EPL Statistical Analysis and Team Valuation

May 13, 2023

Zachary Knudsen

(760) 421-7732 | zachknudsen@icloud.com

BS in Finance/Data Science, Leonard N. Stern School of Business

Lucas Osiecki

(347) 401-2720 | losiecki3@gmail.com

BS in Finance/Entrepreneurship, Leonard N. Stern School of Business

ABSTRACT

Soccer is quite a difficult sport. It's about being patient, cohesive, methodical, and calm- yet it's also about being tenacious, mentally tough, and physically overpowering the other team. In this data report, we examine some of the statistics that correlate to success in the English Premier League; specifically, we analyze the importance of shots, corners, fouls, and yellow cards to ultimately predict how each team performs. We derived many interesting insights from this project, and we're excited to share our findings with you. Also, while we believe this is a comprehensive and thorough analysis, we are open to any advice or tips on how we can improve our model. You can reach out to us at our phone numbers or emails above to get in touch.

1 INTRODUCTION

We chose to tackle this project because of our passion for sports. As current student-athletes on the NYU men's volleyball team, sports take up a big chunk of our lives. We both played soccer throughout our childhood prior to going to college, and so we are quite acquainted with the skills and strategies necessary to win soccer games. Teams that have more shots and corner kicks tend to score more goals, while fouls and yellow cards both speak to a team's physicality, aggression, and overall desire to win. We chose to analyze these statistics specifically because we wanted to see 1) how important they were relative to each other in differentiating a win vs. a loss, and 2) how accurate of a model they could create. We had several other statistics available to us, but these stood out because they didn't indicate how much a team actually scored- they just provided a snapshot into how offensively driven and physical each team played. With this in mind, we wanted to see if the data supported our preconceptions on what contributed to a team's success.

In this project, we will first examine the relative importance of each statistic in determining a team's performance, or win proportion. Then, we will weigh each statistic by its importance to derive an equation that compares each team's average statistics to the overall averages of each statistic across teams from seasons 2011-2021.

BIG IDEA: The purpose of this project is to create a model that predicts each team's relative performance/win proportion given its average shots, corners, fouls, and yellow cards per game.

2 DESCRIPTION OF DATA

We created an original dataset by concatenating English Premier League data from seasons 2011-2021. The data set displayed the home team name, away team name, statistics such as home team goals scored "FTHG", away team goals scored "FTAG", home team shots on target, away team shots on target, etc. for every single game played from 2011-2021. The data frame distinguished who won or lost in a column called "FTR" or full-time result, which displayed either H "home", A "away"", or D "draw". We've included a data slice below to show what the data frame we originally created looked like.

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR
0	E0	13/08/11	Blackburn	Wolves	1	2	A	1	1	D
1	E0	13/08/11	Fulham	Aston Villa	0	0	D	0	0	D
2	E0	13/08/11	Liverpool	Sunderland	1	1	D	1	0	H
3	E0	13/08/11	Newcastle	Arsenal	0	0	D	0	0	D
4	E0	13/08/11	QPR	Bolton	0	4	A	0	1	A
...	...	...	...	...	...	...	...	...	...	...
375	E0	22/05/2022	Crystal Palace	Man United	1	0	H	1	0	H
376	E0	22/05/2022	Leicester	Southampton	4	1	H	0	0	D
377	E0	22/05/2022	Liverpool	Wolves	3	1	H	1	1	D
378	E0	22/05/2022	Man City	Aston Villa	3	2	H	0	1	A
379	E0	22/05/2022	Norwich	Tottenham	0	5	A	0	2	A

4180 rows × 23 columns

From this original data set, we filtered our data to only display wins and losses, therefore eliminating draws. We found that draws would create ambiguity in our data, as it would become much harder to find strong relationships between each team's statistics and relative performance.

We also reformulated our data set to display each team's shots, corners, fouls committed, and yellow cards, as well as whether the team won or lost. The way we went about this was through creating win data frames and loss data frames, and later concatenating them to have each tuple display an individual team's statistics per game, and whether the result was a win or a loss for that team.

This was somewhat complicated because of how the data was originally presented to us: we created win data frames by displaying the home team's statistics when the FTR was "H" and the away team's statistics when the FTR was "A". We did the exact opposite for our loss data frames, as we displayed the away team's statistics when the FTR was "H", and then displayed the home team's statistics when the FTR was "A". We then concatenated our data frames to display all wins and losses from 2011-2021, as well as the statistics for each team in separate tuples. We've included a slice of this data frame below.

