Sport Analysis

The prediction of the final credits and each game scores of the Premier League

Stanley Zheng Gefei Wu Zhengtao Zhang Yufanxing Liu Zihao Zhang*

January 7, 2024

Abstract

This research utilizes sophisticated analytical tools, including Multiple Linear Regression (MLR) models and simulations, applied to an extensive ten-year dataset of English Premier League (EPL) seasons. Across five MLR models, Liverpool consistently emerges as the frontrunner for the current season's championship. The incorporation of Monte Carlo simulation and Poisson distribution enhances the analysis by providing nuanced insights into likely score differentials for individual matches, contributing to a comprehensive understanding of match dynamics. This study not only reaffirms Liverpool's favoritism for the EPL title but also significantly advances the field of football analytics, offering a detailed perspective on team performance and match-specific nuances. In essence, the research positions Liverpool as the prime candidate for EPL glory in the 23-24 season.

1 Introduction

In the vast tapestry of global sports, football stands tall as a unifying force, drawing in an immense and diverse fan base. Our team takes pride in being an integral part of this dynamic community that shares an unwavering passion for the beautiful game. Amidst the multitude of football leagues, the English Premier League (EPL) emerges as a beacon of influence, where each match becomes a spectacle, captivating the hearts and minds of millions of fans worldwide. The burning anticipation of which club will emerge triumphant as the EPL champion becomes a narrative that unfolds with each thrilling match, igniting fervor and enthusiasm among spectators.

The analytical lens through which we examine the EPL season involves the sophisticated use of multivariable linear regressions (MLRs). This statistical approach allows us to meticulously

*__Stanley Zheng__: Conceptualization, Methodology, Software, Data Curation, Investigation, Formal analysis, Writing, Supervision, Project administration. **Gefei Wu**: Methodology, Software, Investigation, Formal analysis. **Zhengtao Zhang**: Data Curation. **Yufanxing Liu**: Methodology, Software. **Zihao Zhang**: Methodology.

scrutinize data spanning the last decade, enabling us to make informed predictions about the final point standings of each team at the conclusion of the ongoing season. The interconnectedness of these predictions with the overarching question of championship conquest adds a layer of intrigue to our analysis. Concurrently, alternative methodologies are employed to forecast goal tallies for both home and away teams in each match, further enriching our understanding of the potential outcomes.

This research embarks on a journey to unravel the complexities of the English Premier League's competitive landscape. Drawing on a thorough examination of historical data, our aim is to provide reasoned and insightful predictions for the current season, offering fans a nuanced perspective on each team's performance. Beyond the realm of passionate supporters, the practical implications of our findings extend to club managers, dedicated fans, and professionals in the burgeoning field of sports analytics. By shedding light on the potential trajectories and competitive dynamics within the English Premier League, our research strives to contribute valuable insights to the broader discourse surrounding the sport.

2 Dataset

The dataset comprehensively covers match statistics spanning the English Premier League (EPL) football matches. Each entry meticulously details specific match facets, encompassing vital information such as the match date, participating home and away teams, full-time goal counts (FTHG and FTAG for the home and away teams respectively), denoting the match outcome (FTR indicating H for home team win, A for away team win, and D for draw), as well as halftime goal tallies (HTHG and HTAG).

Moreover, the dataset encapsulates intricate betting odds provided by bookmakers, presenting probabilities for diverse outcomes including the likelihood of a home win (B365H), a draw (B365D), and an away win (B365A). This extensive repository doesn't limit itself to match results but also incorporates additional statistical insights such as shots on target (HST and AST for home and away teams respectively), contributing to a multifaceted understanding of match dynamics. Notably, the dataset spans an extensive timeline, encompassing data from the 2012-2013 season to the 2023-2024 season, offering a rich and comprehensive historical perspective on EPL match performances and betting trends.

You can see part of our raw data in Appendix.A.

