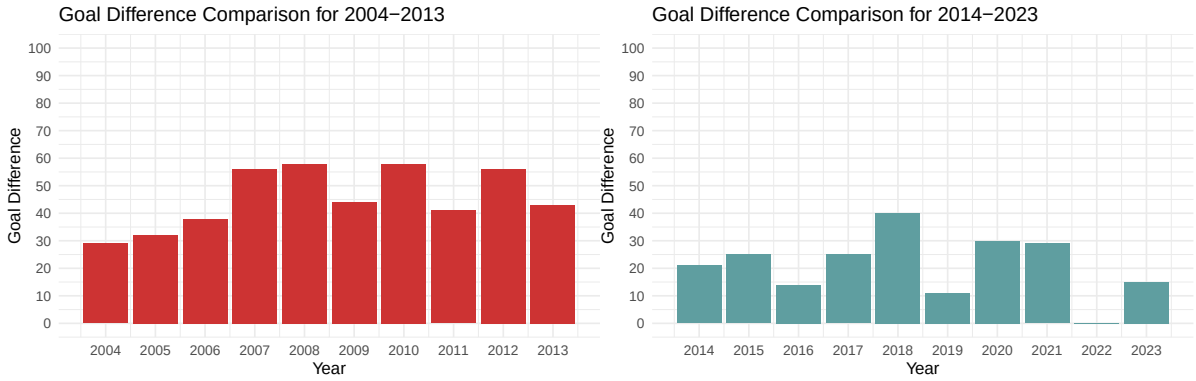
Finally, Figure 3 shows a side-by-side comparison of the Goal Difference for when Manchester United was managed by Sir Alex Ferguson (Figure 3a) and when it was managed by the



(a) Goals Scored Under Sir Alex (2004-2013) (b) Goals Scored Under Subsequent 8 Managers (2014-2023)

Figure 2: Comparing the Goals Scored for eras between 2004-2013 and 2014-2023

subsequent 8 managers (Figure 3b). We observe that for Sir Alex's era the goal difference ranges from 30-60 goals, most of which are above 40 goals (in difference). On the other hand, for the subsequent period of time, the goals scored are lesser than 25 in approximately 7 out of 10 years. We also see that for 2022, the goal difference is 0. This means that the defense was not as effective to win games, which is seen by the extremely low win percentage that year (42.11%). This indicator is extremely necessary in communicating how the whole team, including offense and defense performed that season.



(a) Goal Difference Under Sir Alex (2004-2013) (b) Goal Difference Under Subsequent 8 Managers (2014-2023)

Figure 3: Comparing the Goal Difference for eras between 2004-2013 and 2014-2023

In Section 4.2 we will communicate the model summary, attempt to predict how Sir Alex would have fared in the years 2014-2023, and discuss a comparison between the increase in league points scored over the subsequent years when compared with the other 8 managers.

4.2 Predicting League Points Under Extended Ferguson Era

As mentioned earlier in the paper, we run two models, for the time periods 2004-2013 and 2014-2023. This is essentially to be able to model Sir Alex's last 10 years at Manchester United, as well as the performance of the other 8 managers in the subsequent years. Creating two models allows us with inference and also comparison of predictions of how both models would fare in the dark era for Manchester United i.e. 2014-2023. Section B.1 provides the Model Summary for the two models we want to study in this paper (Table 5).

We observe that both our models show a high positive correlation between the *Win Percentage* and the *end-of-season League Points*, where Sir Alex's model gives us a higher correlation than the other model. This satisfies our hypothesis that as the Win Percentage increases, so will the League Points. We see a surprising negative correlation between the *Goals Scored* and the *end-of-season League Points*. This could be the case as we only looked at 10 years in each era, making the number of observations insufficient to find an appropriate correlation. However, the coefficients along with the standard error might suggest that these results for Goals Scored are not necessarily significant in this model. Finally, as expected, we see a positive correlation between *Goal Difference* and the *end-of-season League Points*. This again satisfies our hypothesis that as the Goal Difference increases, so will the League Points.

