Strathmore
UNIVERSITY

A COMPARATIVE STUDY OF FOOTBALL PREDICTION MODELS

RAPANDO REGINALD TABUCHE 101792

Submitted in partial fulfillment of the requirements for the Degree of Bachelors of Business Science Actuarial Science at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University Nairobi,
Kenya

July, 2020

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

RAPANDO REGINALD TABUCHE ..... [*Name of Candidate*]

R.A. ..... [*Signature*]

17/02/2021 ..... [*Date*]

This Research Project has been submitted for examination with my approval as the Supervisor.

MELEAH OLECHE ..... [*Name of Supervisor*]

..... [*Signature*]

17/02/2021 ..... [*Date*]

Strathmore Institute of Mathematical Sciences

Strathmore University

Table of Contents

Table of figures

ABSTRACT

This study investigates the effect of tiers on football prediction models. The tiers in this study are grouped according to the value of the teams. Many football prediction models do not take tiers as a factor when making predictions. The study used a Poisson Regression Model which assumes that the number of goals scored by a team follow a Poisson distribution. It also employed a Bayesian approach. The parameters used in the Poisson Regression Model are strength of attack, defensive strength, home advantage and club value (the tiers). From the study it was found that teams in a higher tier were given a higher probability of scoring more goals and winning games as compared to teams in a lower tier. This shows that the value of a team has a significant effect on the predicted outcome in a football prediction model. From this, it was concluded that the value of a team is an important aspect that needs to be considered in football prediction models since it has a significant effect on the outcome of prediction.

1. INTRODUCTION

1.1. Background Information and Problem Statement

Football has become the most popular sport in the world estimating a 4 billion person following and a global sphere of influence. Due to this there has arisen a craze to predict the scores of various matches. This can be either as a super fan who just wants to see their team win (most times they are wrong in their predictions since they predict using emotions) or even as football analysts, who most times are retired professional footballers, who give their insight and predictions on the game before it starts or even bookmakers who create odds for gambling and the right prediction of the outcome of a game means more money for them. Prediction of football scores can be done in many ways. Some say they use their intuition (a gut feeling), mostly fans, others study teams and their performances over various games and make an informed guess, mostly gamblers, and others, which will be the main method analyzed in this study, use predictive models, mostly bookmakers.

Are football prediction models really accurate for all team levels: top, middle and bottom teams?

There have been various models developed over the years to predict football scores. Most of these models draw from various statistical concepts. For example, Bayesian networks which are a type of probabilistic graphical model that uses Bayesian inference for probability computations. They aim to model conditional dependence, and therefore causation, by representing conditional dependence by edges in a directed graph. Such a model that combines dynamic ratings and hybrid Bayesian networks to predict football match outcomes is the Dolores (Constantinou A. C., 2018). Another example is the Poisson distribution which is a statistical distribution that shows how many times an event is likely to occur within a specified period of time. It is used for independent events which occur at a constant rate within a given interval of time. Such a Poisson regression model was used to predict the football scores of the 2012/2013 English Premier League and the 2015 Brazilian Serie A (Saraiva et al. , 2016). Other models have even drawn from the biological neural networks in the form of artificial neural networks. They "learn" to carry out tasks by considering samples generally without being guided with specific task rules. A system from this has been used to predict results for the Iran pro league (Arabzad et al. , 2014).

According to (Constantinou & Fenton, 2010) although association football forecasting models are growing in popularity and importance, no method of evaluating their accuracy has been agreed upon. They grouped the evaluators used into two categories: those considering only the prediction for the observed outcome; and those considering predictions for both the unobserved and observed outcome. They highlighted underlying differences between them and demonstrated that they produced wildly contrasting conclusions about the exactness of four different predictive systems based on Premier league data.

In Robust fitting of football prediction models (Karlis & Ntzoufras, 2010), they proposed a new way to accurately predict results in football games. They claimed that existing models focused on modelling the numbers of goals scored by two competitors with parameter estimation of the assumed model usually based on the maximum likelihood approach. This to them was not sufficient enough and they proposed a weighted likelihood approach which allows the modeler to allocate less weight to a specific football score if there is a perception that the result was not typical and makes (in any way) the parameter estimates less accurate.

Although there is growth in football as a sport and an increase in various prediction models, there are no existing accepted evaluators in that domain (Constantinou & Fenton, 2010). From all the evaluators that they used they established underlying differences between the evaluators and demonstrated that they produce wildly dissimilar conclusions about the exactness of the four predictive systems.

There was a way proposed by (Karlis & Ntzoufras, 2010) on how to improve the accuracy of football models. This was a different approach as compared to the evaluators approach taken by (Constantinou & Fenton, 2010). However, in both cases the underlying issue was the accuracy of football models.

Various models have been used to improve on the accuracy of football prediction models. However, no known model has put team levels into consideration. Therefore, this paper seeks to fill the gap that team levels is an important aspect of football prediction.

