International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-8, Issue-6, November 2021
https://doi.org/10.55524/ijirem.2021.8.6.7
www.ijirem.org

Survey on Football League Table and Player Performance Prediction Using Data Science

Swapneel Deshpande
Department of Computer Engineering, NBN Sinhgad School of Engineering, Pune, India,
swapd912@gmail.com

Varsha Rasal
Department of Computer Engineering, NBN Sinhgad School of Engineering, Pune, India,
varsharasal.nbnssoe@sinhgad.edu

ABSTRACT
This article focuses on team performance as well as player performance prediction, with team performance being evaluated using a variety of machine learning algorithms and web scraping methodologies. Data is refined and modified efficiently to get the desired accurate results. Advanced Statistics is used to get results. The prediction includes final league table of teams, whether a team is going to have a better season than the previous one. Prediction is also done to evaluate the rating of a defender.

Keywords
Sports analytics, Data mining, Web scrapping, Machine Learning

1. INTRODUCTION
Due to the ever changing and demanding nature of sports, the use of data analytics is on a great rise. Data analytics is the study and prediction of data and thus giving a favorable outcome from the analysis. Sports analytics is the study of data of various fields in sports and analyzing every aspect of it and forecasting its outcome.

Performance prediction is a very common task in today's world of sports analysis. Various clubs hire analysts pre and post season to introspect the performance of the team. Due to the high availability of data from many sources, analysis is done from different aspects of the sport. High amount of GPS data and in-match data makes the research easy.

Clubs collect and analyse data supplied by players using sophisticated gadgets and software (such as GPS tracking systems).

The proliferation of online football data is a benefit, but it requires filtering and proper data to anticipate team and player success. Regrettably, this is not always simple. Additionally, situations not depicted in the data collected might alter team and player performance. When a player returns to action after a significant injury, he or she may have a poor rating performance. Finally, due to the confusing character of football, statistical recording of match occurrences, as well as individual and team ratings, is a difficult task. Performance prediction is difficult, and long–term performance prediction is even more difficult, yet it has not been thoroughly investigated until now.

The usage of sports analytics has risen in recent years for all of he reasons stated above. There are unique challenges that make long–term football prediction tough.

2. OVERALL HISTORY
Instead of attempting to forecast goals scored or points won, academics have concentrated their efforts in recent years on explicitly predicting victories, draws, and defeats. Lago-Penas et al. found that shots on goal, crosses, match location, ball possession, and opponent team ability were the most discriminating factors, based on a ranking system.

Studying the Dutch Football Championship, Tax and Joustra suggested that combining public data and betting odd in a hybrid model might improve the accuracy. [1].

Hucaljuk and Rakipovic proved that Neural Networks were the best over any other technique[2].

Eggels et al. employed xG(Expected Goals) in 2016 to try to create a model that could classify each scoring opportunity into a scoring likelihood. They took advantage of geospatial data and used a variety of classification approaches. They also suggested that xG may be used to evaluate players and seasons, but cautioned that goal scoring opportunity probability estimates could have a high standard deviation [3].

Oberstone created a multiple regression model, resulting in six independent factors that he deemed enough for forecasting the EPL final league table in terms of points rather than exact positions [4]. He was able to obtain excellent outcomes.

There have been some noteworthy studies that have concentrated on football clubs' financial aspects rather than their on-field performance. Kringstad and Olsen surveyed about the financial capacity of the clubs and if it affects the team performance by studying Norwegian League[5]. After a lot of research, they came to the conclusion that focusing on athletics is still important because money is a big element in success, but only to a degree.

Sîrb et al. established a set of 54 performance criteria for various playing positions in order to evaluate player performance, taking into account each player's natural position as well as the tactical configuration used by the team during a match [6].

Finally, Pappalardo et al. examined player performance over a period of several years in 18 different competitions and presented PlayeRank, a data–driven framework. 31 million matches and 21,000 players were taken into account. Competitive prediction algorithms were proven to outperform PlayeRank. They also spoke about what sets top players apart from the rest and discovered patterns for great performances. One of PlayeRank's flaws is that it doesn't take into account off–ball activities like pressing. The authors also stated that a future version of the system should be able to incorporate data from other sources such as wearables, GPS, and video monitoring data [7].

