PREDICTING THE OUTCOMES OF FOOTBALL MATCHES USING MACHINE LEARNING TECHNIQUES

Authored by:

Ms Nicole D'Silva*
Mr Ganesh Narayanan**
**Assistant Professor, International Institute of Sports Management*
Mumbai, India
nicole.dsilva@iismworld.com
***Master of Sports Management*
International Institute of Sports Management
Mumbai, India
ganesh.0102@outlook.com

ABSTRACT

Analytics has been playing a major role in the arena of Sports. Machine Learning techniques are widely used in Sports as a part of predictive analytics for predicting the outcome of a match. Today, predictive analytics has been instrumental in predicting the outcome of a professional football match with the help of historical data. In this research, based on the number of goals and with the full-time results, the outcome of the football match is predicted using Generalised Linear Model and Naïve Bayes Model, respectively. The key objective of this project is to explore various Machine Learning methods to predict the outcome of some famous English Premier League derbies.

Keywords: *Predictive Analytics; Football predictions, Machine Learning; GLM in R; Naïve Bayes in R; Premier League Predictions*

INTRODUCTION

Analytics in Football

Football is widely played across the globe and has always been followed very closely by many people. After the emergence of Analytics in sports, in recent years, match data have been collected across all the major top-flight competitions for analyzing the various intricate details of the match.

There is a difficulty in predicting the outcomes of sports matches owing to its unpredictability. Football is one such sport in which the matches have a fixed duration of 90 minutes plus the injury time. In any knock-out matches, if the match results in a tie, an additional 30 minutes is given. Post 30 minutes, in case the match

results in a tie, the match ends up with penalties. Such instances make the matches played in a football sport highly unpredictable.

In a football match, there are mostly three possible outcomes viz., win, lose or a draw. The traditional predictive methods have simply used match results to evaluate team performance and build statistical models to predict the results of future games.

The collection of the data has placed Data Science at the forefront of the football industry with many possible uses and applications which includes calculation of betting odds, identifying players' playing styles, match strategy, tactics, analysis; player acquisition, player valuation, team spending; Performance management and predictions. Thus, analytics is a widely used mechanism to determine the outcomes of a match.

However, the analytical tool such as the R CRAN helps in accurately predicting the outcomes of matches in a highly unpredictable sport such as football. A potential solution to this problem can be explored by using in-game statistics to dive deeper than the simple match results. In the last few years, in-depth match statistics have been made available, which creates the opportunity to look further. The emergence of new Machine Learning techniques in recent years allows for better predictive performance in a wide range of classification and regression problems. The exploration of these different methods and algorithms have enabled the development of better models in both predicting the outcome of a match and the actual score.

English Premier League

The Premier League is the top tier of England's football pyramid, with 20 teams battling it out for the honour of being crowned English champions. Home to some of the most famous clubs, players, managers and stadiums in world football, the Premier League is the most-watched league on the planet with one billion homes watching the action in 188 countries. The league takes place between August and May and involves the teams playing each other home and away across the season, a total of 380 matches.

Since the League began in 1992, there have been six different winners: Manchester United, Arsenal, Chelsea, Manchester City, Blackburn Rovers and Leicester City. Manchester Utd has had the most success with 13 titles in the 25 seasons so far. Manchester City has the Premier League record for the biggest winning margin when they finished 19 points ahead of second-placed Manchester United in 2017/18. (1)

METHODOLOGY

Machine Learning Techniques

Generalized Linear Models

Generalized Linear Models are a set of regression methods for which the output value is assumed to be a linear combination of all the input values. The mathematical formulation of the regression problem is:

$$y = {}_0 + {}_1x_1 + \ldots + {}_mx_m + \text{✏}$$

where:

- y is the observed value/dependent variable
- the x_k are the input values/predictor variables
- the ${}_k$ is the weights that have to befound
- ✏ is a Gaussian-distributed error variable

