How reliable is Pythagorean Expectation to predict the result of the 2023/24 EFL Soccer Championship?

Introduction

As the impact of data analytics in sports, soccer in specific, continues to increase at an exponential rate, predictive models are constantly evolving to provide insights into team performance and outcomes. The focus of this study is to apply a well-grounded analysis of the data in hand using a method called the Pythagorean Expectation, borrowed from a fellow sport, baseball. The model aims to predict the result of the 2023/24 EFL Championship after verifying the reliability of the process using data from previous seasons.

Pythagorean Expectation Explained

Pythagorean Expectation is the formula that is used to estimate the number of games a team should win based on the number of runs they've scored and the number of runs they've conceded. The formula was developed by renowned baseball statistician Bill James, one of the pioneers of sports analytics.

$$(Runs\ Scored^2)\ /\ (Runs\ Scored^2\ +\ Runs\ Allowed^2)$$

Applying it to soccer will be relatively different since it's a low-scoring game and has a lesser number of games per season, thus forming a comparatively smaller dataset. Below is the formula for Pythagorean Expectation modified for soccer:

$$(Goals\ Scored^2)\ /\ (Goals\ Scored^2\ +\ Goals\ Conceded^2)$$

A bit about the EFL Soccer Championship

The EFL Championship is England's second-tier soccer league, featuring 23 teams competing in a 46-match season spanning over a year. It is renowned for its tight races,

competitiveness and dramatic finishes (“Here’s Hogg… DEENEY!”). The top two finishers of the league are directly promoted to the Premier League, England’s first-tier soccer league, while there’s a playoff for the third spot between the following four teams. The bottom three teams are relegated to League One, England’s third-tier soccer league.

Data Collection and Cleaning

The dataset consisted of the matches played by each of the 23 teams from each of the past five EFL Championship seasons, that is, from 2019/20 to 2023/24. The last entry of the matches played in the 2023/24 season was on January 22, 2024 which makes it 28 games played per team. The cleaning process involved filtering out required attributes for analysis, modifying filtered attributes and adding necessary attributes.

Finally, before beginning the analysis process, the dataset was split based on the season the set of games belonged to, based on the game date. An example of the state of the data at this stage is shown below from season 2019/20:

	Date	Team	Opponent Team	H/A	Goals	Goals Conceded	Possession	Opponent Possession	Duration	PTs	W	D	L	Goal Difference	Matches
0	20200722	Barnsley	Brentford	A	2	1	36.61	63.39	101.0	3	1	0	0	1	1
1	20200719	Barnsley	Nottingham Forest	H	1	0	61.89	38.11	98.0	3	1	0	0	1	1
2	20200716	Barnsley	Leeds United	A	0	1	48.77	51.23	98.0	0	0	0	1	-1	1
3	20200711	Barnsley	Wigan Athletic	H	0	0	50.97	49.03	100.0	1	0	1	0	0	1
4	20200707	Barnsley	Luton Town	A	1	1	63.91	36.09	100.0	1	0	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1109	20190824	Wigan Athletic	Queens Park Rangers	A	1	3	42.47	57.53	98.0	0	0	0	1	-2	1
1110	20190820	Wigan Athletic	Middlesbrough	A	0	1	47.78	52.22	99.0	0	0	0	1	-1	1
1111	20190817	Wigan Athletic	Leeds United	H	0	2	25.69	74.31	95.0	0	0	0	1	-2	1
1112	20190810	Wigan Athletic	Preston North End	A	0	3	44.32	55.68	96.0	0	0	0	1	-3	1
1113	20190803	Wigan Athletic	Cardiff City	H	3	2	56.98	43.02	97.0	3	1	0	0	1	1