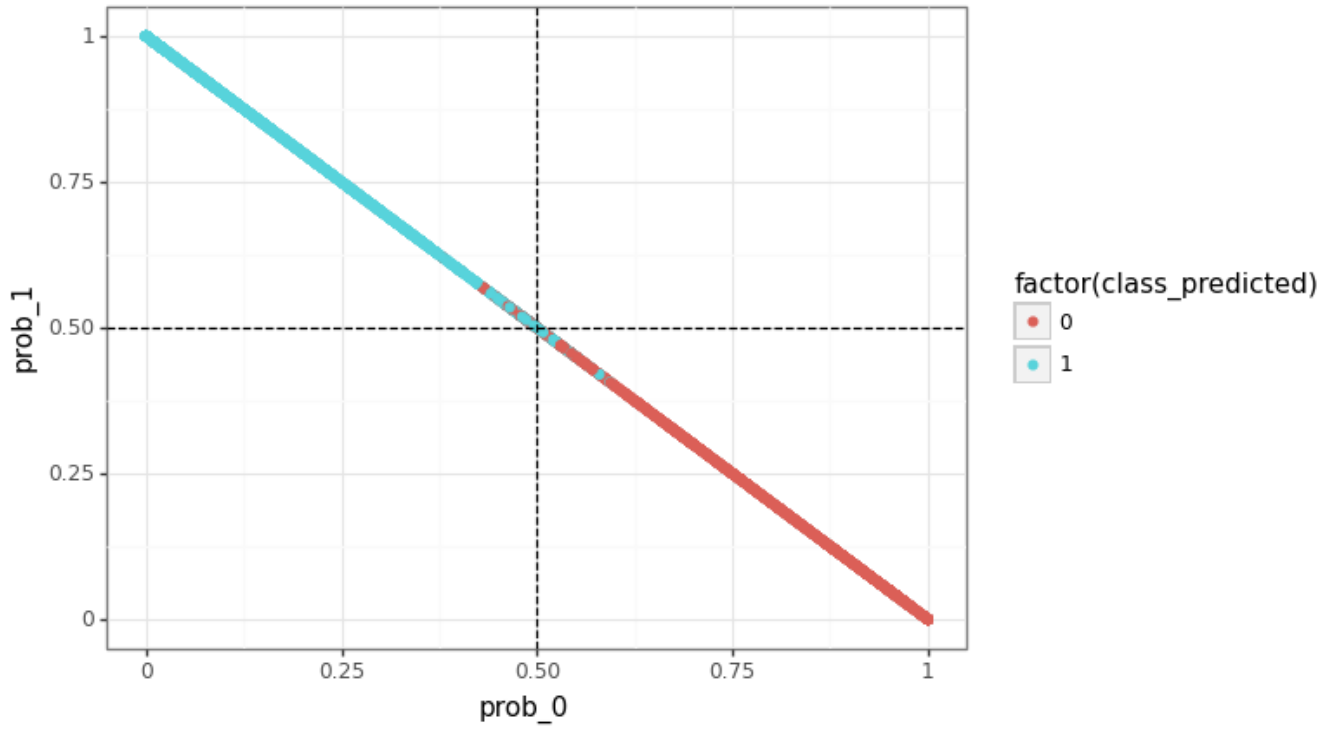
	Shots	Fouls Committed	Corners	Yellow Cards	Team	Result
8	26	9	7	0	Man City	W
10	9	9	5	3	Aston Villa	W
11	22	12	7	2	Chelsea	W
17	17	7	7	2	Wolves	W
18	27	12	11	1	Man United	W
...	...	...	...	...	...	...
364	10	8	5	0	Everton	L
366	4	6	1	2	Southampton	L
371	14	11	3	2	Brentford	L
373	12	8	8	1	Burnley	L
379	9	13	3	3	Norwich	L