1.2. Research Objective

The objective of this study is to investigate the effect of team level(tier) on football prediction modelling

1.3. Significance of the Research

There are many beneficiaries to this study. Some examples are sport betting companies who are setting their odds. If they can predict the matches at a higher accuracy they can make more profit. Also, for investors investing in football clubs, if they are able to predict the performance of a specific club in a coming season accurately, they would be able to predict the profitability of that club.

There is an expectation of two possible results. If the introduction of team value increases the accuracy of the model it shows that team value is an important variable in football prediction. However, if the introduction of the variable has no effect on accuracy of the football prediction then the variable has no effect on football prediction.

2. LITERATURE REVIEW

2.1. Introduction

There have been various works on the prediction of football scores and even the accuracy of those scores done by various scholars. This section separates the past research into two sections. The various models that have been proposed and studies that focus on the improvement of football prediction model accuracy.

2.2. The models

In the study by (Hvattum & Arntzen, 2010), they examined the value of assigning ratings to teams according to their past results so as to forecast match outcomes in association football. According to (Cengiz et al. , 2012) a Bayesian model was proposed to test its forecasting strength on the Turkish Super League in the 2008/2009 season. They did a comparative analysis of the accuracy of the Bayesian approach against a Poisson Log Linear and Artificial Neural Network approaches.

Also, (Chandran & Sujatha, 2013) proposed a prediction technique that used a neural network approach for prediction of football match outcomes. It analyzed patterns from a variety of factors which affected the outcome of a game making use of past data. They described the inputs and outputs and compared the results of such a system. According to (Owramipur et al. , 2013) a Bayesian Network (BN) was proposed to predict results of football matches. They studied the 2008/2009 season in La Liga and tested its performance.

In the work by (Arabzad et al. , 2014) they suggested a machine learning approach, that is, Artificial Neural Networks (ANNs), in predicting match outcomes for one week in the Iran Pro League for 2013/2014 season. For (Lid'en, 2016) he implemented and ran tests with historical data, different models predicting results of football games. Models using a bi-variate distribution for number of home goals and away goals were fitted and tested in practice. He also tested profitability against some bookmakers.

According to (Eggels et al. , 2016) a method to predict the winner in a match in elite soccer was proposed. The result of the match is determined by estimating the probability of scoring for the individual chances created by the team. The result of a match is then got by

integration of the probabilities. A Poisson regression model was also proposed by (Saraiva et al. , 2016) for prediction. They applied it to two national competitions: the 2012/2013 English Premier League and the 2015 Brazilian Football League.

For (Razali et al. , 2017) they looked at the prediction of match results in terms of home win, away win and draw by use of Bayesian Networks. They considered The English Premier League (EPL) for the 2010/2011,2011/2012 and 2012/2013 seasons. The study by (Pettersson & Nyquist, 2017) investigated the deep learning method Recurrent Neural Networks (RNNs) in football match prediction. Also, (Constantinou A. C., 2018) described a model for prediction of football match outcomes in one country by observing football matches in other multiple countries. The model was a combination of two methods: a) dynamic ratings and b) Hybrid Bayesian Networks.

In the study by (Ganesan & Harini, 2018) they predicted the match outcomes of the English Premier League, by studying past football matches and observing the most important attributes that are likely to decide the conclusion. They used algorithms such as Support Vector Machines and then selected the best one to give them the target label. For (Azhari et al. , 2018) they proposed a Poisson regression model to predict the results of football matches. They predicted the average goals scored by each team by assuming that the number of goals scored followed a univariate Poisson distribution. The model was formulated from four covariates: average goals scored in a match, the home advantage, the team's attacking strength, and the opposing team's defensive strength. The model was applied to the 2017/2018 English Premier League.

An Artificial Neural Network model was developed by (Balogun & Ogunseye, 2019) with Manchester United as a case study. Past results by Manchester United were used to train and validate the network. Also,(Rahman, 2019) proposed a comprehensive framework developed by deep neural networks (DNNs) and artificial neural network (ANNs) for football match prediction. A dataset is used with the rankings, team performances, all previous international football match results and so on. ANN and DNN are used to explore and process the sporting data to generate prediction value.

2.3. Accuracy

In the study by (Aslan & Inceoglu, 2007) they highlighted the efficiency of black-box modeling capacity of neural networks in football match prediction. They also opened up several debates on the prediction nature and input parameters selection. For (Karlis & Ntzoufras, 2010) they proposed a weighted likelihood approach allowing the modeler to allocate less weight to a specific football score if there is a perception that the result was not typical and makes (in any way) the parameter estimates less accurate.

According to (Constantinou & Fenton, 2010) although football prediction models were becoming really relevant, there was no agreed method of their accuracy evaluation. The evaluators were classified into two categories: those considering only the prediction for the observed outcome; and those considering the predictions for both unobserved and observed outcome. They highlight underlying differences between them and demonstrate that they produce wildly dissimilar conclusions about the exactness of four different predictive systems built on recent Premier league data. Also, (Constantinou & Fenton, 2017) looked at what they considered smart-data; a method supporting data and knowledge engineering approaches that emphasize applying causal knowledge and real-world factors in model development, taking into great consideration what data is most important for forecasting rather than what is available.