Data Transformation and Analysis

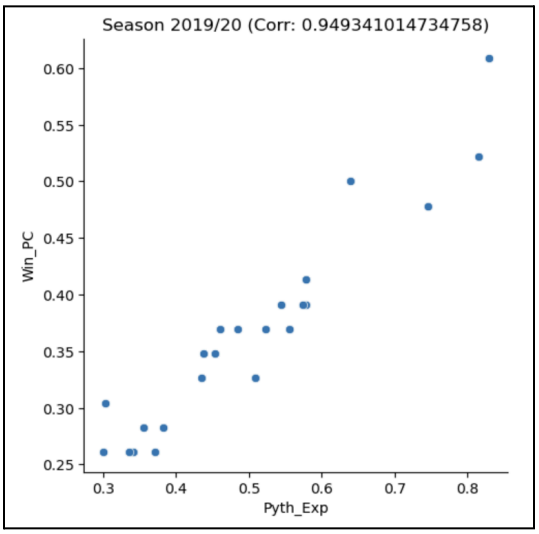
This step involved transforming the match-by-match data for each season into a table format for better readability and interpretation. It also led to finding some unseen

errors in the analysis process, which had to be accounted for before moving on to further data mungling. An example of the 2020/21 season's match-by-match data transformed to a table-format is shown below:

	Team	MP	W	D	L	GF	GA	GD	PTs	Team_Actual	PTs_Actual
0	Norwich City	46	29	10	7	75	36	39	97	Norwich City	97
1	Watford	46	27	10	9	63	30	33	91	Watford	91
2	Brentford	46	24	15	7	79	42	37	87	Brentford	87
3	Swansea City	46	23	11	12	56	39	17	80	Swansea City	80
4	Barnsley	46	23	9	14	58	50	8	78	Barnsley	78
5	Bournemouth	46	22	11	13	73	46	27	77	Bournemouth	77
6	Reading	46	19	13	14	62	54	8	70	Reading	70
7	Cardiff City	46	18	14	14	66	49	17	68	Cardiff City	68
8	Queens Park Rangers	46	19	11	16	57	55	2	68	Queens Park Rangers	68
9	Middlesbrough	46	18	10	18	55	53	2	64	Middlesbrough	64
10	Millwall	46	15	17	14	47	52	-5	62	Millwall	62
11	Luton Town	46	17	11	18	41	52	-11	62	Luton Town	62
12	Preston North End	46	18	7	21	49	56	-7	61	Preston North End	61
13	Stoke City	46	15	15	16	50	52	-2	60	Stoke City	60
14	Blackburn Rovers	46	15	12	19	65	54	11	57	Blackburn Rovers	57
15	Coventry City	46	14	13	19	49	61	-12	55	Coventry City	55
16	Nottingham Forest	46	12	16	18	37	45	-8	52	Nottingham Forest	52
17	Birmingham City	46	13	13	20	37	61	-24	52	Birmingham City	52
18	Bristol City	46	15	6	25	46	68	-22	51	Bristol City	51
19	Huddersfield Town	46	12	13	21	50	71	-21	49	Huddersfield Town	49
20	Derby County	46	11	11	24	36	58	-22	44	Sheffield Wednesday	47
21	Wycombe Wanderers	46	11	10	25	39	69	-30	43	Derby County	44
22	Rotherham United	46	11	9	26	44	60	-16	42	Wycombe Wanderers	43
23	Sheffield Wednesday	46	12	11	23	40	61	-21	41	Rotherham United	42

Proof of reliability of Pythagorean Expectation from the past four seasons (2019/20 to 2022/23)

The reliability of Pythagorean Expectation as a predictor of the result of the 2023/24 season can be seen in the regression and correlation analysis results from the previous seasons. The graphs and regression results for those seasons are shown below:



