Predicting Football Match Results with Logistic Regression

Darwin Prasetio
Department of Informatics Engineering
Institut Teknologi Bandung
Bandung, Indonesia
prasetiodarwin@gmail.com

Dra. Harlili, M.Sc.
Department of Informatics Engineering
Institut Teknologi Bandung
Bandung, Indonesia
harlili@informatika.org

***Abstract*— Many efforts has been made in order to predict football matches result and selecting significant variables in football. Prediction is very useful in helping managers and clubs make the right decision to win leagues and tournaments. In this paper a logistic regression model is built to predict matches results of Barclays' Premier League season 2015/2016 for home win or away win and to determine what are the significant variable to win matches. Our work is different from others as we use only significant variables gathered from researches in the same field rather than making our own guess on what are the significant variables. We also used data gathered from video game FIFA, as Shin and Gasparyan [8] showed us that including data from the video game could improve prediction quality. The model was built using variations of training data from 2010/2011 season until 2015/2016. Logistic regression is a classification method which can be used to predict sports results and it can gives additional knowledge through regression coefficients. The variables used are "Home Offense", "Home Defense", "Away Offense", and "Away Defense". We conducted experiments by altering seasons of training data used. Prediction accuracy of built model is 69.5%. We found our work improved significantly compared to the one Snyder [9] did. We concluded that the significant variables are "Home Defense", and "Away Defense".**

Keywords—regression coefficients, variables, logistic regression, prediction accuracy.

I. INTRODUCTION

Football match result prediction is such a huge phenomenon. A lot of fans and analyst always give prediction about who is going to win the match before it starts. Prediction is done by calculating a number of variables, such as home advantage, recent team performance, team strength, and other variables [2]. Prediction can be used by managers and club directors to decide what is needed to win matches [7]. It becomes very important because the money involved in football is getting bigger by the years. On 2012/2013 season, total money spent by Premier League Clubs was approximately 918.7 million poundsterling, and it kept growing until 2015/2016, which was more than 1 billion poundsterling [10].

One method can be used to solve prediction problem is logistic regression [4][7]. Logistic regression is a classification method where its dependent variable has only 2 possible value, such as a student graduating or not, someone was diagnosed with a disease or not [6]. Logistic regression model is used to predict probability of the binary response based on one or more variables. This method's advantages are it's very suitable to explain the relationship between output variable and input, and it can solve the problem which ordinary least squares regression cannot, which is the plotting of data will form a S-shaped curve that cannot be explained with a linear equation [6].

In this paper, we propose a model of football match prediction with data from Barclays' Premier League and sofifa.com using 4 variables, Home Offense, Home Defense, Away Offense, and Away Defense. The 4 variables are most mentioned as significant by other researchers. We implement this method in a software called *Football Predictor*. Our work predicts who is going to win a match (home/away), and give information on the odds, probability, and the regression coefficients. Users can also add their own training data to the application to alter the prediction model.

Second section of this paper will discuss related work on this field, serving as some of the base for the proposed method. Section 3 explains our proposed method to solve the problem, and section 4 elaborate our experiments and result analysis. The last section contains our conclusion on football match prediction using logistic regression.

II. RELATED WORK

II.1 Predicting Football With Data from FIFA Football Game

Shin and Gasparyan [8] conducted a research on predicting football results by mixing real data and data collected from football video game, FIFA 2015. They integrated real data with the players' attribute data they collected from FIFA 2015 such as heading, passing, shooting, and strength, to predict matches results and it shows better prediction results compared to only real data. They also stated that the method of using data from video games can save a lot of time and energy as it can be very expensive for some data to be calculated or collected from the real world.

II.2 Predicting Matches Results

There has been several previous works on predicting football match results, such as in [1][2][4][5][7][8][9].

Bailey [1] predicted Australian Football League using match records of 100 seasons prior to 1997, and tested it to matches from season 1997 until 2003 using multiple linear

978-1-5090-1636-5/16/$31.00 ©2016 IEEE

regression model. He used home advantage, travel fatigue measured from distance traveled by the away team, ground familiarization, and measures of team quality and current form. He built 3 models, one named as benchmark, one used team measures, and one used individual measures. His work produced 66.7% accuracy and he stated significant variables are team's attack strength, home advantage, traveled distance, and ground familiarization.