3 Implementation of MLR for Prediction of seasonal credits

3.1 Data Preprocessing

Within the code(the first code block), a primary focus is dedicated to the initialization of a meticulously structured dataframe, named "total_data_12_13," serving as the repository for comprehensive statistics pertaining to the English Premier League (EPL) season of 2012-2013. This foundational step involves the extraction and incorporation of unique team names, strategically placed as the initial column of the dataframe. To ensure an all-encompassing representation of team performance, an array of columns is meticulously crafted to encompass various statistical dimensions. These include essential metrics like wins (W), draws (D), losses (L), total goals scored (Agoal), total goals conceded (Lgoal), as well as performance indicators such as shots (S), shots on target (ST), fouls (F), corners (C), yellow cards (Y), and red cards (R). This comprehensive structure not only sets the stage for in-depth analysis but also lays the groundwork for a holistic understanding of team dynamics throughout the specified EPL season.

```
# to-do
unique(data_12_13["HomeTeam"])
total_data_12_13 = as.data.frame(unique(data_12_13["HomeTeam"]))
names(total_data_12_13)[1] <- "Team"
column_names <- c("W","D", "L", "Agoal", "Lgoal", "S",
                  "ST", "F", "C", "Y", "R")
new_columns <- data.frame(matrix(ncol = length(column_names),
                           nrow = nrow(total_data_12_13)))
colnames(new_columns) <- column_names
total_data_12_13 <- cbind(total_data_12_13, new_columns)
total_data_12_13
```

Moving forward, the code(the second code block) is intricately designed to dissect the initial 16 matches of the season, placing a specific emphasis on calculating the number of wins for each

participating team. This task is streamlined through the definition of a specialized function named "count_win," strategically crafted to extract pertinent data pertaining to a specific team and subsequently tally the number of victories. The orchestration of this function is then executed through an iteration process, where the unique team names are systematically traversed. The culmination of this iterative journey results in the meticulous calculation and incorporation of the count of wins for each team, seamlessly populating the dedicated "W" column within the "total_data_12_13" dataframe. This code segment not only lays the groundwork for subsequent analytical endeavors but also exemplifies a structured and purposeful approach towards extracting and utilizing critical match data for insightful outcomes.

```
# win
count_win<-function(x){
    t = data_12_13[data_12_13$HomeTeam==x|
                   data_12_13$AwayTeam==x, ][1:16,]
    t1 = table(t[t$HomeTeam==x, ]$FTR)["H"]
    if (is.na(t1)) {
    t1 = 0
    }
    t2 = table(t[t$AwayTeam==x, ]$FTR)["A"]
    if (is.na(t2)) {
    t2 = 0
    }
    win = as.integer(t1 + t2)
    return(win)
}
team = unique(data_12_13["HomeTeam"])
team = team[,1]
W = c()
for (i in seq_along(team)) {
    res = count_win(team[i])
    W = c(W, res)
}
total_data_12_13$W = W
```

3.2 Models

As the Figure 1 showns, we use the history data to train our model, then we put the nowadays' data into this model we trained, and then we get the prediction. This is how the data-driven model works.

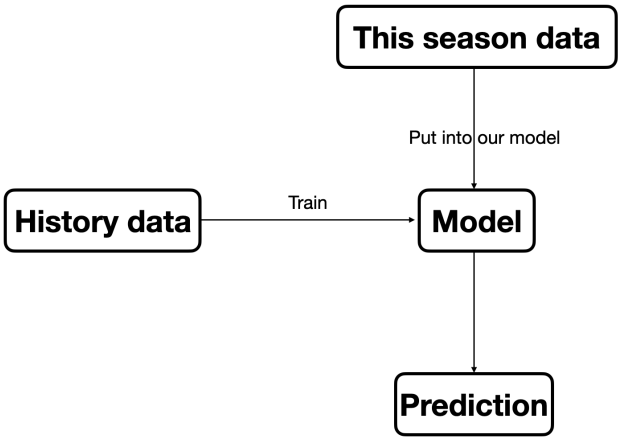

Figure 1: Data-Driven Model.

The esstial point of our MLR models is that we only use the first 16th rounds to predict the final credits. We also only use the first 16th rounds data of the current season to predict final credits of the current season. This is like a mask, because in the end, when we use the data to predict, we only use the first 16th data, so to keep the consistency, when we train our data, even though we have the full data, we only use the first 16th round. This idea is as same as the idea of GPT in NLP area.

Now, we have the data we need, we can do the analysis.

First, we just assumed that every explanatory variable could affect the final credits. And we apply end-to-end structure, which means we do no transformation to the explanatory variables or response variable. Then we just use the lm() function in R to get our multivariable linear regression models. The results you can see in Appendix.B.