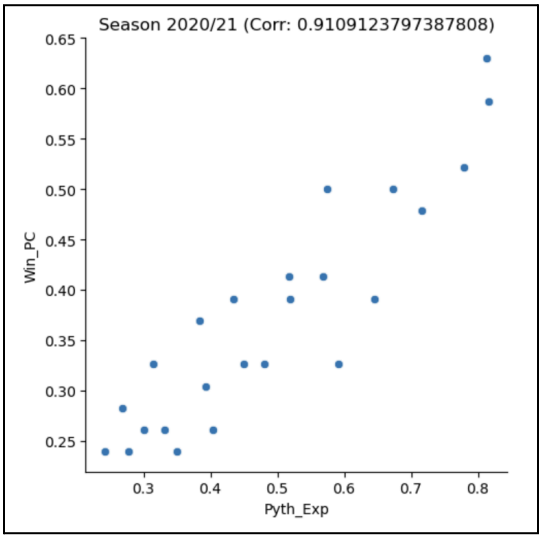
Season 2019/20

OLS Regression Results

Dep. Variable:	Win_PC	R-squared:	0.901
Model:	OLS	Adj. R-squared:	0.897
Method:	Least Squares	F-statistic:	200.8
Date:	Wed, 20 Mar 2024	Prob (F-statistic):	1.54e-12
Time:	21:55:13	Log-Likelihood:	52.097
No. Observations:	24	AIC:	-100.2
Df Residuals:	22	BIC:	-97.84
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0762	0.021	3.590	0.002	0.032	0.120
Pyth_Exp	0.5748	0.041	14.170	0.000	0.491	0.659

Omnibus:	2.074	Durbin-Watson:	2.034
Prob(Omnibus):	0.354	Jarque-Bera (JB):	1.604
Skew:	0.619	Prob(JB):	0.448
Kurtosis:	2.728	Cond. No.	8.66



Season 2020/21

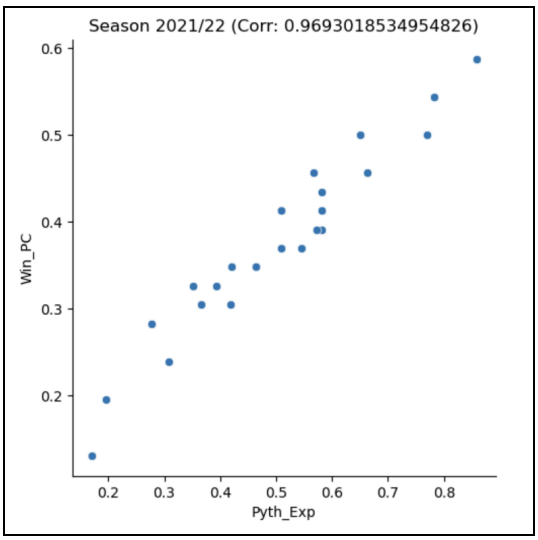
OLS Regression Results

Dep. Variable:	Win_PC	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.822
Method:	Least Squares	F-statistic:	107.2
Date:	Wed, 20 Mar 2024	Prob (F-statistic):	6.37e-10
Time:	21:55:20	Log-Likelihood:	39.983
No. Observations:	24	AIC:	-75.97
Df Residuals:	22	BIC:	-73.61
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0880	0.029	3.003	0.007	0.027	0.149
Pyth_Exp	0.5800	0.056	10.355	0.000	0.464	0.696

Omnibus:	0.282	Durbin-Watson:	1.761
Prob(Omnibus):	0.869	Jarque-Bera (JB):	0.385
Skew:	-0.219	Prob(JB):	0.825
Kurtosis:	2.562	Cond. No.	7.18

On observation, it can be seen that in the graphs for all four seasons the correlation values between Pythagorean Expectation and Win % (0.949, 0.910, 0.969, 0.915), are above 0.9. Also on observation of the regression results, the coefficient of determination (R^2) values are all relatively high, thus signifying that Pythagorean Expectation explains Win % considered over the entire season quite well statistically.