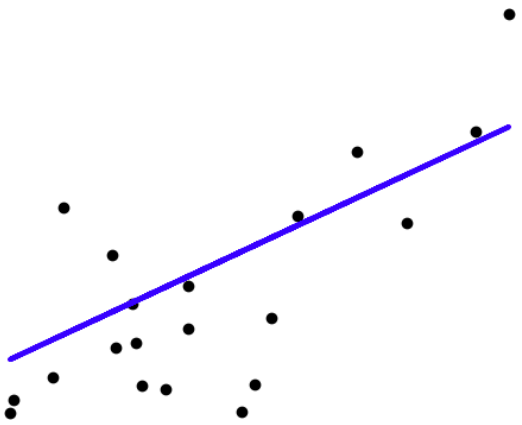


Figure 1: Generalized Linear Model

Generalized Linear Models are widely known and used technique in analytics. It is known for its ease of implementation and, in many classifications or regression problems, assuming linearity between predictor variables and the outcome variable is sufficient to generate accurate predictions. In addition to this, the fitted coefficients are interpretable, which implies, we can understand the direction and magnitude of the association between each predictor variable and the outcome variable. However, using Generalized Linear Models has its setbacks, significant among them being, the assumption that the input variables and output variable are linearly connected does not always fit the problem and can be too simple. Furthermore, if the input variables are highly correlated, the performance of the model can be quite poor.

Naïve Bayes Model

The reason that the Naive Bayes algorithm is called Naïve is that the algorithm makes a very strong assumption about the data having features independent of each other while in reality, they may be dependent in some way. In other words, it assumes that the presence of one feature in a class is completely unrelated to the presence of all other features. If this assumption of independence holds, Naive Bayes performs extremely well and often better than other models. Naive Bayes can also be used with continuous features but is more suited to categorical variables. If all the input features are categorical, Naive Bayes is recommended. However, in the case of numeric features, it makes another strong assumption which is that the numerical variable is normally distributed.

R supports a package called 'e1071' which provides a naive Bayes training function. As we know, Bayes theorem is based on conditional probability and uses the formula

P (A | B) = P(A) * P (B | A) / P(B)

We now know how this conditional probability comes from multiplication of events so if we use the general multiplication rule, we get another variation of the theorem that is, using P(A AND B) = P(A) * P(B | A), we can obtain the value for conditional probability: P(B | A) = P(A AND B) / P(A) which is the variation of Bayes theorem.

Since P (A AND B) also equals P(B) * P (A | B), we can substitute it and get back the original formula: P (B | A) = P(B) * P (A | B) / P(A)

The most significant advantages of Naïve Bayes are that they are quick and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real-life cases, the predictors are dependent, this hinders the performance of the classifier. (2)