For (Hessels, 2018), he forecasted the outcome of games in the English Premier League, Championship, League One, and League Two through machine learning and data science. The objective this study was to improve upon the prediction results of Ulmer et al. (2013) by using a bigger dataset. According to (Rosli et al. , 2018) different mining techniques needed to be taken into consideration for match prediction. The main thoughts behind this research was finding the most appropriate data mining technique fitting the nature of football data.

In the study by (Eryarsoy & Delen, 2019) they built a model to forecast the results of football games and determine what factors influenced their outcomes. Also, (Wheatcroft, 2020) considered the use of observed and predicted match statistics as inputs to forecast the results of matches. It was seen that if it were possible to know the match statistics in advance, very revealing forecasts of the results could be made.

2.4. Summary of the literature review

From the above, most researchers have proposed models but none of them consider team value as an important variable in prediction. They have also insisted on the need for checking for accuracy in various models and researchers such as Constantinou, A. have highlighted the deficiency of various evaluators to check accuracy in models. They have put forth their work as a foundation for the building up of works on evaluation of accuracy in models. The study seeks to contribute to the gap by introducing a variable that is very important in all football leagues, the value of the club. Its importance is evidenced by the fact the resources available to a club determine the quality of players the club will have, the quality of manager and even the quality of staff supporting the club. Value of a club affects all other variables that are used in football prediction.

In most of the models described above, the main variables that have been considered are the attacking strength of the team and the defensive strength of the team. Also, home advantage (assumption that a team playing in its home stadium would be stronger) is a variable that is considered. These are then used to predict the goals scored by a team and the goals scored against a team. When you introduce the aspect of team value, it is clear to see that it affects all these other variables. A team with high value will be able to buy high quality strikers and hence increase the attacking strength as compared to a team with lesser value. Same case with the ability to buy high quality defenders and goalkeepers increasing the defensive strength of a team. This is even evidenced by the fact that UEFA introduced Financial Fair Play regulations where clubs are not allowed to spend more than they make in the pursuit of success showing that value is a big factor in club performance that is overlooked.

3. METHODOLOGY

3.1. Research Design

3.1.1.1. Choice of design and justification

The model was based on a replication of a Poisson Regression Model (Saraiva et al., 2016). It was applied on the 2012/2013 English Premier League and the 2015 Brazilian Football League. The alteration which is made is the inclusion of club value as a variable. The reason for choosing this model is the fact that it assumes the number of goals scored follows a Poisson distribution.

3.2. Population and Sampling

3.2.1.1. Definition of population and sample

The study tries to establish whether the value of a club affects the prediction accuracy of a football model. The target population was various football clubs in the world in any league. More specifically, the accessible population was the clubs in the Spanish league (Campeonato Nacional de Liga de Primera División, La Liga). The study grouped the league into 3 sections according to the value of the teams. There were teams in the upper, middle and lower value. From this the sample were 3 teams in each tier. For the upper value we had Real Madrid, FC Barcelona and Atlético Madrid. For the middle value we had Athletic Bilbao, Getafe and Celta De Vigo. Finally, for the lower value Eibar, RCD Mallorca and Leganes were taken into consideration.

3.2.1.2. Sampling technique and procedure

The sampling technique used was biased sampling, more specifically purposive sampling. The team values in La Liga were researched and ranked from first to last and from this they were divided into 3 groups those with high value, those with medium value and those with lower value.

3.2.1.3. Representativeness of sample

The reason for choosing 3 teams in each rank was that it increases the sample size and also the teams chosen are the most consistent in the league. The sample was also a good representation of the target population since in all the leagues the teams vary in value. A team with the highest value, middle value and least value can be got in every league.

3.3. Data Collection

3.3.1.1. Type of data

The data used was quantitative in nature. This was the value of the various clubs and the number of goals scored by the clubs.

3.3.1.2. Source of data

The main source of data was secondary data collection through the internet. The data was got from https://www.football-data.co.uk.

3.3.1.3. Data collection method and instruments

Since all of the data was got from the internet, the data collection method was secondary data collection. Since the nature of data was, data that has already been compiled by other individuals and sources, the most appropriate tool was web articles. This was adequate in giving the information required for the study.

3.4. Data Analysis

3.4.1.1. Analytic technique

The analytic tool chosen for this study was predictive analytics since there was prediction of future scores of the various clubs. This technique was appropriate because the study was predictive in nature and the results were compared against current results in real time.

3.4.1.2. The Model

In this case the model was based on the Spanish League, La Liga, hence the number of teams was 20. Assume the number of teams to be n. The number of games played by each team will be $2(n-1)$, since they play each team home and away. The total number of games in the league will be $N = n(n-1)$, since each match involves two teams. The games are played in two phases. For the first phase between two teams h and a, the match will occur in $h's$ home stadium and in phase two it will occur in $a's$ home stadium.