As we all know, we you use multivariable linear regression, it's actually build a high dimensional space, and try to find a hyperplane to fit our data point in this space. However, we know when the dimension increases, the data points we need increases exponentially. So, it's obvious that our data from last 10 seasons are not enough to support a 12 dimensional space. Intuitionally, this

will lead to the "Curse of Dimensionality", Thus this, we need to reduce the dimension of our data.

From the results of model1 we know W, D and S are the most related variables. We just keep these variable in our Model2.

Besides using the statistic way to decide which variable we should keep, we can apply some other methods. Like we can combine two variables into one index, and use this index to represent the original variable. Here, we combine the S and ST to get STR(Shoot-on-Target Rate), and we combine the Agoal and S to get GTR(Goal-on-Transfer Rate). The convert formula is

$$STR = \frac{ST}{S} \qquad GTR = \frac{Agoal}{S}.$$

We also find that the value of R is very low, and only have very small amount of number. So, we just treat it as dummy variable.

Nevertheless, we can see that the data we use are not on the same scale, so sometimes we think that the scale will also affect the results. Thus, we will do some normalization to the data.

In the Model3, I apply the Z-score normalization to our explanatory variables. The convert formula of Z-score normalization is

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

In the Model4, I apply the Min-Max normalization to our explanatory variables. The convert formula of Min-Max normalization is

$$x_i' = \frac{x_i - \text{Min}(x)}{\text{Max}(x) - \text{Min}(x)}$$

In the Model5, I only make one change to Model4, which is I also apply the Min-Max normalization to the response variable.

You can see the summaries of these give models in Appendix.B.

3.3 Results

From the summary of these models, we can compare the statistic indices between these models.

The statistic indices in the Table 1 shows that, all these models have a very high R-squared and the P-values of theses models are very low. This means the model fit the sample points well, and these models all are reliable statistically. The residual standard error shows the bias between

	Model1	Model2	Model3	Model4	Model5
Residual Standard Error	7.081 on 208 DF	7.092 on 216 DF	7.285 on 206 DF		0.08673 on 206 DF
Multiple R-squared	0.8519	0.8457		0.8448	
Adjusted R-squared	0.8441	0.8436		0.835	
F-statistic	108.8 on 11	394.8 on 3		86.23 on 13	
P-value			$< 2.2 \times 10^{-16}$		

Table 1: Statistic Indices

the real value and the prediction. You can see it's kind of high, since at the end of the season, even 1 credit can change the ownership of the champion. But in all model the residual standard error is already over 7 credits. In model 5, since we do some normalization to the credit, the range of the the converted credits is [0, 1], so in Model5 the residual standard error also relatively high.

After we get these model, we can use these model to predict the final credits of each team at the end of this season. You can the prediction in Appendix.C. We can see the all five model predict that Liverpool. We only use first 16th round data, and now is the 20th round, Liverpool is actually on the top of the standings. Maybe it is the coincides, we will see the final results when the season over.

4 Implementation of other models for the Prediction of scores

4.1 Data Preprocessing

The code of data preprocessing snippet performs data preprocessing for football match data. It iterates through a list of file paths representing different seasons, reads CSV files, extracts relevant information, and calculates average goals for both home and away teams. The processed data is then stored, and all resulting data frames are combined into a single data frame. The final dataset is written to a CSV file named "combined_data.csv". The preprocessing includes handling column names, addressing team name variations, and aggregating match statistics.

The fully application you can find in the folder called "Implementation of Monte Carlo simulation and Poisson distribution for the Prediction of scores".

4.2 Model

This model combines Monte Carlo simulation and Poisson distribution to predict football match outcomes. The function takes as input the home team, away team, the number of simulations, and historical match data. It selects the first 100 matches from the dataset for simulation, conducting 10,000 simulations for each match.

In the simulation of each match, historical scoring averages drawn from the dataset are utilized to compute the average goals scored and conceded by both the home and away teams. This computation takes into consideration the amalgamation of historical data and overall team averages. Following this, through the application of the Poisson distribution, we generate random scores for the simulation. The resulting outcomes capture the most frequently occurring score and its corresponding probability.