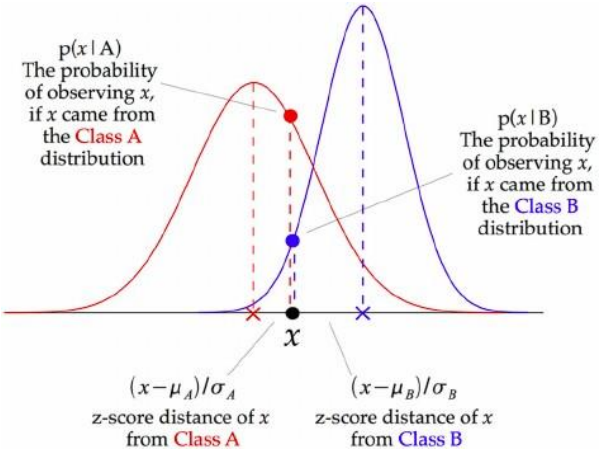


Figure 2: Gaussian Naive Bayes method

DATASET

We have obtained the dataset from the football data portal website. The website has the data from 1993 to date. The data includes detailed match events such as home team, away team, half time result, full-time result, goals, fouls, total shots, shots on target etc. (3)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	Div	Date	Time	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	BWD	BWA
2	E0	09-08-2019	20:00	Liverpool	Norwich	4	1	H	4	0	H	M Oliver	15	12	7	5	9	9	11	2	0	2	0	0	1.14	10	19	1.14	8.25	18.5
3	E0	10-08-2019	12:30	West Ham	Man City	0	5	A	0	1	A	M Dean	5	14	3	9	6	13	1	1	2	2	0	0	12	6.5	1.22	11.5	5.75	1.26
4	E0	10-08-2019	15:00	Bournemouth	Sheffield United	1	1	D	0	0	D	K Friend	13	8	3	3	10	19	3	4	2	1	0	0	1.95	3.6	3.6	1.95	3.6	3.9
5	E0	10-08-2019	15:00	Burnley	Southampton	3	0	H	0	0	D	G Scott	10	11	4	3	6	12	2	7	0	0	0	0	2.62	3.2	2.75	2.65	3.2	2.75
6	E0	10-08-2019	15:00	Crystal Palace	Everton	0	0	D	0	0	D	J Moss	6	10	2	3	16	14	6	2	2	1	0	1	3	3.25	2.37	3.2	3.2	2.35
7	E0	10-08-2019	15:00	Watford	Brighton	0	3	A	0	1	A	C Pawson	11	5	3	3	15	11	5	2	0	1	0	0	1.9	3.4	4	1.9	3.4	4.33
8	E0	10-08-2019	17:30	Tottenham	Aston Villa	3	1	H	0	1	A	C Kavanagh	31	7	7	4	13	9	14	0	1	0	0	0	1.3	5.25	10	1.3	5.5	10
9	E0	11-08-2019	14:00	Leicester	Wolves	0	0	D	0	0	D	A Marriner	15	8	1	2	3	13	12	3	0	2	0	0	2.2	3.2	3.4	2.25	3.3	3.3
10	E0	11-08-2019	14:00	Newcastle	Arsenal	0	1	A	0	0	D	M Atkinson	9	8	2	2	12	7	5	3	1	3	0	0	4.5	3.75	1.72	4.5	3.75	1.78
11	E0	11-08-2019	16:30	Man United	Chelsea	4	0	H	1	0	H	A Taylor	11	18	5	7	15	13	3	5	3	4	0	0	2.1	3.3	3.5	2.15	3.3	3.5
12	E0	17-08-2019	12:30	Arsenal	Burnley	2	1	H	1	1	D	M Dean	16	18	9	5	13	11	10	7	2	1	0	0	1.3	5.5	10	1.3	5.5	10
13	E0	17-08-2019	15:00	Aston Villa	Bournemouth	1	2	A	0	2	A	M Atkinson	22	12	7	4	10	13	10	5	0	2	0	0	2.3	3.4	3.1	2.25	3.5	3.1
14	E0	17-08-2019	15:00	Brighton	West Ham	1	1	D	0	0	D	A Taylor	16	8	4	3	11	10	8	6	0	2	0	0	2.55	3.25	2.87	2.5	3.3	2.9
15	E0	17-08-2019	15:00	Everton	Watford	1	0	H	1	0	H	L Mason	12	8	2	2	11	11	4	7	2	3	0	0	1.72	3.8	4.75	1.67	4	5
16	E0	17-08-2019	15:00	Norwich	Newcastle	3	1	H	1	0	H	S Attwell	15	10	8	3	9	11	7	5	1	3	0	0	2.25	3.3	3.3	2.25	3.3	3.3
17	E0	17-08-2019	15:00	Southampton	Liverpool	1	2	A	0	1	A	A Marriner	14	15	3	6	10	6	5	9	2	1	0	0	6.5	4.75	1.44	6.25	4.75	1.48
18	E0	17-08-2019	17:30	Man City	Tottenham	2	2	D	2	1	H	M Oliver	30	3	10	2	14	4	13	2	1	0	0	0	1.36	5.25	8	1.35	5.5	7.75
19	E0	18-08-2019	14:00	Sheffield United	Crystal Palace	1	0	H	0	0	D	D Coote	15	6	3	4	16	11	8	4	3	1	0	0	2.55	3.1	2.9	2.55	3.2	2.9
20	E0	18-08-2019	16:30	Chelsea	Leicester	1	1	D	1	0	H	O Langford	14	12	5	3	9	14	4	5	1	0	0	0	1.7	3.75	5	1.72	3.7	5
21	E0	19-08-2019	20:00	Wolves	Man United	1	1	D	0	1	A	J Moss	6	9	2	2	17	8	4	6	2	2	0	0	3.3	3.3	2.25	3.3	3.25	2.3