For a game g between two teams h and a, let X_{hg} and X_{ag} be random variables denoting the number of goals for both the home and away team, for $g = 1, \dots, N$. Assuming that,

$$X_{hg} \sim Poisson(\lambda_{hg}) \text{and } X_{ag} \sim Poisson(\lambda_{ag})$$

So as to link the number of goals for both teams h and a with their value as a club(v), strength of attack(t), defensive strength(d) and the home advantage(s), we consider,

$\lambda_{kg} = e^{U_{kg}\beta_k}$,

for $g = 1, \ldots, N$, where $k = h, a$, $U_{kg} = (1,1,1,1)$ if the game is played at the home stadium of k team and $U_{kg} = (1,1,1,0)$ otherwise, and $\beta_k = (\beta_{k0}, \beta_{kv}, \beta_{kt}, \beta_{k^c d}, \beta_{ks})'$ is the vector of parameters for the k team, where k^c represents the opposite team. Therefore, if $k = h$, then $k^c = a$. β_{kv} measures the value of the team, β_{kt} measures the attacking strength of the team, $\beta_{k^c d}$ measures the defensive strength of the opposing team and β_{ks} measures the home advantage of the teams which we assume to be similar for each of the teams. However, in this a team with a good defense will have a negative defense effect and a team with positive defense effect means an increase in the number of expected goals of the opponent.

Suppose that a game g is played in $(r+1)^{th}$ round of the league, $1 \leq r \leq N$. Assume that n_r are the number of games played by teams h and a before the $(r+1)^{th}$ round. Consider $x_k = (x_{k1}, \ldots, x_{kn_r})$ be the goals scored by k team in the n_r games, where, x_{km} is number of goals scored by the k team in the m^{th} game, for $k = h, a$ and $m = 1, \ldots, n_r$. Thus, the log-likelihood function for (β_h, β_a) is given by

$$l(\beta_h, \beta_a; x_h x_a) = \sum_{k \in \{h,a\}} \sum_{m=1}^{n_r} (-e^{U_{km}\beta_k} + x_{km} U_{km} \beta_k - \log(x_{km}!))$$

Some constraints have to be imposed on team-specific parameters to avoid nonidentifiability. There is use of a sum-to-zero constraint from (Karlis & Ntzoufras, 2003) and (Baio & Blangiardo, 2010) , that is,

$$\sum_{h=1}^{n+1} \beta_{ht} = 0, \sum_{h=1}^{n+1} \beta_{hd} = 0 \text{ and } \sum_{h=1}^{n+1} \beta_{hs} = 0$$

Implying that the sum of the strength of the attack, defense and home advantage of all n teams is equal to zero.

For development of the Bayesian approach, prior distributions for the parameters β_k, $k = h, a$ need to be specified. There is an assumption that priors are a priori independent, that is, $\pi(\beta_h, \beta_a) = \pi(\beta_h)\pi(\beta_a)$, in which, $\pi(\beta_k) = \pi(\beta_{k0})\pi(\beta_{kv})\pi(\beta_{kt})\pi(\beta_{kd})\pi(\beta_{ks})$ for $k = h, a$. Therefore, the following prior distributions are taken into consideration: $\beta_{k0} \sim N(0, 10^{-4})$, $\beta_{kv} \sim N(0, 10^{-3})$, $\beta_{kt} \sim N(0, 10^{-3})$, $\beta_{k^c d} \sim N(0, 10^{-3})$ and

$\beta_{ks} \sim N(0, 10^{-3})$, for $k = h, a$, where $N(0, b)$ denotes the normal distribution with mean 0 and precision b.

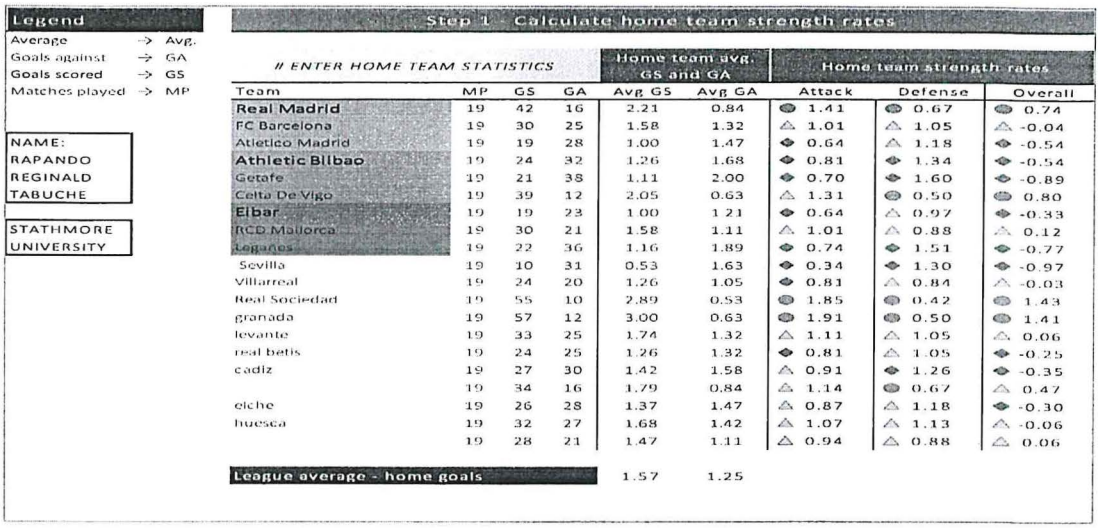
4. RESULTS AND FINDINGS

This chapter looks at the results obtained from the application of the methodology raised from the literature review chapter. The data used in this study was 2018/2019 data because this was the most recent uninterrupted season. The most recent season data should have been the 2019/2020 season but due to the effect of Corona on the season causing it not to be played for a while the 2018/2019 season was deemed more appropriate.