Now we will move on to the main focus and aim of the paper, i.e. to predict Sir Alex's performance, using the model, between years 2014-2023 and comparing it with the actual performance and the modeled for performance by the other 8 managers between 2014-2023.

Table 2: Comparison of Predicted League Points for Sir Alex's Continued Era and Other 8 Mangers with Actual League Points (2014-2023)

Year	Sir Alex	Other 8 Managers	Actual Points
2014	69	67	64
2015	71	70	70
2016	69	67	66
2017	67	66	69
2018	82	82	81
2019	69	65	66
2020	67	67	66
2021	73	72	74
2022	62	58	58
2023	78	74	75

Table 2 contains 4 columns. It shows seasons from 2014-2023, the predicted end-of-season league points for Sir Alex Ferguson, the modeled end-of season league points for the other 8 managers and finally the actual league points obtained that year by Manchester United.

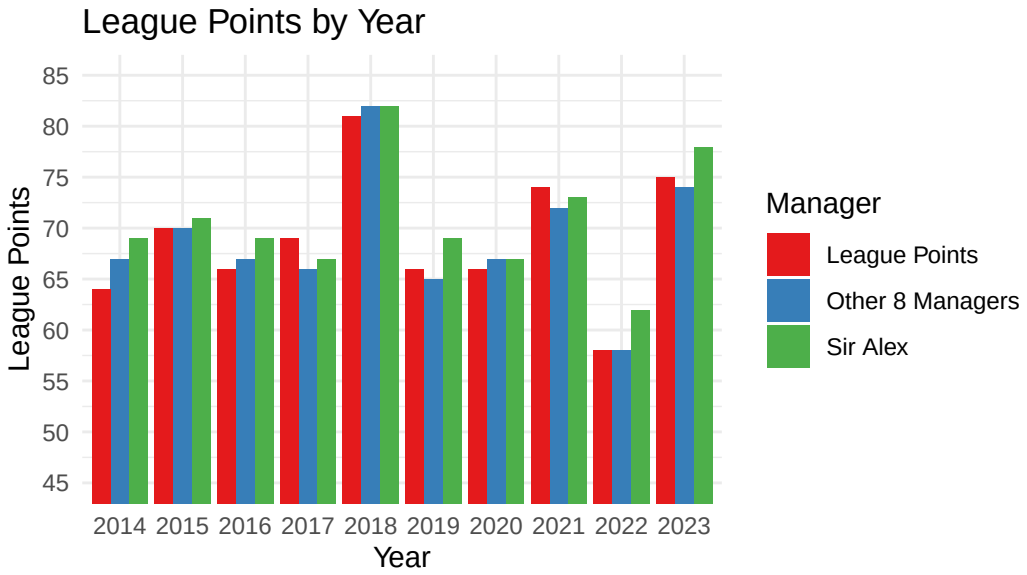


Figure 4: Comparing League Points for predicted Sir Alex extended tenure, modeled other 8 managers and actual league points (2014-2023)

Figure 4 shows a season by season comparison of how well Sir Alex would have done in terms of league points, with how well the other 8 managers are predicted to do, and how well the team did in real. It is important to note here that looking at Table 2 and Figure 4, we realise that although Sir Alex is seen to would have done better on average between 2014 and 2023, the difference is not as much as we expected. For 8 out of 10 seasons, Sir Alex is predicted to have done better than the team performed. But as we will observe in Figure 5 and Table 3, the difference is not a lot.