Legends			
FTHG	*Full-Time Home Goal*	***HC***	*Home Crosses*
FTAG	*Full-Time Away Goal*	***AC***	*Away Crosses*
FTR	*Full-Time Result*	***HY***	*Home Yellow*
HTHG	*Half Time Home Goal*	***AY***	*Away Yellow*
HTAG	*Half Time Away Goal*	***HR***	*Home Red*
HTR	*Half Time Result*	***AR***	*Away Red*
HS	*Home Shots*	***B365H***	*Bet 365 Home*
AS	*Away Shots*	***B365D***	*Bet 365 Draw*
HST	*Home Shots on Target*	***B365A***	*Bet 365 Away*
AST	*Away Shots on Target*	***BWH***	*Betway Home*
HF	*Home Fouls*	***BWD***	*Betway Draw*
AF	*Away Fouls*	***BWA***	*Betway Away*

The following table is built from the data that is obtained from the Premier League website https://www.premierleague.com/tables and it is used for Generalized Linear Model.

Team	Matches	Shots	ST	Crosses	CA	AVGS	AVGST	Key Cross
Man Utd	1068	8164	2923	12497	23%	7.64	2.74	2.69
Liverpool	1068	8895	3008	11843	22%	8.33	2.82	2.44
Man City	876	8400	2954	11017	22%	9.59	3.37	2.77

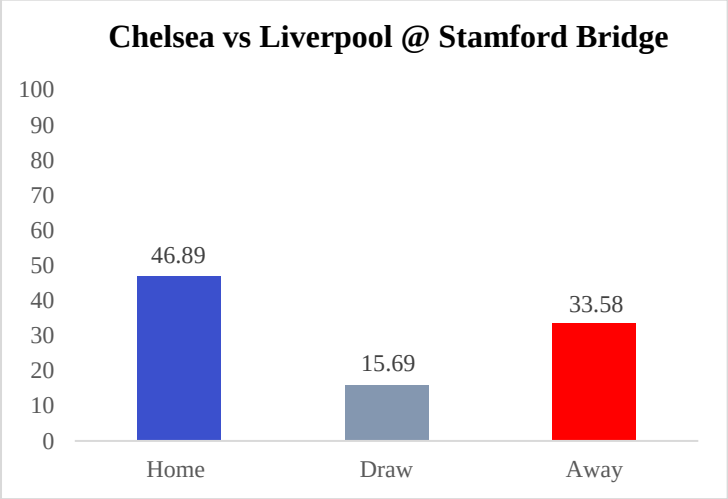
Chelsea	1067	8970	3034	11697	23%	8.41	2.84	2.52
Arsenal	1066	8127	2935	11440	21%	7.62	2.75	2.25
Spurs	1067	8294	2851	11440	22%	7.77	2.67	2.36
Everton	1067	7038	2390	11873	23%	6.60	2.24	2.56
Newcastle	949	5549	1778	9561	22%	5.85	1.87	2.22
Sunderland	608	4532	1402	7994	22%	7.45	2.31	2.89

Table 1: Premier League Table

Generalized Linear Model's Output

Team	Stadium	Home	Draw	Away
Chelsea vs Liverpool	Stamford Bridge	46.89	15.69	33.58

Table 2: GLM Predictions for Chelsea vs. Liverpool

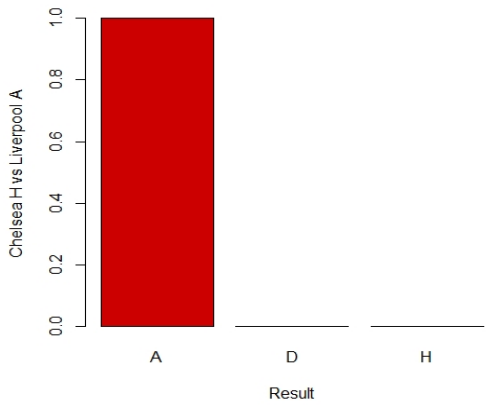


Analysis

At Stamford Bridge, Chelsea has a 46.89% winning rate, whereas Liverpool has 33.58% chance of winning. In the last five meetings at Stamford Bridge, Chelsea have won only once in May 2018, whereas Liverpool has won thrice. (4)