The outcome for the predictions is arrived from the goals scored, attacking strengths, defense strengths overall strength rates and the team value. The different outcomes also depend on whether the team is away or home. The home team strengths are calculated for the different teams in the different levels of the league which will be later be used for the determination of the outcome.

The games played depend on the point at which the league has reached. The home strengths and away strengths depend on the scoring during the games played. In the calculation of the Poisson distribution then there is need for the calculation of the goals that a team is expected to score and also those that are expected to be scored against the team. When measuring these values, it is crucial to use a relevant date range depending with the league and the type of game. Too long and the data would be meaningless due to the dynamic shifts in non-playing and playing personnel in modern football, whereas too short leaves the data vulnerable to outliers. For example, getting a big data range between 2000 and 2020 would give wrong predictions as much have changed in the teams' management, personnel and support base.

Figure 1: home team strength rates



Legend

Average -> Avg.
Goals against -> GA
Goals scored -> GS
Matches played -> MP

NAME:
RAPANDO
REGINALD
TABUCHE

STATHMORE
UNIVERSITY

Step 1 Calculate home team strength rates

# ENTER HOME TEAM STATISTICS				Home team avg. GS and GA		Home team strength rates		
Team	MP	GS	GA	Avg GS	Avg GA	Attack	Defense	Overall
Real Madrid	19	42	16	2.21	0.84	1.41	0.67	0.74
FC Barcelona	19	30	25	1.58	1.32	1.01	1.05	-0.04
Atletico Madrid	19	19	28	1.00	1.47	0.64	1.18	-0.54
Athletic Bilbao	19	24	32	1.26	1.68	0.81	1.34	-0.54
Getafe	19	21	38	1.11	2.00	0.70	1.60	-0.89
Celta De Vigo	19	39	12	2.05	0.63	1.31	0.50	0.80
Eibar	19	19	23	1.00	1.21	0.64	0.97	-0.33
RCD Mallorca	19	30	21	1.58	1.11	1.01	0.88	0.12
Leganes	19	22	36	1.16	1.89	0.74	1.51	-0.77
Sevilla	19	10	31	0.53	1.63	0.34	1.30	-0.97
Villarreal	19	24	20	1.26	1.05	0.81	0.84	-0.03
Real Sociedad	19	55	10	2.89	0.53	1.85	0.42	1.43
granada	19	57	12	3.00	0.63	1.91	0.50	1.41
levante	19	33	25	1.74	1.32	1.11	1.05	0.06
real betis	19	24	25	1.26	1.32	0.81	1.05	-0.25
cadiz	19	27	30	1.42	1.58	0.91	1.26	-0.35
	19	34	16	1.79	0.84	1.14	0.67	0.47
elche	19	26	28	1.37	1.47	0.87	1.18	-0.30
huesca	19	32	27	1.68	1.42	1.07	1.13	-0.06
	19	28	21	1.47	1.11	0.94	0.88	0.06
League average - home goals				1.57	1.25			

Generally, the away teams attack strength rates are lower than when they are at home, which also applies to the defense strengths. These are dynamics that are considered in the distribution before determining the outcome for a match. The overall strength rates are dependent on the attack and defense strengths. The teams that are used in this study are divided into three as discussed in the methodology chapter. In normal cases their strengths should reduce as one goes down in the table. The individual attacking and defense strengths may vary for different teams in the table. For example, the overall strength for Barcelona could be high but it doesn't mean that its defense strength will be higher than that of Eibar.

Figure 2 below also shows the value of the specific teams that were considered in this study. The data was obtained only from a secondary source and is indicative even for future determinations. The value of the top team doesn't necessarily have to be the highest considering that the team can have a good form for a given season. For this case the highest value team is the second on the table.