Table 3: Comparison of Differences between League Points for Sir Alex's Continued Era and Other 8 Mangers and between Actual League Points (2014-2023)

Year	Sir Alex - Other 8 Manager Points	Sir Alex - Actual Points
2014	2	5
2015	1	1
2016	2	3
2017	1	-2
2018	0	1
2019	4	3
2020	0	1
2021	1	-1
2022	4	4
2023	4	3

As we can see in Table 3, the predicted League Points for Sir Alex's extended tenure are always greater than the modeled for other 8 managers League Points between 2014 and 2023. On average, Sir Alex would would score **1.9** more points per season than the other 8 managers modeled prediction over the period of 2014-2023. Similarly, the predicted League Points for Sir Alex's extended tenure are greater than the actual league points received in 8 out of 10 seasons. For the seasons 2016-2017, and 2020-2021, we see that Sir Alex would have probably done worse than how the team performed in real. On average, Sir Alex would score **1.8** more points per season than the team did in real. We see that although there is an improvement in average league points, this difference is not as much as we have would have expected it to be. This will be explored in full depth in Section 5.

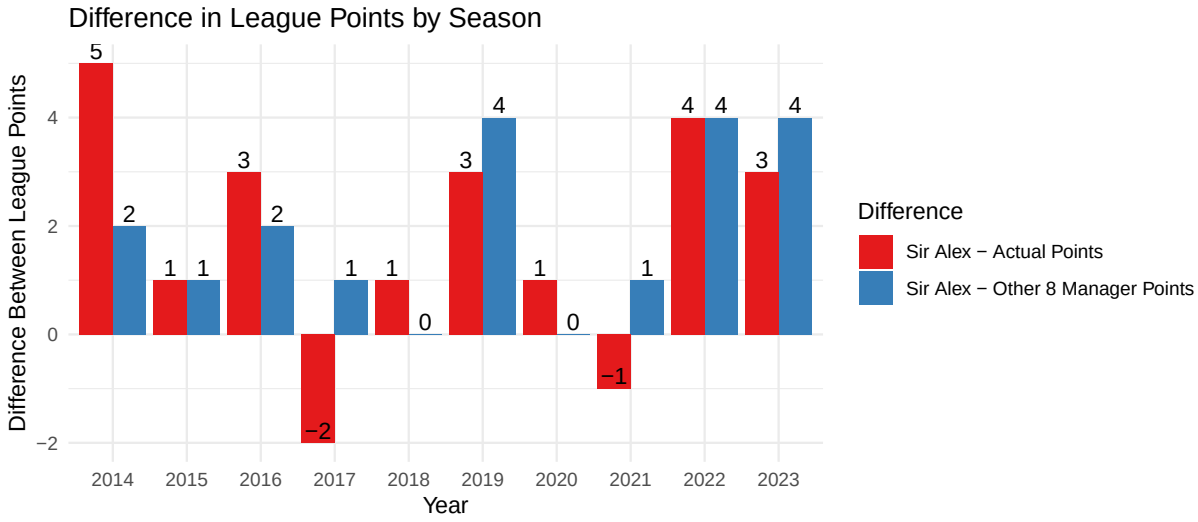


Figure 5: Plot for comparison of differences between League Points for Sir Alex's Continued Era and Other 8 Mangers and between Actual League Points (2014-2023)

Hence, finally in Section 5, we will explore the reasons behind why we observed a marginal improvement in Sir Alex's Performance in recent times and not as much as we expected we would. We will also discuss biases, weaknesses and further research in this field.

5 Discussion

5.1 Findings: Exploring Sir Alex's Only Marginal Improvement to Manchester United between 2014-2023

For the first part of our research, our results tell us about the state of the club. With the comparisons in Section 4.1, we can certainly affirm that Manchester United is in a worse state than it was 11 years ago under Sir Alex Ferguson. This can specifically be seen by the sudden drop in all three of Win Percentage, Goals Scored and Goal Difference between 2014-2023.