The justification for employing Poisson distributions in modeling football scores stems from its aptitude in emulating the unpredictability of goal occurrences. The Poisson process, marked by its stationary, independent, and infrequent incremental nature, serves as an effective framework for mirroring the randomness inherent in goal events during football matches. By assuming that goals within short time intervals adhere to a Poisson distribution characterized by a rate parameter λ, this modeling approach adeptly encapsulates the essential aspects of sporadic and independent goal-scoring instances. As a result, it furnishes a practical approximation for the erratic nature of goal-scoring trends observed in football matches.

4.3 Results

In the specific code provided, the first 100 matches from the final_data dataset are simulated, and the results are written to a CSV file named "simulate_outcome.csv." The probabilities in the prob column are divided by 100 to correct the scaling. Part of the results you can see in the Appendix.D.

5 Discussion

In this comprehensive analysis of football match prediction models, our primary focus was on forecasting the English Premier League (EPL) champion for the current season. The foundational

model, Multiple Linear Regression (MLR), demonstrated interpretability and ease of implementation. However, its reliance on a linear relationship between predictor variables and outcomes may oversimplify the intricate dynamics of football matches. To address this limitation, future iterations could explore the integration of non-linear models, such as polynomial regression or advanced machine learning algorithms like neural networks, to capture more nuanced patterns in the data.

Additionally, we employed Monte Carlo simulation and the Poisson distribution to predict match scores, providing a probabilistic outlook. While effective in offering a range of possible outcomes, this approach simplifies the complex interplay of factors influencing match results. Recommendations include exploring alternative probability distributions, such as the negative binomial distribution, to better account for overdispersion in the data. Moreover, expanding the scope beyond goal counts to incorporate additional features, like possession percentages or team strategies, could enhance the overall predictive accuracy.

Looking forward, our study suggests several avenues for model refinement. Experimentation with non-linear models and exploration of alternative probability distributions beyond Poisson can contribute to more accurate predictions. The incorporation of time-dependent variables and dynamic factors, such as player form and injuries, will likely improve model performance. Ensuring robustness through validation with out-of-sample data and addressing ethical considerations, such as potential biases in referee decisions, are crucial steps in advancing the reliability and fairness of football match prediction models. In conclusion, continuous refinement and the integration of advanced statistical techniques will be pivotal in developing models that provide nuanced and reliable insights into the unpredictable world of football.