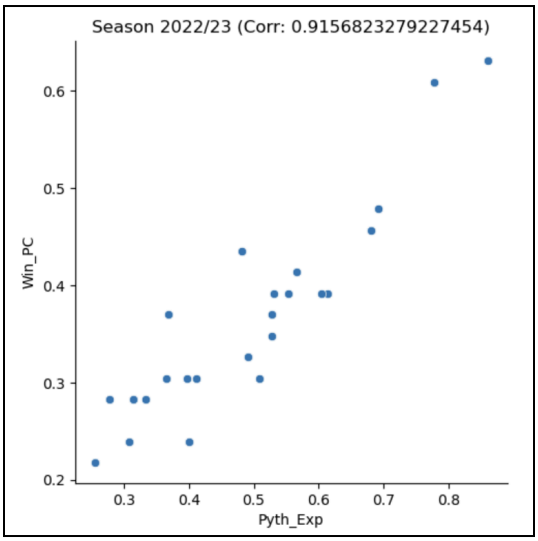
Season 2021/22

OLS Regression Results

Dep. Variable:	Win_PC	R-squared:	0.940
Model:	OLS	Adj. R-squared:	0.937
Method:	Least Squares	F-statistic:	341.9
Date:	Wed, 20 Mar 2024	Prob (F-statistic):	6.82e-15
Time:	21:55:26	Log-Likelihood:	53.878
No. Observations:	24	AIC:	-103.8
Df Residuals:	22	BIC:	-101.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0817	0.017	4.880	0.000	0.047	0.116
Pyth_Exp	0.5850	0.032	18.491	0.000	0.519	0.651

Omnibus:	1.117	Durbin-Watson:	1.908
Prob(Omnibus):	0.572	Jarque-Bera (JB):	0.847
Skew:	0.109	Prob(JB):	0.655
Kurtosis:	2.106	Cond. No.	7.27



Season 2022/23

OLS Regression Results

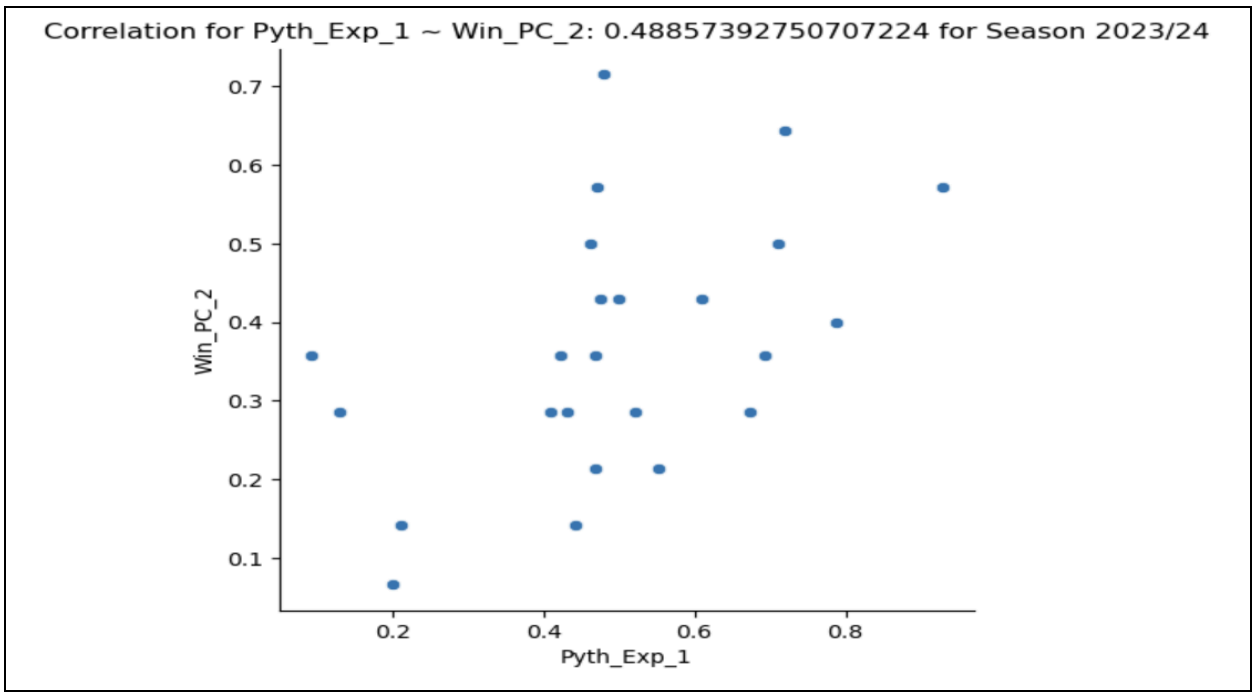
Dep. Variable:	Win_PC	R-squared:	0.838
Model:	OLS	Adj. R-squared:	0.831
Method:	Least Squares	F-statistic:	114.2
Date:	Wed, 20 Mar 2024	Prob (F-statistic):	3.56e-10
Time:	21:55:30	Log-Likelihood:	42.528
No. Observations:	24	AIC:	-81.06
Df Residuals:	22	BIC:	-78.70
Df Model:	1		
Covariance Type:	nonrobust		