Baio [2] proposed Bayesian hierarchical model to predict matches outcome to predict 1991-1992 Serie A matches so that he can compare his work to Karlis & Ntzoufras (2003). He chose the variables home advantage, team attack, and team defence. His work produced each team's effect of attack and defence, and shown that the team with the biggest attack won the league at the end of the season while for the predictive accuracy, it is good for the usual team but for the relegated team, it is very poor because it overly estimate the team's performance. Min [5] also used the same technique, but he added rule-based reasoning, and using not only the score or results data, but also work rate, aggressiveness, and others. However, he reached the similar conclusion, that the significant variables are attack and defense. Another variables were ball possession and traveled distance.

Igiri [4] used 9 sets of features from home and away data, which is goals, shorts, corner, odds, attack strength, players' performance index, managers' performance index, managers' win, and teams' win streak. He employed ANN and logistic regression in his research. He built his model and tested it using English Premier League season 2014-2015 matches records, and produced a very high prediction accuracy with logistic regression, 95%. Reddy [7] also included logistic regression in one of the methods he used to predict English Premier League season 2012-2013 matches. His research concluded that the significant variables were number of away goals scored, number of red cards, home team position, and shots on target.

Snyder [9] conducted a research to predict Barclays' Premier League season 2011/2012 matches by using all matches in season 2010/2011. He used a lot of variables, stadium capacity, distance traveled by a team before match, and statistics of a team on previous season, including ranking, amount of wins, draws, losses, scored and conceded goals, goals difference in each match, points, money spent on players' wages, money spent on 2011 summer transfer market, and number of games a manager of the team has played in the league. He built his model with logistic regression and its prediction accuracy was 51.06%. He also attempted to guess what was the deciding factor in football matches. Out of all the variables he used, he concluded that it were the 2 previous matches, some events that occur a few times in a game, and player evaluation in attack, defence, midfield, and goalkeeper were the most important factor.

Table I shows a summary of all variables stated significant by researchers mentioned.

Table I. Significant variables

Num.	Variables	Researcher(s)
1.	Attack (Integer)	1. Shin [8] 2. Bailey [1] 3. Min [5] 4. Baio [2] 5. Snyder [9]
2.	Home advantage (Real)	1. Shin [8] 2. Bailey [1] 3. Baio [2] 4. Reddy [7]
3.	Traveled distance (Real)	1. Bailey [1] 2. Min [5]
4.	Defence (Integer)	1. Shin [8] 2. Baio [2] 3. Min [5] 4. Bailey [1] 5. Snyder [9]
5.	*Ground Familiarisation (Real)*	1. Bailey [1]
6.	*Ball Possession (Real)*	1. Min [5]
7.	Players' index (Integer)	1. Igiri [4]
8.	Managers' index (Integer)	1. Igiri [4]
9.	Away Goal Scored (Integer)	1. Reddy [7]
10.	Red Card (Integer)	1. Reddy [7]
11.	Shot on Goals (Integer)	1. Reddy [7]
12.	Goalkeepers' Skill (Integer)	1. Shin [8] 2. Snyder [9]
13.	Midfield (Integer)	1. Snyder [9]

III. PROPOSED METHOD

Logistic regression is a classification method where the dependent variable can only have 2 possible values [3]. It is chosen over linear regression because data plotting shows that it does not suit linear regression pattern [6]. In this paper, we chose 4 variables from Table I referenced by 3 or more researchers in the logistic regression equation. Although home

advantage is referenced by 4 researchers, it is not used because data on home advantage is not readily available. The logistic regression equation then will be

$$\Pi(x) = 1/1+ e^{-y} \qquad (1)$$

Where $y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4$ and e being Euler's number. β_0 is the constant, β_1 is coefficient for variable X_1 (Home Attack), β_2 coefficient for X_2(Home Defense), β_3 coefficient for variable X_3(Away Offense), and β_4 coefficient for variable X_4(Away Defense). If $\Pi(x)$ shows value greater than 0.5, it will be classified as Home Win, otherwise it's an Away Win.