6368 rows × 6 columns

3 METHODOLOGY TO CREATE VALUATION MODEL

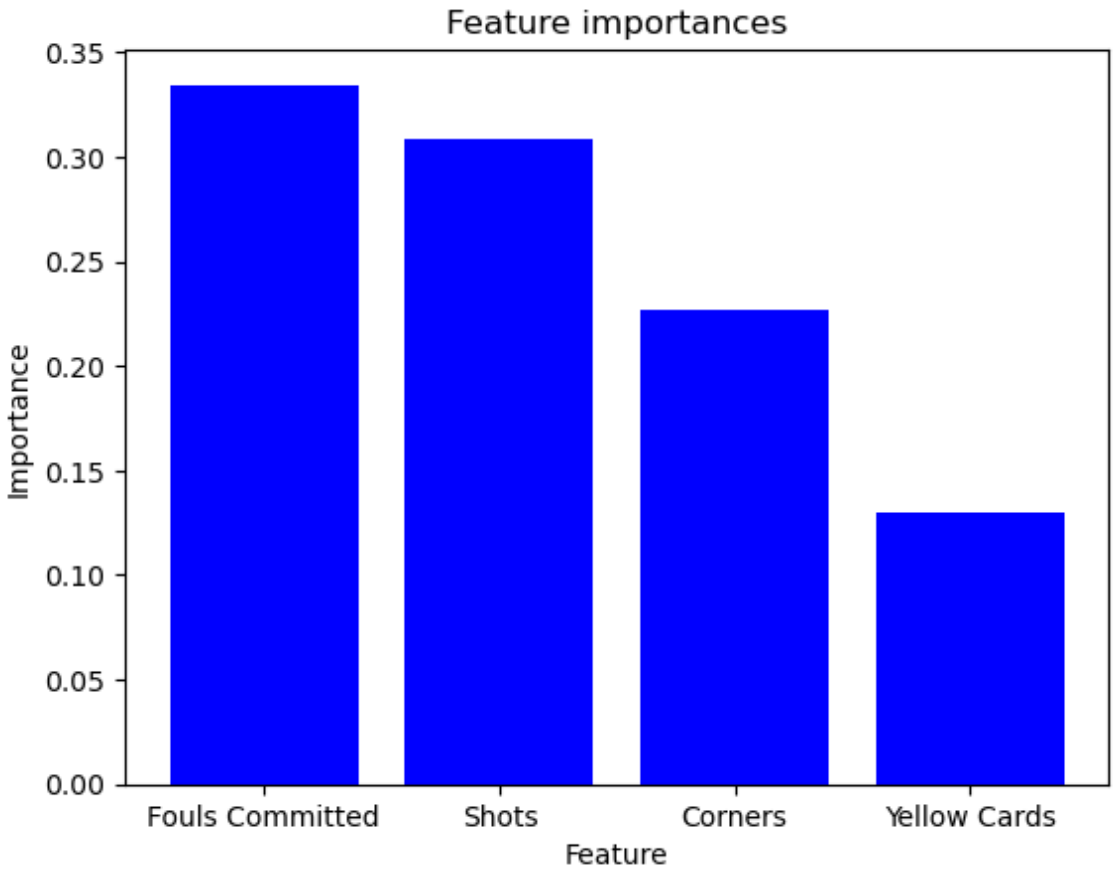
Organizing our data by showing only wins and losses made it extremely convenient to work with, as our data frame displayed a win-loss rate of 50%. Finding the averages for each statistic across teams would therefore equate to a predicted 50%-win rate, which we later use in our valuation model to help predict a team's relative performance/win proportion.

To weigh each statistic by its relative importance in determining a team's performance, we utilized a machine learning algorithm known as random forests. We converted our "Result" column to binary characters, denoting "L" as 0 and "W" as 1. We then executed a train-test split to see how well random forests could predict wins and losses based on each team's statistics.



We found that random forests predicted wins/losses correctly with 89% accuracy, per our classification report. With our train-test split successfully executed, we were finally ready to find our feature importance scores by using the *.feature_importances_* function.

This led us to perhaps our most intriguing finding in our data exploration. As experienced soccer players, we thought that shots or corners would have been the most relevant features in predicting a team's win proportion because they represented goal-scoring opportunities; however, we found that the number of fouls a team committed had the strongest impact.



This was quite shocking to us at first. However, after further reflecting on strategy in soccer, we realized that many high-level teams often commit fouls out of necessity- a phenomenon known as "strategic" fouling. This is especially prevalent in highly competitive situations, as teams with high levels of soccer IQ may commit fouls to prevent other teams from scoring goals. Teams commit strategic fouls for many different reasons, ranging from instances as dire as a 1 vs. 1, or just to prevent a counterattack. The more we thought about this finding, the more it really did make sense to us.

To find the data frame's average statistics per game, we used the *.sum()* function to add up each statistic, and then divided each sum by the number of tuples present in the table.
To find each individual team's average statistics per game, we used the *.sum()* and the *.groupby()* functions. This allowed us to add up all statistics and group them by different teams.