Figure 2; Away team strength rates

Step 2 - Calculate away team strength rates ► Sheet2 - Match predictions

# ENTER AWAY TEAM STATISTICS				Away team avg. GS and GA		Away team strength rates		
Team	MP	GS	GA	Avg GS	Avg GA	Attack	Defense	Overall
Real Madrid	19	31	35	1.63	1.84	1.30	1.17	0.13
FC Barcelona	19	26	45	1.37	2.37	1.09	1.51	-0.42
Atletico Madrid	19	16	32	0.84	1.68	0.67	1.07	-0.40
Athletic Bilbao	19	21	36	1.11	1.89	0.88	1.21	-0.33
Getafe	19	13	31	0.68	1.63	0.55	1.04	-0.49
Celta De Vigo	19	24	27	1.26	1.42	1.01	0.91	0.10
Eibar	19	32	30	1.68	1.58	1.34	1.01	0.34
RCD Mallorca	19	24	25	1.26	1.32	1.01	0.84	0.17
Leganes	19	12	45	0.63	2.37	0.50	1.51	-1.01
Sevilla	19	12	45	0.63	2.37	0.50	1.51	-1.01
Villarreal	19	27	28	1.42	1.47	1.13	0.94	0.19
Real Sociedad	19	34	12	1.79	0.63	1.43	0.40	1.03
granada	19	38	11	2.00	0.58	1.60	0.37	1.23
levante	19	32	29	1.68	1.53	1.34	0.97	0.37
real betis	19	18	23	0.95	1.21	0.76	0.77	-0.02
cadiz	19	18	35	0.95	1.84	0.76	1.17	-0.42
	19	33	23	1.74	1.21	1.39	0.77	0.61
elche	19	26	31	1.37	1.63	1.09	1.04	0.05
huesca	19	20	28	1.05	1.47	0.84	0.94	-0.10
	19	19	25	1.00	1.32	0.80	0.84	-0.04
League average - away goals				1.25	1.57			

CLUB VALUE
0.74
0.81
0.77
0.21
0.18
0.12
0.07
0.04
0.04

1. **First situation**

This situation considers two teams that are at the top of the league. As indicated the expected number of goals to be scored are determined. This is the third step in the determination of the outcome. The teams feed determines the data that will be looked up in the teams' average goals and strengths. At this point the team value is also looked up and helps in the prediction. A higher team value is likely to favor the team to score more and win considering that soccer is a business in most cases.

Step 3 - Pick match, and calculate expected goals (xG)

# ENTER THE MATCH YOU WANT TO PREDICT		Strength rates		Expected number of goals scored
	Team	Attack	Defense	xG
Home	Atletico Madrid	0.64	1.18	1.16
Away	FC Barcelona	1.09	1.51	1.30

UPPER VALUE TEAMS

Step 4 - Predict correct score market

Away team goals \ Home team goals	0	1	2	3	4	5	6	7	8	9	10
0	8.5%	9.9%	5.8%	2.2%	0.6%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%
1	11.1%	12.9%	7.5%	2.9%	0.8%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%
2	7.2%	8.4%	4.9%	1.9%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
3	3.1%	3.6%	2.1%	0.8%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
4	1.0%	1.2%	0.7%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
5	0.3%	0.3%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
6	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Figure 3: Poisson regression

The figure 3 above shows that the Poisson regression gives the percentage with ranges from 0 to 10. The more the green color the higher the percentage. For example, for this case a score of 1-1 is likely to happen for a game between Barcelona and Atletico Madrid. These are arrived at with consideration of the different strengths. The other outcomes can happen but they have less probability of happening so the model gives the best of all the outcomes. In this case there are 100 outcomes but only one will be arrived at.

Figure 4: Regression

Expected number of goals scored		Home team goals								
	#	0	1	2	3	4	5	6	7	8
xG	0	=POISSON.DIST(N$5,$F$24,FALSE)*POISSON.DIST($I23,F25,FALSE)								
1.31	1	9.9%	12.9%	8.5%	3.7%	1.2	POISSON.DIST(x, mean, cumulative)			
0.92	2	4.5%	6.0%	3.9%	1.7%	0.6%	0.1%	0.0%	0.0%	0.0%

Figure 5: Outcome display

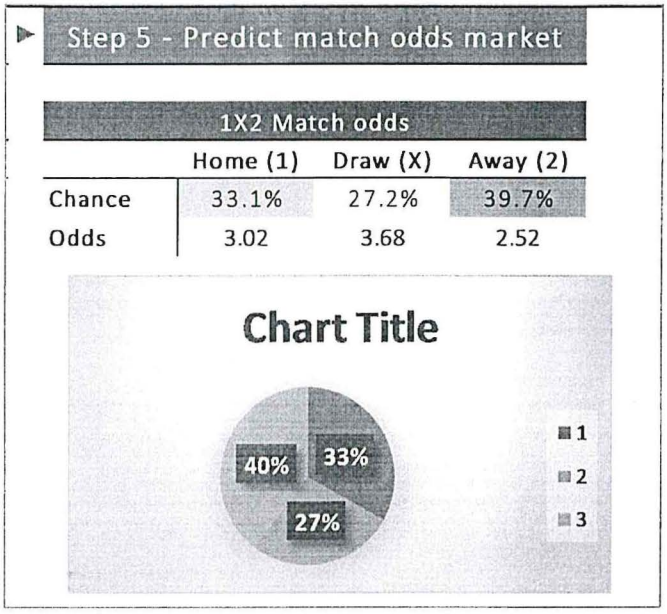
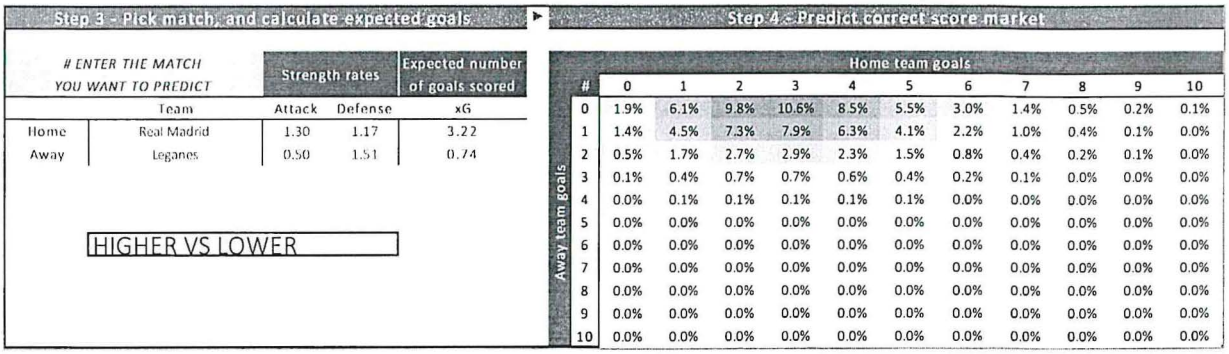