3. RESEARCH METHODOLOGY

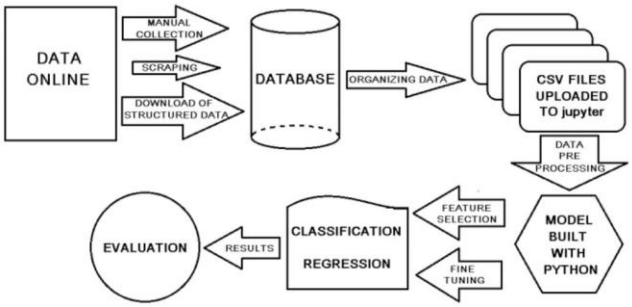

Figure 1: Block diagram of the process followed for the experiments

There are numerous websites that provide data and statistics about football matches and events. Some of them, on the other hand, were scraped off the internet using a variety of scraping programs. A free database was acquired. The database was collected from a popular manager simulation game and contains data from thousands of participants. It shows player ratings for a variety of football talents. Domain specialists rate the players.

Following the data collection procedure, there was a vast database that needed to be organised. The csv files were then uploaded to jupyter, which was mostly utilised for data processing.

Naturally, the data had to be preprocessed first. They were examined for null values, duplication, and noise, among other things. The data was cleaned and the models were built using Python.

4. EXPERIMENTS AND RESULTS

4.1 1st experiment: Team Performance Prediction

The first experiment is divided into two sections.

1st section: With a dataset containing over 40 variables for each team from four major European football national leagues for each of the last four years (2015-18), forecast whether a team will have a better or worse season in terms of points than the previous year. Every prior season serves as a training set, while the current season (2017–18) serves as a test set. It's treated as a binary classification issue, with AccuracyP being measured in the following way:

$$AccuracyP = \frac{Number\ of\ teams\ with\ correct\ perfromance\ prediction}{Number\ of\ total\ teams} \quad (2)$$

Further, a model was developed, using nearly the same features as the first part, to simulate every match of the 2018–19 season for the same championships (i.e. 380 matches per championship).Virtual points of every team are then used to determine final league table.

RMSE is calculated:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y_i} - y_i\right)^2} \quad (3)$$

where:

- n is the number of teams participating in the championship.
- $\hat{y_i}$ is the predicted points for the i-th team.
- y_i is the actual points for the i-th team.

The evaluation metric is *AccuracyM*, defined as follows:

$$AccuracyM = \frac{Number\ of\ games\ with\ correctly\ predicted\ outcome}{Number\ of\ total\ games} \quad (4)$$

The traits were linked to team success and were divided into three categories:

- Historical data from the previous five years. This mostly refers to past season's performance metrics (e.g. team average points).
- Team statistics from the most recent season.
- Information that isn't measurable in terms of team performance (e.g. financial).

These characteristics are formed during the summer break, therefore the majority of them are unrelated to the previous season's success, but they are quite likely to influence the next season's performance.

It is binary and indicates whether the team will have a better or worse season in terms of points won than the previous one.

Some features were eliminated from the original datasets after data preprocessing because they were irrelevant to the research or noisy, and so added little value to the outcome. Cards, interceptions, offsides, fouls, and other team statistics were included.

It achieved a very high accuracy and the results were satisfying.

In the second experiment, the data was derived from the results of every match in the four championships stated in the experiment's first part. Every unnecessary attribute was removed once again, and the datasets were combined with the datasets from the first half of the experiment. This approach generated a new dataset including every match from a football season, as well as statistical, financial, and historical data on the two teams involved in each match. The attributes of the home and away teams were then blended by removing the appropriate pairings.

Some qualities from the first part of the issue were removed from the second since the subtraction was useless. Finally, the first three seasons of each national title were combined, and dummy variables were used to encode each team.

2 sets were created from the original dataset: training and validation. The aim (i.e. the full-time outcome) was disguised and utilised as a test set for the previous season, which was kept separate from the others. There were 1140 rows in the training set with 40 different qualities.

