FiveThirtyEight's club football predictions

Sports modelling reading group

Clement Lee

2023-04-28

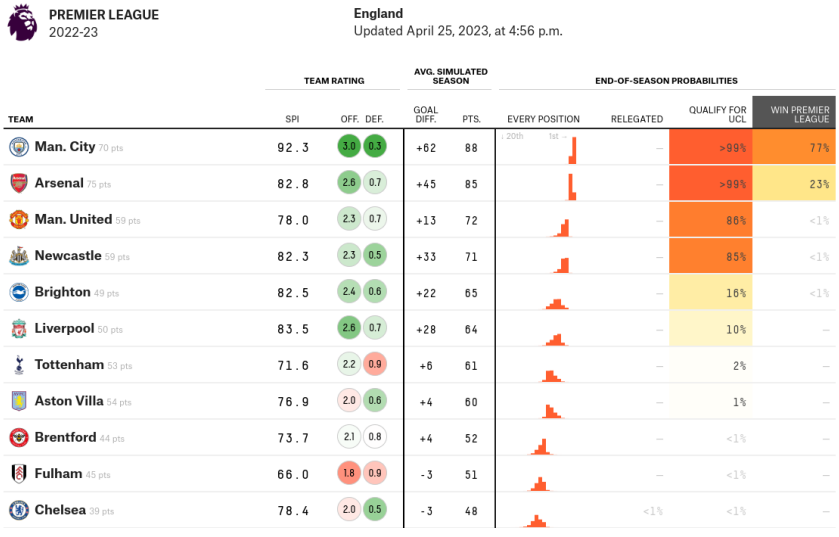
Nate Silver

Nate Silver

- Predicted the outcomes in all but one states in 2008 U.S. presidential election
  - 2008: All states but one
  - 2012: All states
  - 2016: Least worst of major forecasts
  - 2020: All states
- 538 electors in the Electoral College

FiveThirtyEight

- Does other forecasts as well



PREMIER LEAGUE
2022-23

England
Updated April 25, 2023, at 4:56 p.m.

		TEAM RATING		AVG. SIMULATED SEASON		END-OF-SEASON PROBABILITIES			
TEAM	SPI	OFF.	DEF.	GOAL DIFF.	PTS.	EVERY POSITION (20th – 1st)	RELEGATED	QUALIFY FOR UCL	WIN PREMIER LEAGUE
Man. City 70 pts	92.3	3.0	0.3	+62	88		–	>99%	77%
Arsenal 75 pts	82.8	2.6	0.7	+45	85		–	>99%	23%
Man. United 59 pts	78.0	2.3	0.7	+13	72		–	86%	<1%
Newcastle 59 pts	82.3	2.3	0.5	+33	71		–	85%	<1%
Brighton 49 pts	82.5	2.4	0.6	+22	65		–	16%	<1%
Liverpool 50 pts	83.5	2.6	0.7	+28	64		–	10%	–
Tottenham 53 pts	71.6	2.2	0.9	+6	61		–	2%	–
Aston Villa 54 pts	76.9	2.0	0.6	+4	60		–	1%	–
Brentford 44 pts	73.7	2.1	0.8	+4	52		–	<1%	–
Fulham 45 pts	66.0	1.8	0.9	-3	51		–	<1%	–
Chelsea 39 pts	78.4	2.0	0.5	-3	48		<1%	<1%	–

Outline

1. Soccer Power Index (SPI)
   - Nate originally devised for ESPN
   - Revised substantially over the years
2. Forecasting
   - Poisson model
   - Monte Carlo simulations
3. Leagues and tiers
   - For e.g. Champions League

Soccer Power Index (SPI)

What is SPI?