This has consequently affected the performance of the team in recent times. This can be because of the following reasons:

Firstly, as the English Premier League is a league based (or points based competition), success or the finishing position depends a lot on how well the other clubs are doing in the table too. For instance, Jürgen Klopp joined Liverpool in 2015, since when he has registered exceptional statistics every year. He was won the Premier League once in 2020, and in 328 matches has won 206 games, drawn 76 and lost just 46 (*Jürgen Klopp Manager Profile, Record & Stats* 2024). Similarly, Pep Guardiola joined Manchester City, Manchester United's city rivals, in 2016. Since then, he has won the Premier League 5 times in 2018, 2019, 2021, 2022, 2023, and in just 298 matches has register 219 wins, 41 draws and just 38 losses. Manchester City's dominance can be extended by also taking into account that they won the treble in 2023, which includes winning the Premier League, the UEFA Champions league and the FA cup all in one season (*Pep Guardiola Manager Profile, Record & Stats* 2024). With Manchester United's troubles of finding an appropriate manager for the club, it gets tough to compete with clubs like Liverpool and Manchester City which are on an exceptional run. Some individuals in the football community compare Pep Guardiola's success to that of Sir Alex because of his incredible tenure at Manchester City.

This leads to the second point of why Manchester United might not be doing as well right now as it was when Sir Alex was the manager. Since, Sir Alex's departure, the club has been unable to find the appropriate manager. This can be seen by looking at the fact that Manchester United has had 8 managers in the past 10 years. Additionally, there has been a lot of outrage against the current owners of the club who have been claimed to be doing a poor job at running the club. With speculations of a change in ownership in 2023, there was a lot of positivity towards the betterment of the club, hence contributing to the second best season with 75 league points in the past 11 years. However, the ownership did not change leading to the position in the 2024 season being much worse. Along with this, it is also believed that the poor transfer of players into Manchester United, along with overpaying for transfer fee and being exploited in the market has led to the clubs bad financial situation as well as lack of quality.

Moving on to the second part of our research, we see that although our results satisfy our hypothesis of how Sir Alex would on average do better than the other 8 managers, however there is only a very marginal increase seen in the prediction of if were to Sir Alex manage the club in the subsequent time period of 2014-2023. In Section 4.1, we saw that there are significant differences in how well Manchester United performed under Sir Alex when compared to the other 8 managers. We see in Figure 1, Figure 2 and Figure 3 that all three predictors show a much better performance under Sir Alex. On average, Sir Alex would score **1.8** more points per season than the team did in real. This is not a very significant number of points. As mentioned above, a win is 3 points and a draw is 1 point. According to the differences in Win Percentage, Goals Scored and Goal Difference we observed in Section 4.1, the predicted league points should have ideally been higher. When compared to the modeled prediction for

the other 8 managers, we see that Sir Alex would score on average **1.9** more points per season. Again, this improvement in performance is only marginal and not as to what we expected.

There can be several reasons why we do not see the desired extent of increase in points which will explored in Section 5.2.

5.2 Weaknesses

In this section, we will explore the various weaknesses that can probably be the reason for our results not being as high as we would have expected them to be. Firstly, there are a lot of lurking factors that come into play when assessing a team's performance in a season which can not be added to the model as they are not easily quantifiable. These factors would certainly play a part in improving the model. For instance, player morale is a very important factor to take into account. There have been a lot of reports in recent times sharing concerns of Manchester United players feeling like current managers' approach is affecting team morale (Reporter 2023). There have been numerous reports of unhappy players back lashing on the management in recent times. I believe that this indicator can have a major influence on the performance of the squad. However, it is very tough to quantify this indicator hence it was not added to our model.

Moving on, I believe that possession stats for the whole season on average can be an indicator of how well the squad is being able to dominate the field when playing. It is believed that higher the possession is the more goals a team will score. However, there were two issues when we considered adding this indicator to the model. First, for data that dated a decade ago, this statistic was not reported. Due to this, this indicator could not be added as a predictor in the model. Second, I believe that possession can vary depending on playing style as well as the managerial style too. Some managers believe in keeping the possession and playing out from the back line, however, the others might believe in counter-attack football. For instance, for the season 2020-2021, Manchester United was being managed by Ole Gunnar Solskjær (*Ole Gunnar Solskjær - Manager Profile* 2024), who preferred to play a more counter-attack style of football. That season Manchester United finished the season in 2nd place with a possession of 55.7%. Chelsea finished the season in 4th place with a possession of 60.9% (Zivkovic 2022). Hence, looking at this data it seemed after primary analysis that this predictor might not actually result with a significant correlation.