To anticipate the match outcome and, as a result, the leagues' final rankings, multiple classifiers were used. The most essential features were team market value percentage, expected points, and non–penalties xG, but not by a large amount over other attributes.

They are comparable to the findings of prior studies, with the added benefit that the trials can be completed at the start of the season, with no official matches played and recorded. The English Premier League had the best AccuracyM for match results, with 57 percent, and the Spanish La Liga had the shortest RMSE for team points, with 9. In terms of AccuracyM and RMSE, the French Ligue 1 had the worst results. Tables 1 through 4 show the results from each league. Best outcome is highlighted in green, worst in red.

Table 1: Results from the English Premier League

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	55	17
Decision Tree	45	12.9
Random Forest	56	14.3
KNN	48	15.3
SVM (rbf kernel)	54	18.2
SVM (poly kernel)	57	11
XGBoost	52	17.3

Table 2: Results from the Spanish La Liga.

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	47	23.7
Decision Tree	39	14.9
Random Forest	48	17.7
KNN	46	13.3
SVM (rbf kernel)	51	13.8
SVM (poly kernel)	47	9
XGBoost	45	17.4

Table 3: Results from the Italian Serie A.

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	53	19.7
Decision Tree	41	11.3
Random Forest	40	14.4
KNN	47	14.7
SVM (rbf kernel)	52	19
SVM (poly kernel)	50	12.2
XGBoost	42	14.5

Table 3: Results from the French Legue 1.

CLASSIFIER	AccuracyM	RMSE
Naïve Bayes	42	28.2
Decision Tree	39	17.6
Random Forest	45	24.8
KNN	39	20.9
SVM (rbf kernel)	43	22.2
SVM (poly kernel)	43	17.3
XGBoost	44	21.8

SVM with polynomial kernel is clearly found to consistently achieve good results in every league tested, hence it will now be used as a benchmark for this research. The results from Spanish La Liga were best overall.

ACTUAL TABLE		PREDICTED TABLE	
1. Barcelona	87	1. Barcelona	83
2. Atletico Madrid	76	2. Atletico Madrid	75
3. Real Madrid	68	3. Real Madrid	64
4. Valencia	61	4. Valencia	58
5. Sevilla	59	5. Sevilla	57
6. Getafe	59	6. Getafe	57
7. Espanyol	53	7. Real Betis	57
8. Athletic Bilbao	53	8. Eibar	57
9. Real Sociedad	50	9. Celta Vigo	57
10. Real Betis	50	10. Villarreal	55
11. Alaves	50	11. Athletic Bilbao	54
12. Eibar	47	12. Real Sociedad	54
13. Leganes	45	13. Leganes	54
14. Villarreal	44	14. Espanyol	51
15. Levante	44	15. Alaves	51
16. Celta Vigo	41	16. Levante	51
17. Valladolid	41	17. Valladolid	51
18. Girona	37	18. Girona	51
19. Huesca	33	19. Vallecano	50
20. Vallecano	32	20. Huesca	49

Figure 2: Spanish La Liga 2018-19 actual vs predicted table

Green was used for the teams that won Champions league qualifications, blue that won Europa League qualifiers and red that were demoted at the end of the season.
As indicated in Table 5, the equivalent accuracy was quite good for teams that won European qualification, particularly those that qualified for the Champions League. The demoted teams' results were likewise satisfactory. Teams in the Europa League were the exception, with poor prediction accuracy.

Table 2: Accuracy in predicting champion, teams that won European qualification and teams relegated.

	Premier League	La Liga	Serie A	Ligue 1	Overall
Championship Winner	71%	71%	57%	57%	64%
European Qualification	86%	76%	82%	46%	75%
Champions League	79%	86%	71%	57%	74%
Europa League	38%	29%	29%	0%	29%
League Relegation	52%	48%	57%	10%	42%

Here, first 10 matches of 2018-19 season were used as training set and the remaining 28 days as the test set. As shown in Fig. 3, AccuracyM of the Spanish La Liga increased from 51% to 70% in that situation. As a result, it is demonstrated that using data from the current season can significantly improve the model's accuracy.