Appendix

A. Dataset

Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR	B365H	B365D	B365A	BWH
E0	18/08/12	Arsenal	Sunderland	0	0	D	0	0	D	C Foy	14	3	4	2	12	8	7	0	0	0	0	0	1.4	4.5	8.5	1.35
E0	18/08/12	Fulham	Norwich	5	0	H	2	0	H	M Oliver	11	4	9	2	12	11	6	3	0	0	0	0	1.8	3.6	4.5	1.8
E0	18/08/12	Newcastle	Tottenham	2	1	H	0	0	D	M Atkinson	6	12	4	6	12	8	3	5	2	2	0	0	2.5	3.4	2.75	2.6
E0	18/08/12	QPR	Swansea	0	5	A	0	1	A	L Probert	20	12	11	8	11	14	5	3	2	2	0	0	2	3.4	3.8	2
E0	18/08/12	Reading	Stoke	1	1	D	0	1	A	K Friend	9	6	3	3	9	14	4	3	2	4	0	1	2.38	3.25	3.1	2.4
E0	18/08/12	West Brom	Liverpool	3	0	H	1	0	H	P Dowd	15	14	10	7	10	11	7	3	1	4	0	1	4.2	3.5	1.91	4.1
E0	18/08/12	West Ham	Aston Villa	1	0	H	1	0	H	M Dean	8	10	4	6	17	8	6	4	0	1	0	0	2.2	3.3	3.4	2.25
E0	19/08/12	Man City	Southampton	3	2	H	1	0	H	H Webb	20	9	15	6	4	6	7	3	1	2	0	0	1.17	7	17	1.18
E0	19/08/12	Wigan	Chelsea	0	2	A	0	2	A	M Jones	12	5	4	3	16	11	7	1	2	2	0	0	6	3.75	1.62	6
E0	20/08/12	Everton	Man United	1	0	H	0	0	D	A Marriner	16	12	7	7	12	11	6	8	1	2	0	0	4.33	3.6	1.83	4.6
E0	22/08/12	Chelsea	Reading	4	2	H	1	2	A	L Mason	23	7	11	5	11	14	1	5	0	2	0	0	1.25	5.5	13	1.25
E0	25/08/12	Aston Villa	Everton	1	3	A	0	3	A	M Oliver	7	19	3	11	10	7	1	9	1	0	1	0	2.9	3.4	2.4	2.95
E0	25/08/12	Chelsea	Newcastle	2	0	H	2	0	H	P Dowd	11	11	6	5	8	10	2	3	1	0	0	0	1.5	4.33	6.5	1.5
E0	25/08/12	Man United	Fulham	3	2	H	3	1	H	K Friend	20	14	11	11	12	7	8	8	0	1	0	0	1.29	5.5	11	1.28
E0	25/08/12	Norwich	QPR	1	1	D	1	1	D	M Clattenburg	13	6	4	4	15	14	6	2	0	2	0	0	2.25	3.4	3.2	2.25
E0	25/08/12	Southampton	Wigan	0	2	A	0	0	D	A Taylor	14	12	9	8	11	9	10	3	0	0	0	0	2.1	3.4	3.5	2.05
E0	25/08/12	Swansea	West Ham	3	0	H	2	0	H	M Atkinson	10	7	7	6	13	17	5	4	2	4	0	0	2.1	3.3	3.6	2
E0	25/08/12	Tottenham	West Brom	1	1	D	0	0	D	M Dean	18	10	10	5	8	7	7	5	1	2	0	0	1.5	4.33	6.5	1.45
E0	26/08/12	Liverpool	Man City	2	2	D	1	0	H	A Marriner	15	11	8	5	10	7	4	4	1	0	0	0	3.1	3.4	2.3	3
E0	26/08/12	Stoke	Arsenal	0	0	D	0	0	D	L Mason	7	16	4	6	9	9	0	11	2	0	0	0	3.5	3.4	2.1	3.7
E0	01/09/12	Man City	QPR	3	1	H	1	0	H	C Foy	19	9	12	5	12	3	8	3	2	1	0	0	1.18	7	15	1.18
E0	01/09/12	Swansea	Sunderland	2	2	D	1	2	A	R East	14	4	10	3	11	7	8	0	1	1	1	0	2.2	3.3	3.4	2.2
E0	01/09/12	Tottenham	Norwich	1	1	D	0	0	D	M Halsey	15	10	9	4	7	12	4	2	1	0	1	0	1.36	5	8.5	1.36
E0	01/09/12	West Brom	Everton	2	0	H	0	0	D	J Moss	14	12	8	5	11	15	6	3	1	4	0	0	3	3.3	2.38	2.9
E0	01/09/12	West Ham	Fulham	3	0	H	3	0	H	A Taylor	17	14	12	13	12	5	6	6	0	0	0	0	2.5	3.25	2.88	2.55
E0	01/09/12	Wigan	Stoke	2	2	D	1	1	D	M Atkinson	9	16	5	8	13	11	2	1	1	2	0	0	2.25	3.25	3.3	2.2
E0	02/09/12	Liverpool	Arsenal	0	2	A	0	1	A	H Webb	17	11	8	7	12	7	10	2	2	2	0	0	2	3.5	3.75	2.05
E0	02/09/12	Newcastle	Aston Villa	1	1	D	0	1	A	L Probert	16	13	6	9	6	20	10	6	1	4	0	0	1.62	3.75	6	1.6
E0	02/09/12	Southampton	Man United	2	3	A	1	1	D	M Dean	15	18	8	9	9	4	4	7	1	0	0	0	6	4.2	1.53	6
E0	15/09/12	Arsenal	Southampton	6	1	H	4	1	H	K Friend	20	9	12	4	4	7	8	3	0	0	0	0	1.36	5	8.5	1.34
E0	15/09/12	Aston Villa	Swansea	2	0	H	1	0	H	L Mason	17	10	13	5	11	4	11	5	3	1	0	0	2.38	3.25	3.1	2.3
E0	15/09/12	Fulham	West Brom	3	0	H	2	0	H	R East	23	10	14	8	6	10	5	4	0	0	0	1	2.1	3.4	3.5	1.95
E0	15/09/12	Man United	Wigan	4	0	H	0	0	D	M Oliver	17	8	10	4	10	14	5	5	2	2	0	0	1.25	6	12	1.25
E0	15/09/12	Norwich	West Ham	0	0	D	0	0	D	C Foy	20	9	14	4	10	12	8	5	0	1	0	0	2.38	3.4	3	2.3
E0	15/09/12	QPR	Chelsea	0	0	D	0	0	D	A Marriner	10	13	6	9	16	17	2	4	0	2	0	0	4.5	3.5	1.83	4.75
E0	15/09/12	Stoke	Man City	1	1	D	1	1	D	M Clattenburg	4	15	3	11	12	12	6	5	0	1	0	0	5	3.6	1.73	5
E0	15/09/12	Sunderland	Liverpool	1	1	D	1	0	H	M Atkinson	6	20	4	9	18	10	2	7	1	2	0	0	3.4	3.3	2.2	3.3