Figure 3: Graph of GLM Predictions for Chelsea vs. Liverpool

Naïve Bayes Model's Output



Analysis

According to the Naïve Bayes model method, we can predict that Liverpool will be winning against Chelsea at Stamford Bridge. This is mainly because of Liverpool's recent success at Stamford Bridge over the past few years. At Stamford Bridge, in the past ten meetings against Liverpool, Chelsea have won just twice. (4)

Figure 4: Graph of NBM Predictions for Chelsea vs. Liverpool

Team	Stadium	Home	Draw	Away
Chelsea vs Liverpool	Stamford Bridge			✓

Table 3: NBM Predictions for Chelsea vs. Liverpool

CONCLUSION

Football is the most-watched, famous sport in the world. The English Premier League is famous across the globe. The English clubs playing in the Premier League have an annual TV rights deal worth over $9 billion, which is the highest paid licensing contract in sports at present. The English Premier League is followed by billions around the world not only because of its recognition and the big-name players playing there, but it is widely popular for the great extent of uncertainty and unpredictability that it carries. This sheer uncertainty and unpredictability of the matches played in the Premier League give scope for Predictive Analytics. The very purpose of this project was to explore the different Machine Learning techniques to predict the outcome of football matches played in the English Premier League using the statistics available on the League's websites and other football forums. Using the Generalized Linear Model and Naïve Bayes Theorem, we could predict the outcomes of the different matches played by the famous Football Clubs in the English Premier League evaluating the attributes that lead a football team to lose, draw or win the match. Different team metrics like how many points each team got and what place each team finished were being kept track of and those variables were used to analyze and compare our models with the available data. We got each team's goal-scoring rate at home and away from home using the Generalized Linear Model. By following a probabilistic approach using Naïve Bayes Model, we can predict the full-time results of the matches. In this research, we have used two machine learning models to find the predictions. By using Generalized Linear Model, we have considered the total goals as the key factor in predicting the outcome. We have taken data from the inception of the Premier League, but the limitation is that the current form of the teams varies a lot. Teams like Manchester United were successful in the past but their present form is nowhere nearby that. Therefore, the accuracy of Generalized Linear Model is lesser compared to the Naïve Bayes Model which predicts the outcomes based on the Full-time results of the matches. This provides an effective predictive reference for the managers, players, and the supporters of Football. An overall observation indicates that the predicted results have a consistent trend, which indicates that the result is reasonable and acceptable.

LIMITATIONS

Every research study is faced with certain challenges and limitations, and the present study is no exception to it. The following were the major challenges, which were on the path of achieving the desired objectives:

Lack of adequate and quality data: The present research required immense data with the with the required statistical depth to predict the full-time result metrics. However, due to privacy concerns, most of the data regarding the football matches are not made publicly available which leads to difficulties of accessing the same. Hence, the suitable data required for the present research were obtained and refined from different public football database. We would be required to refine various public football databases to find one that is suitable for us to use.

Research & Analysis in Predictive Analytics: With advancement in the field of Predictive Analytics, it was crucial to carry out a detailed background research of prediction techniques for designing the models and testing different hypotheses and to develop a mathematical understanding of various Machine Learning algorithms that can be used for the predictions.

FURTHER RESEARCH

Predictive Analytics in Football is still in the nascent stages. Football has more variables which if further researched and with massive amount of data, it will pave way for experts in the field of sports analytics and to the bookmakers to race with probabilities and be more accurate. With further research we aim to come up with the advanced version of the repost soon.

REFERENCES

1. Premier League. premierleague.com. [Online] 2020. https://www.premierleague.com/premier-league-explained.

2. Predicting Football Results using Machine Learning Techniques. Herbinet, Corentin. s.l. : Imperial College London, 2018, Imperial College of London, p. 74.

3. Footballdata. Football-data.co.uk. [Online] 2020. http://www.football-data.co.uk/englandm.php.

4. 11v11. 11v11.com. [Online] 2020. https://www.11v11.com/teams/liverpool/tab/opposingTeams/opposition/Chelsea/.

5. https://www.tips180.com/

6. https://www.scribd.com/document/410412952/Lapse-Team

7. https://www.r-bloggers.com/understanding-naive-bayes-classifier-using-r/