Hence, there are always more predictors that can be used to make the model more nuanced, however, for this paper we chose the predictors according to which ones were actually making sense in having a relationship with the variable being studied, i.e. League Points.

5.3 Further Scope

As a further scope, it would be interesting to look at the impact of financial decisions on the performance of clubs and squads in the league. It is widely known there clubs vary in terms

of the finances that are available and hence have better facilities, stadiums, marketing, and player signings. Some teams just are not able to sign more expensive players, whereas others are able to make teams of the best players in the world. This is of a lot of relevance in recent times with the growing popularity of the Saudi Pro League with the big-name signings and packages of Cristiano Ronaldo, Neymar and Karim Benzema. Similarly, the MLS has grown in popularity as well with the signing of Lionel Messi by Inter Miami. It can be conjectured that in today's football world, sound finances and a wealthy owner can push a team closer to success. It would be very interesting to take into account the moves in the transfer market and how effective they have been. This data can also be retreived from the R package and API `worldfootballR` (Zivkovic 2022).

Appendix

A Additional data details

Table 4: Cleaned Data showing Premier League statistics for Manchester United (2014-2023)

Year	Rank	Wins	Draws	Losses	Win Percentage (%)	Goals Scored	Goals Conceded	Goal Difference	League Points
2014	7	19	7	12	50.00	64	43	21	64
2015	4	20	10	8	52.63	62	37	25	70
2016	5	19	9	10	50.00	49	35	14	66
2017	6	18	15	5	47.37	54	29	25	69
2018	2	25	6	7	65.79	68	28	40	81
2019	6	19	9	10	50.00	65	54	11	66
2020	3	18	12	8	47.37	66	36	30	66
2021	2	21	11	6	55.26	73	44	29	74
2022	6	16	10	12	42.11	57	57	0	58
2023	3	23	6	9	60.53	58	43	15	75

Table 5: Model Summary for Model 1: Sir Alex's Era (2004-2013) and Model 2: Subsequent Periods (2014-2023)

	Sir Alex (2004-2013)	Other 8 Managers (2014-2023)
(Intercept)	29.20	30.67
	(8.49)	(8.29)
'Win Percentage (%)'	0.81	0.73
	(0.17)	(0.13)
'Goals Scored'	−0.02	−0.06
	(0.12)	(0.12)
'Goal Difference'	0.02	0.18
	(0.12)	(0.09)
Num.Obs.	10	10
R2	0.876	0.895
R2 Adj.	0.805	0.869
Log.Lik.	−19.992	−21.006
ELPD	−24.5	−24.6
ELPD s.e.	2.3	1.7
LOOIC	49.0	49.3
LOOIC s.e.	4.6	3.5
WAIC	47.3	48.3
RMSE	1.41	1.55

B Model details

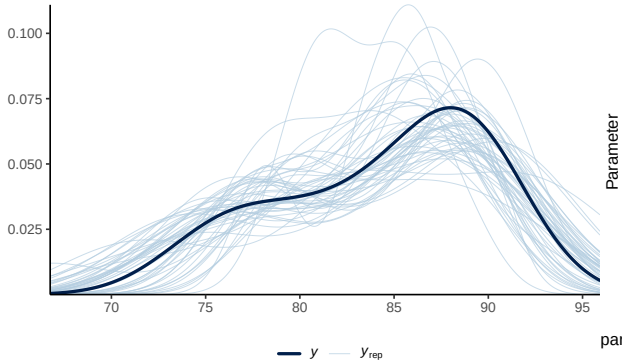
B.1 Model Summary

Table 5 presents the Model Summary for Model 1 which entails Sir Alex's Era (2004-2013) and Model 2 which entails Subsequent Periods (2014-2023).