Figure 2: Dataset.

B. MLR Model Results

```
Call:
lm(formula = Credits ~ W + D + L + Agoal + Lgoal + S + ST + F +
    C + Y + R + BF, data = All_data)

Residuals:
     Min       1Q   Median       3Q      Max
-18.9891  -4.7389  -0.0818   4.8132  18.2781

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.34387    8.86425   0.377  0.70639
W            4.22532    0.54841   7.705 5.28e-13 ***
D            1.58202    0.39521   4.003 8.70e-05 ***
L                 NA         NA      NA       NA
Agoal        0.13406    0.14684   0.913  0.36234
Lgoal       -0.12414    0.15177  -0.818  0.41434
S            0.06976    0.02415   2.888  0.00428 **
ST           0.02774    0.03484   0.796  0.42679
F           -0.03675    0.02621  -1.402  0.16228
C            0.06996    0.04688   1.492  0.13712
Y            0.13018    0.08822   1.476  0.14155
R            0.46921    0.47527   0.987  0.32466
BF          -0.01437    0.01990  -0.722  0.47103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.081 on 208 degrees of freedom
Multiple R-squared:  0.8519,    Adjusted R-squared:  0.8441
F-statistic: 108.8 on 11 and 208 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Credits ~ W + D + S, data = All_data)

Residuals:
     Min       1Q   Median       3Q      Max
-19.0149  -5.3795   0.4652   4.9793  18.0962

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.83594    2.82964  -2.062   0.0404 *
W            4.75354    0.23773  19.996  < 2e-16 ***
D            1.87217    0.29982   6.244 2.23e-09 ***
S            0.10997    0.01591   6.910 5.35e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.092 on 216 degrees of freedom
Multiple R-squared:  0.8457,    Adjusted R-squared:  0.8436
F-statistic: 394.8 on 3 and 216 DF,  p-value: < 2.2e-16
```