	precision	recall	f1-score	support
-1	0.72	0.38	0.50	76
0	0.59	0.78	0.67	79
1	0.83	0.90	0.86	125
micro avg	0.72	0.72	0.73	280
macro avg	0.72	0.69	0.68	280
weighted avg	0.73	0.72	0.71	280

```
[[ 29  35  12]
 [  6  62  11]
 [  5   8 112]]
Accuracy: 0.725
Mean is 0.7043939393939395
Standard deviation is 0.11155148041316686
```

Figure 3: Accuracy in predicting match results after 10 match days from the Spanish La Liga have been analyzed

4.2 2nd experiment: Player Performance Prediction

Here defenders, especially central defenders are considered. Because goals are regarded as the most essential aspect of football, defenders' contributions to a team are frequently overlooked. Midfielders and attackers are often rated higly.As a result, there is a scarcity of studies on central defenders. Furthermore, while it is simple to rank offensive players based on goals, important passes, and assists, determining what constitutes a strong centre defender is more difficult. The problem was first approached by normalising every numeric value in the dataset, transforming each attribute's range to 0 to 1. Despite the simplicity of this approach, some early inferences about which characteristics contribute more to central defender ability were obtained. Interceptions appear to be the most important attribute for the dataset under consideration, followed by team overall rating, as expected. Jumping reach, versatility, acceleration, and first touch on the ball were found to be the best attributes of the players.

Data set was split into 3 categories :

- Player characteristics and attributes.
- Player statistics.
- Team statistics.

Backward element strategy was used. The final model was constructed with seven features for the first category, which appear to be the most influential for a centre defender.

Here 5 regression models were verified. There was evidence of linearity in the model. Furthermore, residuals' expectation (mean) was nearly zero, implying that there was no (perfect) multicollinearity between features. The final assumption, however, was not confirmed; the Durbin–Watson test yielded a value much lower than 2, implying positive autocorrelation between features. Furthermore, the R–squared and adjusted R–squared were both low (under 0.5). The results, however, may be classified as positive, given that the dependent and independent variables came from two separate sources. After Backward Elimination, the final model consisted of 12 features with very low P–values, while R–squared was 0.867 and adjusted R–squared was 0.833, a huge improvement over the original model.

Out[191]:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.833
Method:	Least Squares	F-statistic:	25.03
Date:	Mon, 11 Nov 2019	Prob (F-statistic):	3.39e-16
Time:	02:24:20	Log-Likelihood:	64.550
No. Observations:	59	AIC:	-103.1
Df Residuals:	46	BIC:	-76.09
Df Model:	12		
Covariance Type:	nonrobust		

Omnibus:	1.400	Durbin-Watson:	1.912
Prob(Omnibus):	0.496	Jarque-Bera (JB):	1.161
Skew:	-0.136	Prob(JB):	0.560
Kurtosis:	2.369	Cond. No.	1.41e+05

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.9602	0.142	49.122	0.000	6.675	7.245
x1	-0.0327	0.004	-7.776	0.000	-0.041	-0.024
x2	0.0032	0.001	5.650	0.000	0.002	0.004
x3	-0.0113	0.002	-4.635	0.000	-0.016	-0.006
x4	6.782e-05	2.04e-05	3.317	0.002	2.67e-05	0.000
x5	0.2326	0.039	6.017	0.000	0.155	0.310
x6	0.0972	0.028	3.525	0.001	0.042	0.153
x7	0.1261	0.021	6.127	0.000	0.085	0.168
x8	-0.1559	0.042	-3.749	0.000	-0.240	-0.072
x9	-0.0481	0.019	-2.588	0.013	-0.086	-0.011
x10	0.7259	0.124	5.862	0.000	0.477	0.975
x11	0.9718	0.226	4.292	0.000	0.516	1.428
x12	2.1514	0.743	2.896	0.006	0.656	3.647

Figure 4: Model built with player statistics as independent variables.

The Breusch–Pagan test yielded a p–value of 0.44, indicating that there was no heteroscedasticity, while the Durbin–Watson test yielded a value of 1.91, indicating that the features had nearly little autocorrelation.