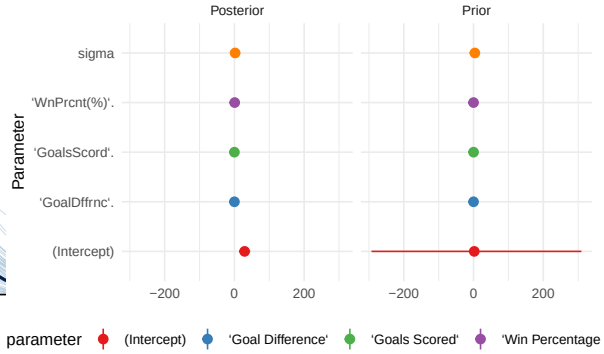
B.2 Posterior predictive check

In Figure 6a and Figure 7a we implement a posterior predictive check for the first model for the era 2004-2013 and the second model for the era 2014-2023 respectively. This compares the actual outcome variable, i.e. *end-of-season league points* with simulations from the posterior distribution. We can see that as the posterior is able to simulate the data, the model does a good job of fitting the data.

In Figure 6b and Figure 6b we compare the posterior with the prior for the first model for era 2004-2013 and the second model for the era 2014-2023 respectively. We do this to see how much the estimates change once data is taken into account.

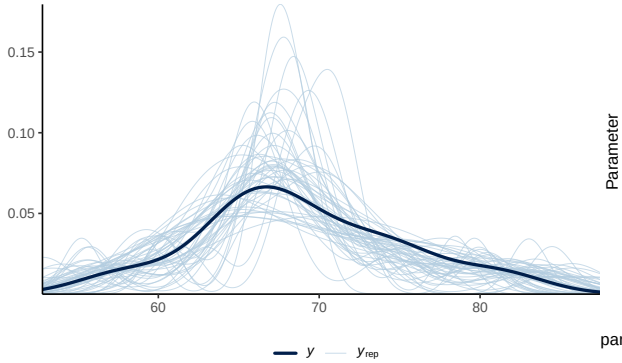


(a) Posterior prediction check (2004-2013)

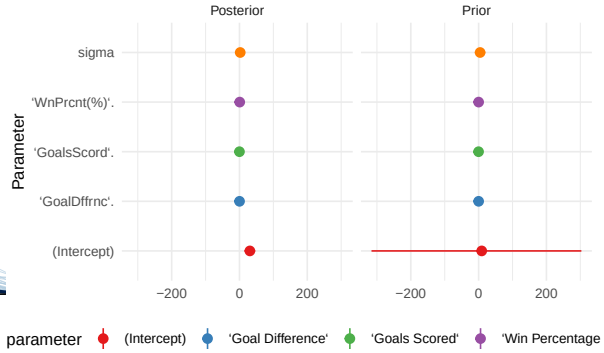


(b) Comparing the posterior with the prior(2004-2013)

Figure 6: Examining how the model fits, and is affected by, the data (2004-2013)



(a) Posterior prediction check (2014-2023)



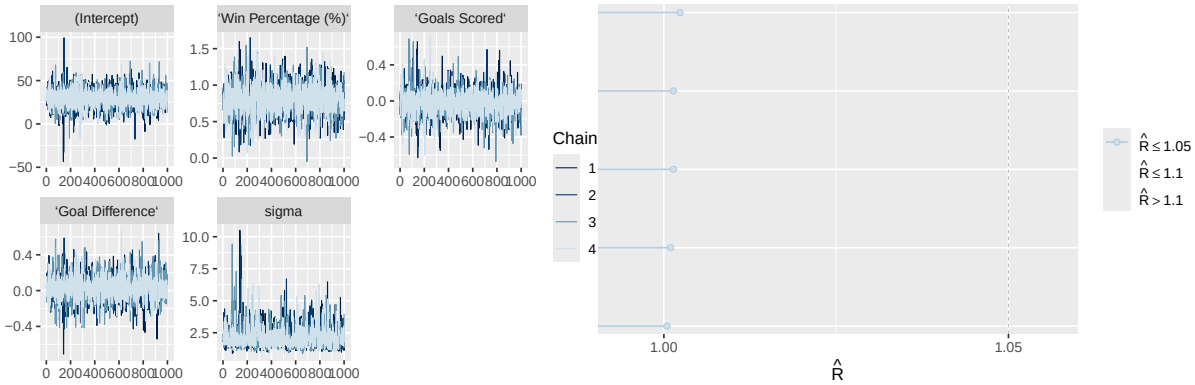
(b) Comparing the posterior with the prior(2014-2023)