Figure 5 above shows how the outcomes are displayed. The chance indicates the outcome of the prediction which is likely to happen. A 33.1% show that Atletico Madrid is less likely to win as compared to Barcelona which has a percentage of 39.7%. The odds which are also a representation of the chance of winning are also arrived at from the distribution. The high club value also contributes to the outcome

2. **Second situation**

The second situation follows similar outcome prediction but for a high-level team vs low level team. Here there is a bigger difference in the strengths thus there will be a bigger margin in the distribution and percentages. The most likely outcome from the Poisson distribution is a score of 3-0 in favor of real Madrid. The expected number of goals to be scored varies between the clubs which is also as a result of their differences in their strengths.

Figure 6: Second situation



Step 3 - Pick match, and calculate expected goals

# ENTER THE MATCH YOU WANT TO PREDICT		Strength rates		Expected number of goals scored
	Team	Attack	Defense	xG
Home	Real Madrid	1.30	1.17	3.22
Away	Leganes	0.50	1.51	0.74

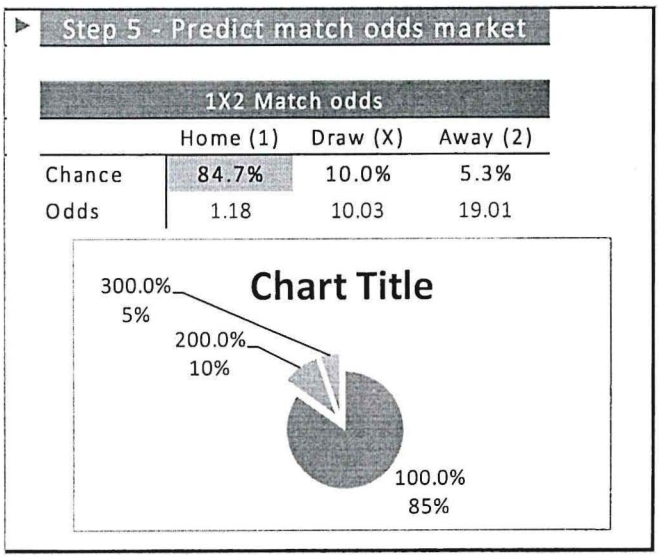
HIGHER VS LOWER

Step 4 - Predict correct score market

Away team goals #	Home team goals 0	1	2	3	4	5	6	7	8	9	10
0	1.9%	6.1%	9.8%	10.6%	8.5%	5.5%	3.0%	1.4%	0.5%	0.2%	0.1%
1	1.4%	4.5%	7.3%	7.9%	6.3%	4.1%	2.2%	1.0%	0.4%	0.1%	0.0%
2	0.5%	1.7%	2.7%	2.9%	2.3%	1.5%	0.8%	0.4%	0.2%	0.1%	0.0%
3	0.1%	0.4%	0.7%	0.7%	0.6%	0.4%	0.2%	0.1%	0.0%	0.0%	0.0%
4	0.0%	0.1%	0.1%	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
6	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

In the figure below the chart shows the outcomes after the prediction. Real Madrid winning is 85%. This arises from the high attack and defense strengths when real Madrid is at home.

Figure 7: Chart for second situation



The figure below indicates the club values which were considered in the model for the prediction. The higher the club value the more likely it will win the game.

Figure 8: Club values

	2021
Team	Value in billions (€)
Real Madrid	0.742
FC Barcelona	0.807
Atletico Madrid	0.769
Athletic Bilbao	0.2116
Getafe	0.1785
Celta De Vigo	0.1192
Eibar	0.0695
RCD Mallorca	0.04301
Leganes	0.03795

In most cases the club value is related to the club levels. Every team with its strengths and weaknesses has a role played in its performance by the management. The management gets different funding according to its level in the league. When the predictions were made without team values the percentages of outcomes happening was different as compared to when the values were incorporated. There are multivariable systems which affect the outcome that are not in control of the level or the value of the club which will ultimately affect the accuracy of the model. Generally, Poisson distribution provides great input when there is the variability in the outcomes. A Poisson distribution can be applied in systems where a large number of events are possible to occur yet where these occurrences are accepted as very rare to occur. In terms of football, what exemplifies this situation best is the example that the possibility of every shot thrown to the goal to be scored is very low.