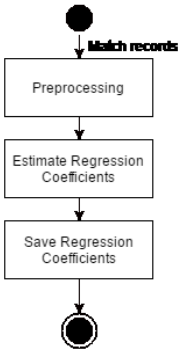

Figure 1. Proposed Training Process Flow

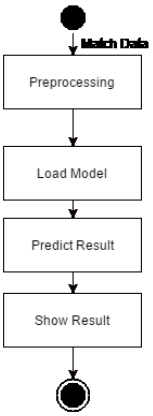

Figure 2. Proposed Testing Process Flow

We separated our work into 2 process flows, one for testing, and one for training. Proposed training process flow can be seen at Figure 1, and proposed testing flow can be seen at Figure 2. We obtained our training and testing data from http://www.football-data.co.uk/, and each team strength from http://sofifa.com/teams. Training process flow is used to estimate coefficients. We built a logistic regression model using Newton-Raphson algorithm to estimate the coefficients (β_0- β_4). Obtained match data from 2010/2011 season until 2015/2016 season is then stored in CSV (comma separated files). After the file is loaded, it will go through preprocessing, in which all data will be loaded and then parsed into its own variables. Home win will be recorded as 1, and away win will be recorded as 0. The next step is estimating regression coefficients, and after the coefficients is estimated, it will be saved on an external .txt file, so it is unnecessary to repeat the process of training if no alteration in training data.

To test the accuracy, we used match records from 2015/2016. Testing process flow can be seen at Figure 2. Match data then go through preprocessing. Using regression coefficients obtained earlier, each match is then predicted and compared to the real result to count the predictive accuracy.

IV. EXPERIMENT

Experiment is conducted to get the best predictive accuracy. In this paper, it is done by using various amount of training data. It is to determine whether amount of training data has any effect on prediction accuracy. We conducted 4 experiments, which details can be seen at Table II.

Table II. Experiments Data

Num	Training Data	Testing Data
1	Match records season 2010/2011 – 2014/2015	Match records season 2015/2016
2	Match records season 2011/2012 – 2014/2015	Match records season 2015/2016
3	Match records season 2010/2011 – 2015/2016	Match records season 2015/2016
4	Match records season 2011/2012 – 2015/2016	Match records season 2015/2016

Each of the testing result produced a different set of coefficients, and the prediction accuracy also varies. Test result for can be seen at Table III. The accuracy comparison is summarised in Figure 3.

Table III. Experiment Result

Num.	Variables	Regression coefficients	Prediction accuracy
1	Constant	0.5644939507569261	0.6951305575158786
	HO	0.008609928926720812	
	HD	0.16115308088163308	
	AO	-0.03197390365516486	
	AD	-0.13813062261521405	
2	Constant	0.6143282059733154	0.6916376306620209
	HO	0.008951012656998934	
	HD	0.161762366586332	
	AO	-0.01757058369792213	
	AD	-0.15511888741122926	
3	Constant	1.6471532044325323	0.6837606837606838
	HO	0.012707671329993115	
	HD	0.13942809324633929	
	AO	-0.035022471166356726	
	AD	-0.13288059283458398	
4	Constant	1.6033905202907694	0.6800584368151936
	HO	0.016387156045137227	
	HD	0.13549027162848515	
	AO	-0.022426475101414173	
	AD	-0.14558209142573333	

HO is short for Home Offense, HD for Home Defense, AO for Away Offense, and AD for Away Defense.

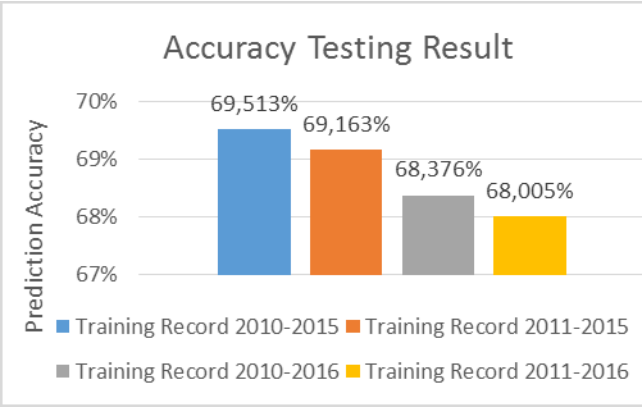

Figure 3. Accuracy Comparison

Highest accuracy reached is 69.513%, by using training data from 2010/2011 season up to 2014/2015, while the lowest was the 4th experiment, only 68.005% when season 2010/2011 match records were not included in building the model but the testing data included.

We were interested in the results produced, where there are 2 variables always showed far greater coefficients than the rest, which are HD and AD, thus we conducted another experiment by using only those 2 variables. Experiment results can be seen at Table IV.