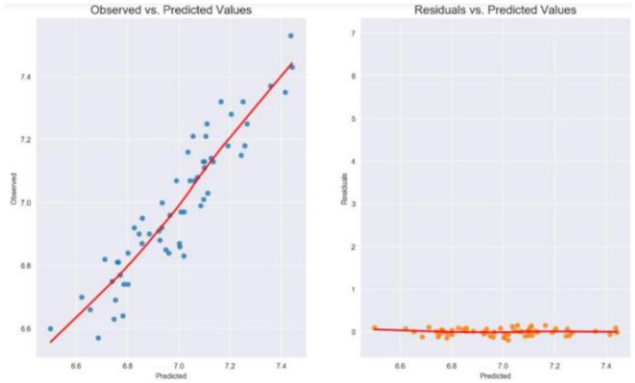

Figure 5: Linearity of the second model.

The third set of features (team statistics) did not contribute to the development of a satisfactory model. There was a hint that "TeamRating" was the only one of the factors worth considering. To take advantage of this feature, it was decided to include "TeamRating" in the second model. R–squared has increased to 0.907, and adjusted R–squared has increased to 0.88.

It's worth noting that, in keeping with how modern defenders are expected to play, attacking skills are included on the list:

Interceptions, Clearances, Won Aerials, Tackles, Jumping Reach, Versatility, Acceleration, First Touch on Ball, Age, Passing, Vision, Determination, Strength, Professionalism, and the ability to perform well in important matches

Caps, Minutes Played, Fouls, Inaccurate Short Passes, Key Passes, Goals, and Team Rating.

5. CONCLUSION

Two basic cases of sports analytics were investigated in this study: team performance prediction and player performance prediction. Tax and Joustra predicted match outcomes with an accuracy of 56 percent, whereas McCabe and Trevathan predicted match outcomes with an accuracy of 54.6 percent. Our findings have the benefit of being achieved without the use of any current official match data. The results were noteworthy, since thirteen features were found to be statistically significant, with a pleasing 0.907 R–squared and 0.88 adjusted R–squared. It was also discovered that some attacking talents, like as passing, and various attacking match behaviours (i.e. important passes made, goals scored) had an impact on centre defence ratings. This fact

emphasises the modern centre defender's changing approach to the game.

6. ACKNOWLEDGEMENTS

My first and biggest attitude goes to my mentor and adviser, Prof. Varsha Rasal. During the long journey of this study, he/she supported me in every aspect. She was the one who helped and motivated me to propose re-search in this field and inspired me with her enthusiasm on research.
Prof. Varsha Rasal, my guide, deserves a heartfelt thank you, for her excellent and valuable advice.
Special mention to the department leader, Prof Shailesh P. Bendale, principal and management inspiring me and providing all the lab and other facilities which made this seminar presentation very convenient.
I'm really thankful to all those who rendered their valuable help for successful completion on seminar presentation.

REFERENCES

[1]Tax, N. and Joustra, Y. Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. 10, 2015, Transactions of knowledge and data engineering, Vol. 10.

[2] Hucaljuk, J. and Rakipovic, A. Predicting football scores using machine learning techniques. 2011. MIPRO, 2011 Proc. 34th Int'l Convention

[3]Eggels, H. - van Elk, R. - Pechenizkiy, M. Explaining soccer match outcomes with goal scoring opportunities predictive analytics, 2016.

[4]Oberstone, J. Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success. 2009, Journal of Quantitative Analysis in Sports, Vol. 5.

[5]Kringstad, M. and Olsen, T.-E. Can sporting success in Norwegian football be predicted from budgeted revenues?

[6]Sîrb, L. - Molcuţ A. - Nastor, F. The Exercise of Prediction Process of Performance within Football Sports Management by Using Fuzzy Logic from the Perspective of Value Analysis on Tactical Compartments of Game of the Football Players. 2015, Journal of Knowledge Management, Economics and Information Technology, Vol. 5

[7]Pappalardo, L. - Cintia, P. - Ferragina, P. -Massucco, E. - Pedreschi, D. - Giannotti, F. PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach, ACM Transactions on Intelligent Systems and Technology September 2019 Article No.: 59.