Although this study was restricted to 9 teams, it doesn't mean that it is limited to 9 teams. More teams can be considered depending with league formations and levels. Models predicting the outcomes of a football match should take the contrasts between the variables of the two clubs to compete into account. The abundance of factors that can influence the outcomes of football matches causes different issues during the prediction. Therefore, it is more significant to appraise the outcome of the matches with inclusion of number of goals scored, goals conceded and team value/level.

5. CONCLUSIONS AND RECOMMENDATIONS

The main aim of this paper was to develop a better model that improves the accuracy of various football prediction models by introducing the aspect of team levels as an important variable. The levels of the teams and their values were looked at and incorporated in the prediction of the outcome. The goal of this paper was to study and develop a criterion that would predict the outcome of matches in the La Liga league on different wagers.

The playing history, attacking and defensive strengths were to be considered. The different aspects were to be incorporated with the value of the teams to go further in the prediction. Different teams at different levels were to be used for the prediction. A Poisson regression model was incorporated with the aspect of Club values to give higher level of predictions. The model can be applied in different leagues provided that the attacking strengths, defense, club values, goal scoring and other aspects are known.

The paper lacks 100% surety considering there as other factors that affect the outcome of a game before and during paly. Incorporation of these many aspects is hard and would lead to a complicated model.

Limitations

The model does not consider changes to squad, manager changes and weather.

It doesn't indicate how the game will happen but only considered with the result of the game.

Many Correlations are ignored in the determination of the outcome.

It's restricted to a given data range.

6. REFERENCES

Arabzad, S. M., Araghi, T., Sadi-Nezhad, S., & Ghofrani, N. (2014). Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. *International Journal of Applied Research on Industrial Engineering Vol. 1, No. 3*, 159-179.

Aslan, B. G., & Inceoglu, M. M. (2007). A comparative study on neural network based soccer result prediction.

Azhari, H. R., Widyaningsih, Y., & Lestari, D. (2018). Predicting final result of football match using poisson regression model. *Journal of Physics: Conference Series*.

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football result. *Journal of Applied Statistics, 37*, 253-264.

Balogun, O., & Ogunseye, A. (2019). Artificial neural network approach to football score prediction. *Global Journal of Artificial Intelligence*.

Cengiz, M. A., Murat, N., Koç, H., & Senel, T. (2012). A bayesian computation for the prediction of football match results using artificial neural networks. *International Journal of Scientific Knowledge Volume 1 Issue 2*.

Chandran, J. A., & Sujatha, K. (2013). Football match statistics prediction using artificial neural networks.

Constantinou, A. C. (2018). Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 1-27.

Constantinou, A., & Fenton, N. (2017). Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowledge-Based Systems, 124*, 93-104.

Constantinou, A., & Fenton, N. E. (2010). Evaluating the Predictive Accuracy of Association Football Forecasting Systems.

Eggels, H., Elk, R., & Pechenizkiy, M. (2016). Explaining soccer match outcomes with goal scoring opportunities predictive analytics.

Erlandson F. Saraiva, A. K. (2016).

Eryarsoy, E., & Delen, D. (2019). Predicting the outcome of a football game: A comparative analysis of single and ensemble analytics methods.

Ganesan, A., & Harini, M. (2018). English football prediction using machine learning classifiers. *International Journal of Pure and Applied Mathematics Volume 118 No. 22*, 533-536.

Hessels, J. (2018). Improving the prediction of soccer match results by means of machine learning.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting 26* , 460-470.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician),52*, 381-393.

Karlis, D., & Ntzoufras, I. (2010). Robust fitting of football prediction models. *IMA Journal of Management Mathematics 22*, 171–182.

Lid'en, J. (2016). Bivariate models to predict football results.

Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football result prediction with bayesian network in spanish league - Barcelona team. *International Journal of Computer Theory and Engineering, Vol. 5, No. 5.*

Pettersson, D., & Nyquist, R. (2017). Football match prediction using deep learning: recurrent neural network applications.

Rahman, A. (2019). A deep learning framework for football match prediction.

Razali, N., Mustapha, A., Yatim, F. A., & Ruhaya, A. A. (2017). Predicting football matches results using bayesian networks for English Premier League.

Rosli, C. C., Saringat, M. Z., Razali, N., & Mustapha, A. (2018). A comparative study of data mining techniques on football match prediction. *Journal of Physics: Conference Series.*

Saraiva, E. F., Suzuki, A. K., Filho, C. A., & Louzada, F. (2016). Predicting football scores via Poisson regression model: applications to the National Football League. *Communications for Statistical Applications and Methods Vol. 23, No. 4*, 297–319.

Wheatcroft, E. (2020). Forecasting football matches by predicting match statistics.