Table IV. Experiment with HD and AD

Num.	Variables	Regression coefficients	Prediction accuracy
1	Constant	0.677906856015421	0.6901905434015526
	HD	0.16677038977902872	
	AD	-0.16906210558019102	
2	Constant	0.6471071562768064	0.6907665505226481
	HD	0.1694810059658868	
	AD	-0.1720845898793583	
3	Constant	1.7022087234163048	0.684981684981685
	HD	0.148579792196321	
	AD	-0.16549066485840724	
4	Constant	1.608989961108896	0.6778670562454346
	HD	0.14983574233944108	
	AD	-0.16616138214633389	

V. Result Analysis

Our analysis is there is so many unexpected result in 2015/2016 records and when incorporated in training data, the model built was twisted and does not produce the expected result. Our work has shown quite significant improvement compared to [9], with our predictive accuracy hitting 69.5% compared to his accuracy on 51%, and we succeed in building a model with less variables but stronger prediction accuracy.

As can be seen from Table III, the two biggest coefficients are always variable HD and variable AD. It is marked as the significant variables, the factors that determine whether a team lose or win. This is similar to other works done, strengthening the assumption on those variables being the thing that really matters in a match.

However, when we experimented using only HD and AD, the result is not as good as using 4 variables chosen on the first experiment. Both variables produced a very close number, with the 2 highest results coefficient differentiating only 0.003.

Looking at Figure 3, we can notice that the model built was not overfit, what was likely to happen in [4] where they used the same testing and training data and showed 95% prediction accuracy, because when we added test data into building model, prediction accuracy went down compared to when it is not included in training.

Another analysis to be made is on constant. It always produces such a high number on every experiment. It is due to imbalanced data in training records. It is recorded that there are much more home win data than away win data. This causes the coefficients estimation on constant tend to have bigger numbers.

VI. Conclusion

In this research, we have built a logistic regression model to predict 2015/2016 Barclays' Premier League matches outcomes. We conclude that the significant variables are Home Defense and Away Defense, but prediction cannot be done using only these two variables, as can be compared from Table III and Table IV. Also, choosing only the significant variables can increase prediction accuracy, as it can be seen from this research compared to Sneider, our prediction is 18% more accurate. Logistic regression is an easy to implement and easy to understand method to apply to prediction problem, moreover it gives us additional insight through the estimated coefficients, but it needs a set of variables that can improve performance of this work. We also concluded that addition of training instance does not affect prediction accuracy, under condition that the additional data does not have any unexpected result.

In the future works, we could develop this into further uses. A prediction result can be used to assists club's manager to form tactic or strategy on upcoming matches, or it could be developed into a system recommender on what player to buy to strengthen the significant variables in football matches.

Acknowledgment

We would like to thank Yani Widyani, S.T., M.T., and Dr. Drs. Judhi Santoso, M. Sc. for giving us correction and improvement suggestion on this work. We would also like to thank our friends and colleagues that gives us motivation to do the research.

References

[1] Bailey, M.J. (2005). *Predicting Sporting Outcomes: A Statistical Approach*. Swinburne University of Technology: Faculty of Life and Social Sciences.

[2] Baio, G., & Blangiardo, M. (2010). Bayesian Hierarchical Model for The Prediction of Football Results. *Journal of Applied Statistics, 253-264.*

[3] Hosmer, D.W. Lemeshow, S. & Sturdivant, R.X. Applied Logistic Regression 3rd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

[4] Igiri, C.P., & Nwachukwu, E.O. (2014). An Improved Prediction System for Football a Match Result. *IOSR Journal of Engineering Volume 04 Issue 12, pp12-20.*

[5] Min, B., et al. (2008). A Compound Framework for Sports Result Prediction: A Football Case Study. *Journal of Knowledge-Based System Volume 21 Issue 7*, 551-562. The Netherlands: Elsevier Science Publishers.

[6] Peng, et al. (2002). An Introduction to Logistic Regression Analysis and Reporting. Indiana University-Bloomington:EBSCO Publishing.

[7] Reddy, V., & Movva, Sai V. K. (2014). The Soccer Oracle: Predicting Soccer Game Outcomes Using SAS® Enterprise Miner™. SAS® GLOBAL FORUM. Washington, D.C.

[8] Shin, J., & Gasparyan, R. (2014). A Novel Way to Soccer Match Prediction. Stanford University:Department of Computer Science.

[9] Snyder, Jeffrey A.L. (2013). *What Actually Wins Soccer Matches: Prediction of the 2011-2012 Premier League for Fun and Profit*. Thesis, University of Washington, WA: Department of Computer Science.

[10] http://www.transfermarkt.co.uk/, accessed on 17th May 2016.