However, we realized that some teams were present in the EPL for only a few years due to relegation. To make sure the statistics we measured were equally comparable, we used the *.to_datetime* function to find how many seasons each team played in the EPL. We then divided the totals of each team's statistics by the number of years the team had been present in the EPL, which allowed us to derive the yearly averages of each statistic per team. Finally, we divided each team's yearly statistical averages by 38, since each team plays 38 games per season in the EPL. This gave us each team's average statistics per game.

4 RESULTS AND IMPLICATIONS

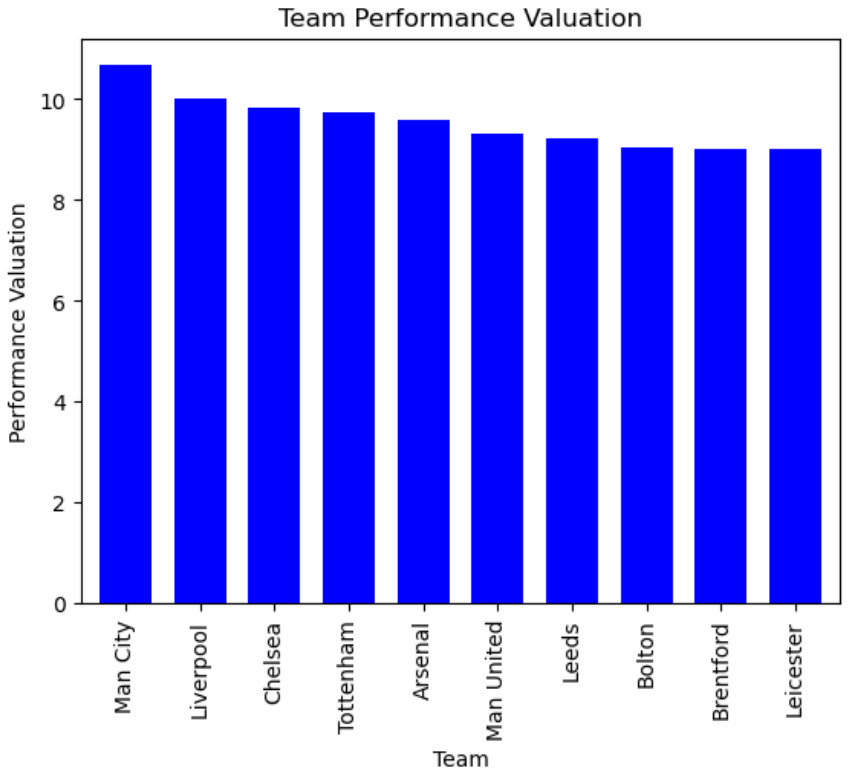
With our deep dive into the data complete, we were finally ready to create our valuation model. We've included the equation we used below to predict each team's relative performance.

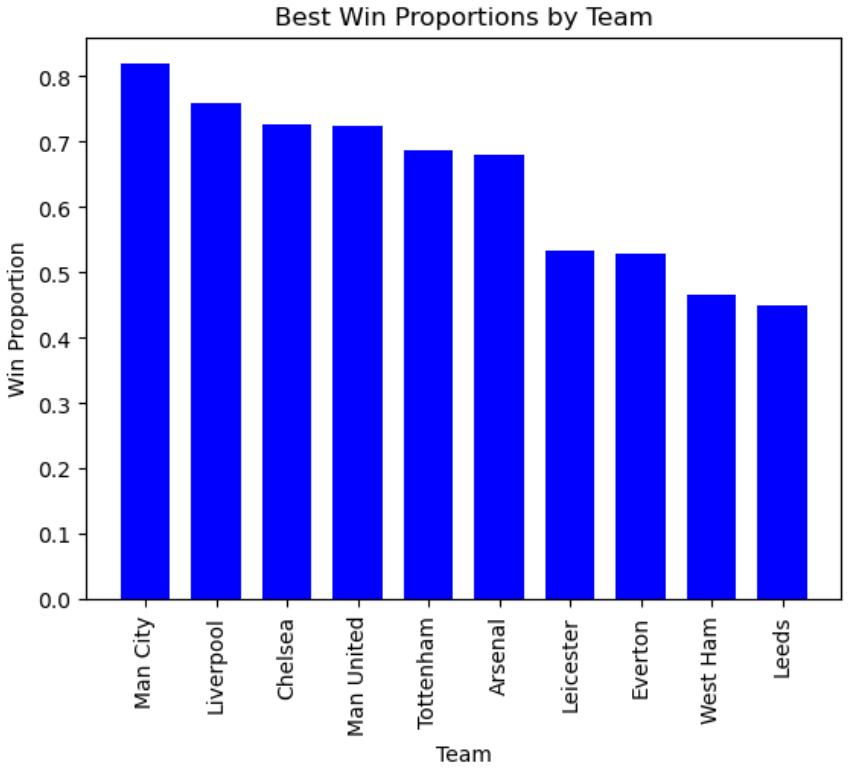
Valuation: ((Team avg shots/game – data frame avg shots/game)(feature importance)+ (Team avg fouls committed/game – data frame avg fouls committed/game)(feature importance)+ (Team avg corner kicks/game – data frame avg corner kicks game)(feature importance) + (Team avg yellow cards/game – data frame average yellow cards/game)(feature importance)) + 10 *