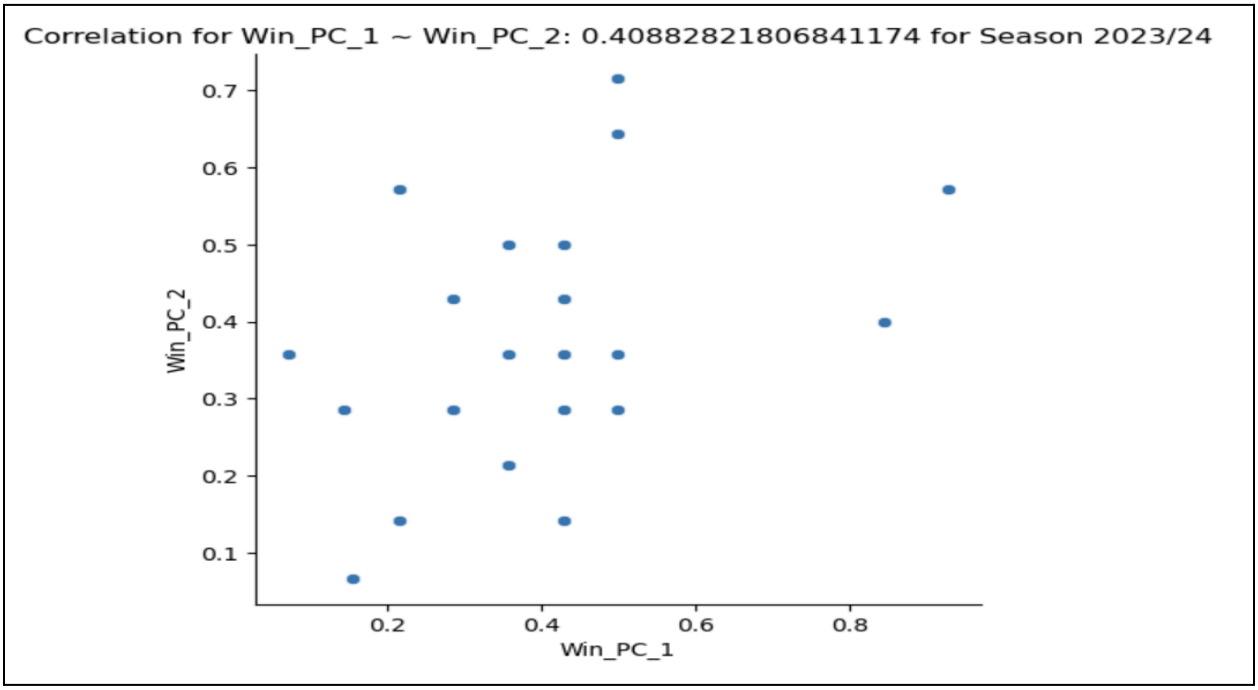
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0684	0.029	2.351	0.028	0.008	0.129
Pyth_Exp	0.6011	0.056	10.686	0.000	0.484	0.718

Omnibus:	0.747	Durbin-Watson:	2.201
Prob(Omnibus):	0.688	Jarque-Bera (JB):	0.744
Skew:	0.359	Prob(JB):	0.689
Kurtosis:	2.524	Cond. No.	8.01

How Pythagorean Expectation is a better predictor compared to Win Percentage?