Figure 7: Examining how the model fits, and is affected by, the data (2014-2023)

B.3 Diagnostics

Figure 8a and Figure 9a are trace plots for the first model for the era 2004-2013 and the second model for the era 2014-2023 respectively. Both figures shows a nice overlap between the chains and lines that appear to bounce around, but are horizontal. This suggests that there are no alarming issues in the models.

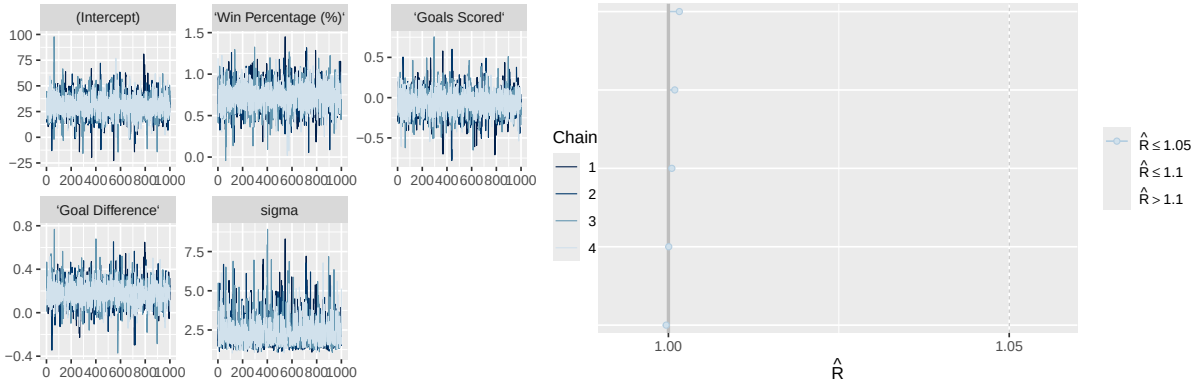
Figure 8b and Figure 9b are Rhat plots for the first model for the era 2004-2013 and the second model for the era 2014-2023 respectively. It shows that in both figures everything is close to 1 and does not exceed 1.1. This again suggests that there are no issues with the models.



(a) Trace plot for (2004-2013 model)

(b) Rhat plot (2004-2013 model)

Figure 8: Checking the convergence of the MCMC algorithm for 2004-2013



(a) Trace plot for (2014-2023 model)

(b) Rhat plot (2014-2023 model)

Figure 9: Checking the convergence of the MCMC algorithm for 2014-2023

C Datasheet for Dataset

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset (API/package) `worldfootballR` (Zivkovic 2022) was created to enable extraction and cleaning of World Football (Soccer) Data.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The authors of this API/package are Jason Zivkovic, Tony ElHabr, Tan Ho and Samuel H.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - Information not provided.
4. *Any other comments?*
    - No.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - The package provides with game-by-game, team and player data accumulated from popular football sources including *FBref*, *Transfermarkt*, *Understat* and *fotmob*.
2. *How many instances are there in total (of each type, if appropriate)?*
    - This is tough to calculate as the package provides with a lot of different data for different leagues in different countries.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

- The datasets retrieved from the package in this paper are **not** samples of a larger set. Instead it provides all the data that exists when the country and league is put in.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - The data contains parameters that can help assess a teams performance such as Wins, Draws, Losses, Goals Scored, Goals Conceded, Goal Difference etc. All of these parameters aid one in analysing team or player performance for each season.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - There is a unique ID for each team provided for each season.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - There were some statistics which were missing for earlier season around the 2010s. This information included possession state, shots taken, shots on target etc. This can be because the technology to track such indicators was not available or fully developed back then.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - No.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - No.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - There maybe but the indicators used in this paper were cross checked to be appropriate with other sources too.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply*

to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

- This package provides datasets that are self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No, all the data is freely available to the public.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - Yes, the package provides with player data which includes the names for the actual players. This is to aid individuals in analysing player performance.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No.