Figure 3: Model 1 & 2.

```
Call:
lm(formula = Credits ~ W + D + Lgoal + STR + GTR + F + C + Y +
    R + BF, data = new1_All_data)

Residuals:
     Min       1Q   Median       3Q      Max
-20.1394  -4.7430   0.0159   4.7158  19.3817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.0294     3.3071  16.942  < 2e-16 ***
W            16.5933     1.4120  11.752  < 2e-16 ***
D             3.7294     0.7160   5.209 4.60e-07 ***
Lgoal         0.6393     0.9301   0.687    0.493
STR           0.3391     0.5667   0.598    0.550
GTR         -30.1677    29.4045  -1.026    0.306
F            -0.6742     0.5549  -1.215    0.226
C             3.4854     0.6146   5.671 4.77e-08 ***
Y             0.8259     0.5689   1.452    0.148
R1           -1.0666     1.2350  -0.864    0.389
R2           -0.7033     1.3328  -0.528    0.598
R3            1.0868     2.0426   0.532    0.595
R4            6.0700     3.7828   1.605    0.110
BF           -0.4857     0.5095  -0.953    0.342
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.285 on 206 degrees of freedom
Multiple R-squared:  0.8448,    Adjusted R-squared:  0.835
F-statistic: 86.23 on 13 and 206 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Credits ~ W + D + Lgoal + STR + GTR + F + C + Y +
    R + BF, data = new2_All_data)

Residuals:
     Min       1Q   Median       3Q      Max
-20.1394  -4.7430   0.0159   4.7158  19.3817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6469     5.2246   1.081    0.281
W            83.2514     7.0842  11.752  < 2e-16 ***
D            20.6401     3.9626   5.209 4.60e-07 ***
Lgoal         3.5410     5.1518   0.687    0.493
STR           1.7769     2.9694   0.598    0.550
GTR         -30.1677    29.4045  -1.026    0.306
F            -3.5777     2.9448  -1.215    0.226
C            20.2220     3.5661   5.671 4.77e-08 ***
Y             4.5462     3.1316   1.452    0.148
R1           -1.0666     1.2350  -0.864    0.389
R2           -0.7033     1.3328  -0.528    0.598
R3            1.0868     2.0426   0.532    0.595
R4            6.0700     3.7828   1.605    0.110
BF           -2.8951     3.0374  -0.953    0.342
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.285 on 206 degrees of freedom
Multiple R-squared:  0.8448,    Adjusted R-squared:  0.835
F-statistic: 86.23 on 13 and 206 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Credits ~ W + D + Lgoal + STR + GTR + F + C + Y +
    R + BF, data = new3_All_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.239755 -0.056464  0.000189  0.056140  0.230735

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.123251   0.062198  -1.982   0.0489 *
W            0.991088   0.084335  11.752  < 2e-16 ***
D            0.245715   0.047173   5.209 4.60e-07 ***
Lgoal        0.042155   0.061331   0.687   0.4926
STR          0.021154   0.035351   0.598   0.5502
GTR         -0.359139   0.350054  -1.026   0.3061
F           -0.042592   0.035057  -1.215   0.2258
C            0.240738   0.042454   5.671 4.77e-08 ***
Y            0.054122   0.037281   1.452   0.1481
R1          -0.012698   0.014702  -0.864   0.3888
R2          -0.008372   0.015867  -0.528   0.5983
R3           0.012938   0.024317   0.532   0.5953
R4           0.072262   0.045033   1.605   0.1101
BF          -0.034465   0.036159  -0.953   0.3416
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08673 on 206 degrees of freedom
Multiple R-squared:  0.8448,    Adjusted R-squared:  0.835
F-statistic: 86.23 on 13 and 206 DF,  p-value: < 2.2e-16
```

Figure 4: Model 3, 4 & 5.

C. MLR Model Prediction

Team	Prediction
Liverpool	84.8610359973753
Arsenal	80.3981426428757
Man City	77.4112528042599
Aston Villa	76.2804401146346
Tottenham	71.7932835450727
Brighton	65.4557154934821
Man United	64.4651537677675
Newcastle	60.4423528353376
Everton	54.8326885620026
West Ham	54.1469158625298
Chelsea	51.3520260376589
Fulham	50.948560163298
Brentford	49.7468809715133
Bournemouth	46.1825628024885
Wolves	45.8869816459903
Crystal Palace	40.7337663475764
Nott'm Forest	36.9861277687981
Burnley	29.7701037903299
Luton	28.0290494475282
Sheffield United	23.6973240209496

Team	Prediction
Liverpool	84.7346411891064
Arsenal	78.7934063027455
Aston Villa	76.2613855320309
Man City	75.0296338294697
Tottenham	71.4858106539775
Brighton	64.0734480487234
Man United	63.6697990440501
Newcastle	59.5813103165046
Everton	56.8073096832551
West Ham	53.4006994138117
Brentford	50.604668215811
Chelsea	49.2849724291092
Fulham	48.3172311538341
Bournemouth	47.0854794512729
Wolves	44.9959611223284
Crystal Palace	42.0020111912953
Nott'm Forest	36.4812481819533
Luton	27.4334848694646
Burnley	26.11118523656
Sheffield United	22.3720471742383

Figure 5: Prediction of Model 1 & 2.