This was shown by splitting the 2023/24 EFL dataset consisting of 28 games per team, to two datasets consisting of 14 games per team. Following that the predictability prowess of Pythagorean Expectation and Win % from the first dataset was tested against the actual Win % from the second dataset. The results are shown below and proves that the hypothesis was correct.





From the graphs, it can be inferred that Pythagorean Expectation is a better predictor than Win % since its correlation value is higher, 0.488 > 0.408. Therefore, this analysis and the analysis from the previous four seasons alludes to the fact that Pythagorean Expectation is a reliable predictor.

Predicting the result of the 2023/24 EFL Championship Season:

After assessing and showing the reliability of Pythagorean Expectation to be able to predict results in the previous seasons of the EFL Championship, the final step is to forecast the results of the rest of 2023/24 season and try to decipher any underlying anomalies. For this, a similar process was used and the following table's attribute "X_PTs" signifies the number of points that each team can be expected to reach based on the Pythagorean Expectation calculated from the first 28 games played by each team.

	Team	MP	X_PTs	Pyth
0	Leicester City	46	112	0.866298
1	Southampton	46	97	0.732846
2	Leeds United	46	96	0.787154
3	Ipswich Town	46	95	0.679822
4	West Bromwich Albion	46	82	0.686990
5	Coventry City	46	79	0.672608
6	Watford	46	71	0.583742
7	Sunderland	46	71	0.587554
8	Hull City	46	71	0.549875
9	Norwich City	46	69	0.533670
10	Middlesbrough	46	67	0.512046
11	Bristol City	46	64	0.500000
12	Cardiff City	46	61	0.446103
13	Plymouth Argyle	46	60	0.500000
14	Preston North End	46	57	0.353301
15	Swansea City	46	56	0.427219
16	Millwall	46	53	0.393542
17	Blackburn Rovers	46	53	0.374388
18	Stoke City	46	52	0.386737
19	Birmingham City	46	52	0.373868
20	Huddersfield Town	46	41	0.270345
21	Queens Park Rangers	46	38	0.274678
22	Sheffield Wednesday	46	32	0.186154
23	Rotherham United	46	29	0.187744

Conclusion:

While Pythagorean Expectation provides a solid foundation for predicting team success, its true value lies in its integration with broader analytical and game-specific, tactical frameworks to provide a comprehensive understanding and prediction of performance. Although Pythagorean Expectation is quite reliable within its parameters, it doesn't take non-statistical unpredictable factors such as injuries, weather conditions to name a few, into account. Along with that, soccer being a low-scoring, small dataset game, some of these factors may play a significant role. By utilizing Pythagorean Expectation alongside other contextual and tactical insights, game professionals can make informed decisions on a game-by-game basis to improve team performances.

Sources:

Note: AI softwares, ChatGPT and Gemini were used to assist in the drafting of this article, upon which statistical results and additional content from the author was added

Data for EFL Championship Seasons 2019/20, 2020/21, 2021/22, 2022/23, 2023/24 was provided by Samford University

Home of football statistics and history 11v11. (2024, March 16). League tables. Retrieved from https://www.11v11.com/league-tables/

Samford University. (2024, March 14). Sports Analytics - Fans. Retrieved from https://www.samford.edu/sports-analytics/fans/

StatsBomb. (2024, March 15). Improving soccer's version of the Bill James Pythagorean. Retrieved from https://statsbomb.com/articles/soccer/improving-soccers-version-of-the-bill-james-pythagorean/

About the Author:

This article was written by Sibi Karthik. Sibi is an undergrad junior studying Computer Science at Vellore Institute of Technology, India. Sibi's interest and work lies within the fields of Sports Data Analytics and Football Tactical Analysis. Having recently completed his Level 1 and 2 licenses in Coaching Football with the English FA, Sibi is working to continue his pursuit in the field of football and is looking to make a mark on the future of the game.

LinkedIn - https://www.linkedin.com/in/sibi-karthik-302a8224a/
GitHub Link to Project - https://github.com/sibi15/2023-24-Pythagorean-Expectation