16. *Any other comments?*

    - No.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- Premier League data is collected and reported by Opta, a part of Stats Perform (*Stats Perform* 2024).

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - All the data is collected by a team of three people which cover each match. These three people include two highly trained analysts who go through video-based collection system to gather data, and a quality control analyst who can rewind the video feed frame-by-frame to make sure the data collected is correct (*Statistics Explained* 2024). Additionally, the data collected is then subject to an exhaustive post-match check to ensure accuracy.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - N/A

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Only the team of three people stated above.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data is collected match by match and is a continuous process.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Information not provided.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - It is from popular football sources including *FBref*, *Transfermarkt*, *Understat* and *fotmob*.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- Information not provided.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - Information not provided.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - Information not provided.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - N/A
12. *Any other comments?*
    - No.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - N/A
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - N/A
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - N/A
4. *Any other comments?*
    - N/A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Information not provided.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - No.
3. *What (other) tasks could the dataset be used for?*
    - This package can be used for financial analysis of clubs as well where the wages, transfer fees and investments can be used to analysis the financial stability of a club.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - No.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No.
6. *Any other comments?*
    - No.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - It is a package in CRAN and hence is open for use for everyone.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - It is a CRAN package.
3. *When will the dataset be distributed?*
    - It was released in 2022.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The package already holds a GPL-3 license.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No.

7. *Any other comments?*

    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The author of the package maintains it: Jason Zivkovic

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - Email Address: jaseziv83@gmail.com

3. *Is there an erratum? If so, please provide a link or other access point.*

    - https://github.com/JaseZiv/worldfootballR/issues

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - The package is updated often. Here is the link to the GitHub: https://github.com/JaseZiv/worldfootb

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - Player data is widely available.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Information not provided.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - Contacting the authors is always an option.

8. *Any other comments?*

    - No.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

How Many Trophies Have Manchester United Won? 2023. Sporting News Canada. https://www.sportingnews.com/ca/soccer/news/manchester-united-trophies-won-complete-list-silverware/uuyb6wc5o7isrfwa1pjssb07.

Jürgen Klopp Manager Profile, Record & Stats. 2024. Premier League. https://www.premierleague.com/managers/5119/J%C3%BCrgen-Klopp/overview.

Manchester United. 2024. Encyclopædia Britannica, inc. https://www.britannica.com/topic/Manchester-United.

Manchester United - Manager History. 2024. WorldFootball. https://www.worldfootball.net/teams/manchester-united/9/.

Müller, Kirill, and Lorenz Walthert. 2024. *Styler: Non-Invasive Pretty Printing of r Code*. https://CRAN.R-project.org/package=styler.

Ole Gunnar Solskjær - Manager Profile. 2024. Transfermarkt. https://www.transfermarkt.us/ole-gunnar-solskjaer/profil/trainer/7286.

Pep Guardiola Manager Profile, Record & Stats. 2024. Premier League. https://www.premierleague.com/managers/5285/Pep-Guardiola/overview.

Premier League. 2024. Premier League - facts; history. https://www.footballhistory.org/league/premier-league.html.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reporter, Metro Sport. 2023. *Man Utd Players Concerned Erik Ten Hag Approach Is 'Affecting Team Morale'*. Metro.co.uk. https://metro.co.uk/2023/11/01/man-utd-players-concerned-erik-ten-hag-approach-is-affecting-team-morale-19752874/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Statistics Explained. 2024. Premier League. https://www.premierleague.com/stats/clarification.

Stats Perform. 2024. Stats Perform. www.statsperform.com/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal*

of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zivkovic, Jason. 2022. *worldfootballR: Extract and Clean World Football (Soccer) Data*. https://CRAN.R-project.org/package=worldfootballR.