Team	Prediction
Liverpool	88.0048399944177
Arsenal	81.2345381830642
Aston Villa	74.1850298445887
Man City	73.555930209306
Tottenham	67.6556510485085
Man United	64.4037314225878
Brighton	63.1775155233967
Newcastle	54.2585265961147
West Ham	53.0977038131417
Everton	48.6138753084364
Fulham	48.569664089914
Chelsea	47.020703991938
Bournemouth	46.2660718300574
Brentford	45.2960589477062
Wolves	44.807623446093
Crystal Palace	38.2269876760492
Nott'm Forest	34.5198401250419
Burnley	29.0335846285796
Luton	27.1943414551358
Sheffield United	21.4455444096985

Team	Prediction
Liverpool	122.773448920603
Arsenal	116.571903274242
Aston Villa	105.99610175363
Man City	103.910913196812
Tottenham	92.5221953724945
Man United	89.2736757097921
Brighton	83.5418427913291
Newcastle	72.806910756503
West Ham	69.9174573035227
Everton	63.1826360313858
Fulham	61.9045095620222
Chelsea	57.5290051001169
Bournemouth	56.5932564739253
Brentford	54.5491854456838
Wolves	53.3826672993868
Crystal Palace	43.0183710645113
Nott'm Forest	36.0022542279186
Burnley	27.1299731166892
Luton	24.7486919987967
Sheffield United	16.3680562945351

Team	Prediction
Liverpool	122.773448920603
Arsenal	116.571903274242
Aston Villa	105.99610175363
Man City	103.910913196812
Tottenham	92.5221953724945
Man United	89.2736757097921
Brighton	83.5418427913291
Newcastle	72.806910756503
West Ham	69.9174573035227
Everton	63.1826360313858
Fulham	61.9045095620222
Chelsea	57.5290051001169
Bournemouth	56.5932564739253
Brentford	54.5491854456838
Wolves	53.3826672993868
Crystal Palace	43.0183710645113
Nott'm Forest	36.0022542279186
Burnley	27.1299731166892
Luton	24.7486919987967
Sheffield United	16.3680562945351

Figure 6: Prediction of Model 3, 4 & 5.

D. Prediction of Each Game

Season	HomeTeam	AwayTeam	match	ave_home_scored	ave_away_scored	score_line	prob
2012-2013	Arsenal	Aston Villa	1	2	1	1-3	8.31
2012-2013	Arsenal	Chelsea	1	1	2	0-3	8.46
2012-2013	Arsenal	Everton	1	0	0	0-3	8.35
2012-2013	Arsenal	Fulham	1	3	3	0-3	8.47
2012-2013	Arsenal	Liverpool	1	2	2	1-2	8.61
2012-2013	Arsenal	Man City	1	0	2	0-3	8.6
2012-2013	Arsenal	Man United	1	1	1	1-3	8.37
2012-2013	Arsenal	Newcastle	1	7	3	0-2	8.63
2012-2013	Arsenal	Norwich	1	3	1	1-2	8.69
2012-2013	Arsenal	QPR	1	1	0	0-3	8.71
2012-2013	Arsenal	Reading	1	4	1	1-3	8.53
2012-2013	Arsenal	Southampton	1	6	1	0-2	8.14
2012-2013	Arsenal	Stoke	1	1	0	1-2	8.49
2012-2013	Arsenal	Sunderland	1	0	0	0-2	8.8
2012-2013	Arsenal	Swansea	1	0	2	0-2	8.38
2012-2013	Arsenal	Tottenham	1	5	2	0-2	8.55
2012-2013	Arsenal	West Brom	1	2	0	1-2	8.62
2012-2013	Arsenal	West Ham	1	5	1	1-3	8.35
2012-2013	Arsenal	Wigan	1	4	1	0-3	9.1
2012-2013	Aston Villa	Arsenal	1	0	0	1-1	10.52
2012-2013	Aston Villa	Chelsea	1	1	2	0-0	100
2012-2013	Aston Villa	Everton	1	1	3	0-1	13.73
2012-2013	Aston Villa	Fulham	1	1	1	1-0	27.45
2012-2013	Aston Villa	Liverpool	1	1	2	2-0	27.69
2012-2013	Aston Villa	Man City	1	0	1	1-1	13.92
2012-2013	Aston Villa	Man United	1	2	3	0-0	100
2012-2013	Aston Villa	Newcastle	1	1	2	2-0	22.69

Figure 7: Prediction of Model 3, 4 & 5.