- Overall rating: **the percentage of available points expected to take**
- Used in turn for the simulations

Pre-season SPI

- $2/3 \times$ end-of-season's SPI $+ 1/3 \times$ SPI implied by market value
- Reason: market value strongly correlated with end-of-season SPI

SPI during season

- Updated after each match
- Calculated according to an **offensive rating** and **a defensive rating**
  - Opaque about how

TEAM	SPI	OFF.	DEF.
Man. City	92.3	3.0	0.3
Real Madrid	86.9	2.7	0.6
Inter Milan	77.7	2.3	0.7
AC Milan	74.5	2.1	0.8

Offensive rating

- Goals expected to score against an "average" team
- Using actual goals?
  - Scoreline often disagrees with people's impressions of the quality of each team's play
  - Low-scoring nature $\rightarrow$ prolonged periods of luck (good results despite playing poorly), or vice versa

Offensive rating

Average of

1. ~~Goals~~
2. Adjusted goals
3. Shot-based expected goals
4. Non-shot expected goals

Defensive rating = offensive rating of opposing team

- High defensive rating is bad

Adjusted goals

- Reducing value of goals scored when more players
- Reducing value of goals scored late when leading
- Increasing value of other goals to add up to total # actual goals

Shot-based expected goals

- Each shot assigned a probability based on distance & angle
- Part of the body the shot was taken with
- Adjustment for the player

Non-shot expected goals

- Passes, interceptions, take-ons and tackles
- Intercepting the ball at opposing team's penalty spot leads to a goal $\approx 9\%$ of the time
- Pass completed at the center of the six-yard box leads to a goal $\approx 14\%$ of the time

SPI pipeline

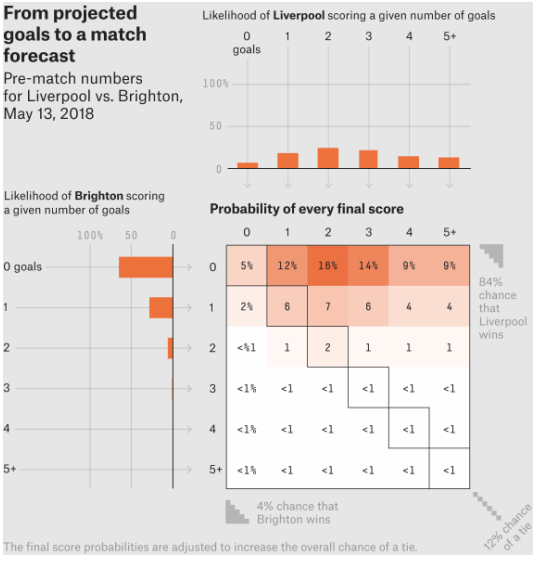
1. End-of-season SPI & market value $\rightarrow$ pre-season SPI
2. Use current SPI to forecast match result (next section)
3. Use match data to calculate offensive & defensive ratings
4. Update current SPI after match using the ratings
   - Unlike ELO, a win doesn't necessarily improve SPI

Forecasting

Simulating number of goals

- Poisson with mean the offensive rating
- Adjusted for
  - league-specific home-field advantage
  - **importance** of the match to each team
- Two independent Poisson distributions
- Inflated for draws (league-specific, around 9%)

Scoring matrix



Importance, quality, match rating

TIME (ET)	LEAGUE	MATCH	WIN	DRAW	QUALITY The harmonic mean of the teams' SPI ratings	IMPORTANCE How much the result will alter the league projections	MATCH RATING The average of quality and importance
Wednesday, April 26 11 a.m.	Super League Greece	Volos Panathinaikos	13% 62%	26%	34	50	42
Wednesday, April 26 12 p.m.	Super League Greece	Aris Olympiacos	32% 41%	27%	51	9	30
Wednesday, April 26 12:30 p.m.	Bundesliga Austria	Sturm Graz Austria Vienna	51% 24%	25%	53	11	32
Wednesday, April 26 1:30 p.m.	La Liga Spain	Getafe Almería	50% 23%	27%	59	70	64
Wednesday, April 26 1:30 p.m.	La Liga Spain	Atlético Madrid Mallorca	62% 13%	25%	73	1	37

Monte Carlo simulations

- Using the scoring matrix
- Teams' SPI can rise and fall depending on simulated matches played
- Widened distribution of possible outcomes

Whole pipeline

1. End-of-season SPI & market value $\rightarrow$ pre-season SPI
2. Use current SPI to forecast match result (next section)
3. Use match data to calculate offensive & defensive ratings
4. (Update current SPI after match using the ratings)
5. Build scoring matrix using offensive ratings & draw inflation
6. Simulate the (remaining) matches of the season