***** We included +10 to make sure our valuation model calculated positive values, which would allow us to make clearer visualizations

We were originally quite skeptical about the strength of this valuation model, as we still did not fully trust that the number of fouls a team committed exhibited the strongest correlation to a team's performance. However, the results from our valuation model demonstrate remarkable success in predicting the relative performance of English Premier League teams. Our model predicted that Manchester City was the winningest team, followed by Liverpool and Chelsea. This is consistent with the actual win proportions of these teams, as they have frequently been at the top of the league table. Our model also accurately included eight out of the top ten teams with the highest win proportions over the span of 11 seasons. This level of accuracy suggests that the features we selected for our model are indeed strongly associated with team success in the EPL.

We've included below two visualizations: (1) our valuation model's top-10 predictions and (2) the top-10 teams with the highest win proportions from 2011-2021.


Team Performance Valuation
Performance Valuation
0
2
4
6
8
10
Man City
Liverpool
Chelsea
Tottenham
Arsenal
Man United
Leeds
Bolton
Brentford
Leicester
Team


Best Win Proportions by Team
Win Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Man City
Liverpool
Chelsea
Man United
Tottenham
Arsenal
Leicester
Everton
West Ham
Leeds
Team

These results prove how the features we've analyzed have strong correlations to team performance. To reemphasize why our results make sense in this context: average shots per game is an intuitive measure of a team's offensive capability. The more shots a team can make, the higher their chances of scoring and ultimately winning the game. Average fouls committed per game and average yellow cards per game may seem counterintuitive as these actions are typically penalized. However, they could also indicate a team's aggressive style of play, which can be an effective strategy in soccer. Average corner kicks per game, while less directly influential, can provide additional scoring opportunities and might also indicate a team's ability to apply pressure on the opposition.

5 DISCUSSION AND SUMMARY

Our project set out to create a model that predicts EPL teams' relative performance based on their average shots, corners, fouls, and yellow cards per game. Our model's predictions aligned significantly with the actual win proportions of the teams we analyzed, thus indicating its effectiveness.

We first had to create our own data frame through concatenating data from several different years in the English Premier League. We then molded our data to disregard home and away factors, and instead focus on wins and losses. This enabled us to utilize machine learning to see how accurate our win/loss predictions could be.

We utilized random forests to find each feature's importance in terms of predicting power. In order from most to least influential on the result of a game: (1) Fouls Committed, (2) Shots, (3) Corners, (4) Yellow Cards.

While conducting this data exploration, we were able to apply the skills we had learned throughout our Data Bootcamp course at NYU. We also learned a lot more about soccer through a statistical lens, as we were shocked that fouls had such a strong relative correlation to team performance.

This finding is supported, however, by a multitude of external sources. Ashwin Phatak, Robert Rein, and Daniel Memmert of the Institute of Exercise Training and Sport Informatics at German Sports University affirm this conjecture, as they studied the relationship between fouling and team success in several different soccer leagues. They found that their "results may suggest that fouling while out of possession is a good strategy for success in every league, but it seems to be the most effective in the EPL."

It is important to note that while our model's predictions were largely accurate, there were some discrepancies. This suggests that there may be other factors influencing team performance that we did not account for in our model. Soccer is a complex game, and many variables can affect the outcome of a match, including player skill levels, team chemistry, the influence of the coach, and even factors such as weather and home-field advantage. Future research might explore these other factors to build a more comprehensive model of team performance.

In summary, our project demonstrates the potential of statistical analysis and machine learning in predicting sports outcomes. Our findings could be of interest to soccer analysts, sports statisticians, and even EPL teams looking for data-driven insights to improve their performance.

REFERENCES

1. https://www.kaggle.com/datasets/thedevastator/uncovering-betting-patterns-in-the-premier-leagu?select=2011.csv
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8607766/