Performance - Ranked Probability Score (RPS)

- Measures how good forecasts are in matching observed outcomes
- RPS $= 0$: wholly accurate; RPS $= 1$: wholly inaccurate
- Current RPS $= 0.1957$ after reductions:
  - by 0.0018 using expected-goals metrics
  - by 0.0011 using market values to pre-season SPI
  - by 0.0006 using match importance

Leagues and tiers

Relative strengths between leagues

- For predictions of e.g. Champions League
- Originally used a tiered system

European countries' soccer leagues, sorted into six tiers by strength

TIER	LEAGUES
1	England, Germany, Italy, Spain
2	France
3	Portugal
4	Belgium, Czech Republic, Netherlands, Russia, Ukraine
5	Austria, Bulgaria, Croatia, Denmark, Finland, Greece, Hungary, Ireland, Israel, Norway, Poland, Romania, Scotland, Slovakia, Slovenia, Sweden, Switzerland, Turkey
6	Albania, Andorra, Armenia, Azerbaijan, Belarus, Bosnia, Cyprus, Estonia, Faroe Islands, Georgia, Iceland, Kazakhstan, Latvia, Lithuania, Luxembourg, Macedonia, Malta, Moldova, Montenegro, Northern Ireland, Serbia, Wales

Top divisions only. Countries are listed alphabetically within each tier.

Current version

1. Calculate SPI using domestic league results only
2. Calculate expected scores of inter-league matches using domestic SPI
3. Apply Massey's method to actual − expected scores to find **league strengths**
4. Regress league strengths toward market-value based ratings
5. Run through all matches again, this time with the league strengths

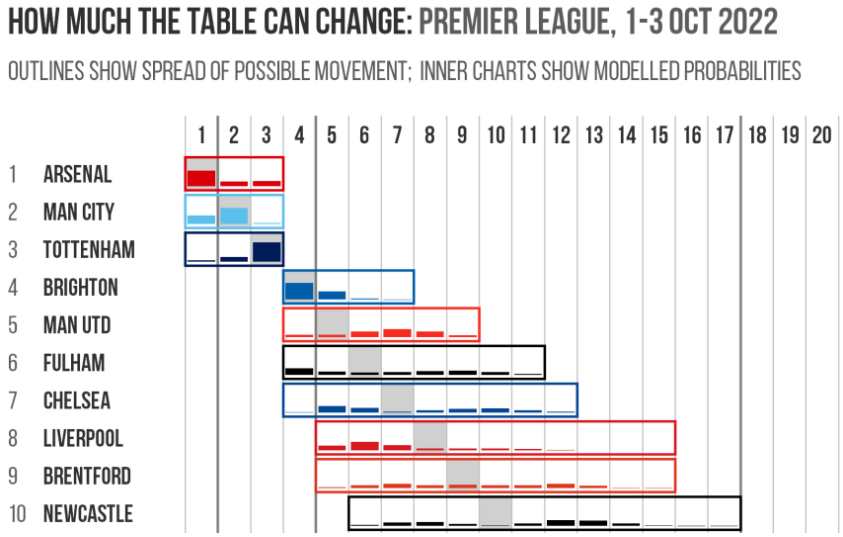
Interpretation: bonus (in goals) given to each team in an inter-league match

Summary

Modularised approach

1. Match data to calculate offensive/defensive ratings
2. Offensive ratings to calculate scoring matrix, with adjustments
3. Scoring matrix to simulate future matches
4. Domestic results to compute league strengths

Extra 1: @Experimental361



Extra 2: Engsoccerdata GitHub repository

https://github.com/jalapic/engsoccerdata

engsoccerdata

This R package is mainly a repository for complete soccer datasets, along with some built-in functions for analyzing parts of the data. Currently I include three English ones (League data, FA Cup data, Playoff data - described below), several European leagues (Spain, Germany, Italy, Holland, France, Belgium, Portugal, Turkey, Scotland, Greece) as well as South Africa and MLS.

Free to use for non-commerical use. Compiled by James Curley.

Please cite as:
James P. Curley (2016). engsoccerdata: English Soccer Data 1871-2016. R package version 0.1.5