Prediction of Premier League 2024 Champion Using Logistic Regression Model[*]

Can Manchester City Continue Their Championship Reign?

Xiyan Chen

December 3, 2024

This study predicts whether Manchester City can maintain their championship status in the 2024 Premier League season. By analyzing historical performance and using models based on match-specific probabilities and the Soccer Power Index (SPI), we assess their chances of continuing success. Our results suggest that Manchester City has a high probability of remaining at the top, supported by their consistent performance. This research highlights the value of statistical models in sports analytics and provides insights into the factors influencing championship outcomes.

1 Introduction

As the 2024 Premier League season unfolds, Manchester City stands as a formidable contender for another championship title. Our predictive analysis suggests that they have a strong likelihood of continuing their dominance, leveraging key factors such as team performance metrics, historical data, and the Soccer Power Index (SPI). By utilizing advanced statistical models, including linear regression and probability analysis, we aim to provide an objective forecast of Manchester City's chances of maintaining their winning streak. This analysis is driven by the growing interest in data-driven sports predictions, offering insights not only for fans and analysts but also for teams looking to refine their strategies. Understanding the factors that contribute to success in the Premier League is crucial for shaping future expectations and enhancing the sports analytics field.

Overview paragraph

Estimand paragraph

[*] Code and data are available at: https://github.com/RohanAlexander/starter_folder.

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

2 Data

2.1 Overview

The dataset used in this analysis was sourced from FiveThirtyEight (FiveThirtyEight 2022). All analyses were conducted in R (R Core Team 2023). Data cleaning and manipulation were performed using `tidyverse` (**tidyverse?**), with specific tasks carried out using `here` (Müller 2020), `readr` (**readr?**), and `dplyr` (Wickham et al. 2023). Data visualization was done using `ggplot2` (**ggplot2?**), while model evaluation were supported by `modelsummary` (Arel-Bundock 2022).

The dataset from FiveThirtyEight includes historical performance data for Premier League teams, featuring key match statistics and team strength indicators. For this analysis, we focused on match outcomes, team performance indicators such as the Soccer Power Index (SPI), and team rankings to predict the 2024 Premier League champion.

2.2 Dataset Overview

The dataset consists of data from multiple seasons of the Premier League, and key variables used in this analysis include:

- **Team Name**: The name of the team playing in the Premier League.
- **Match year**: The year when a match was played.
- **Home Team**: The team playing at home.
- **Away Team**: The team playing away.
- **Home probability**: The probability of winning for the home team in each match.
- **Away probability**: The probability of winning for the away team in each match.
- **SPI (Soccer Power Index)**: A numerical representation of a team's strength, used as an important predictor of future performance.

Because this paper is focusing on if Manchester City can win the championship, so there're variables focus on the perspectives of Manchester City: - **outcome**: The outcome that shows if Manchester City wins the game. - **man_city_spi**: The Soccer Power Index for Manchester City - **man_city_prob**: The probability of winning for Manchester City on that game - **is_home**: The variable shows if Manchester City is play at home on each match

The dataset was cleaned and pre-processed to focus on matches from the last six seasons, with the aim of using recent performance data for predictions.

2.3 Measurement

In this paper, we aim to predict the winner of the 2024 Premier League based on historical match data. The raw data was sourced from FiveThirtyEight's open dataset, which compiles match outcomes and team statistics over several Premier League seasons. This dataset includes key performance indicators such as the Soccer Power Index (SPI), match results, and rankings. The match data is linked to real-world events like football matches, with each match having a specific date, home team, away team, and scores. These variables are measured as follows: the match date is recorded to establish a time-based reference for analyzing team performance over the course of a season; home team and away team are categorical variables indicating which teams were playing at home or away, as location often influences match outcomes; home score and away score represent the number of goals scored by the respective teams, which directly reflect match outcomes (win, loss, or draw); the SPI is a numerical score that measures a team's strength, based on factors like goals scored, goals conceded, and opposition quality, providing insight into team potential beyond just match results. Additionally, team ranking is recorded, which reflects each team's position on the league table based on their performance over the season. These measurements allow us to capture not only the outcome of individual matches but also broader trends in team performance, which are critical for predicting the overall champion. Data preprocessing steps such as imputing missing values and calculating rolling averages for SPI and win/loss streaks were applied to refine the dataset for prediction.

2.4 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from @

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.5 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) & (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i & (2)\\
\alpha &\sim \text{Normal}(0, 2.5) & (3)\\
\beta &\sim \text{Normal}(0, 2.5) & (4)\\
\gamma &\sim \text{Normal}(0, 2.5) & (5)\\
\sigma &\sim \text{Exponential}(1) & (6)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table **??**.

The model results in Table **??** indicates that the logistic regression model shows that The Soccer Power Index (man_city_spi) and whether the team playing at home influencing Manchester City's probability of winning. The intercept is negative (-8.437), representing the log-odds of Manchester City winning when all other variables are zero (i.e., when man_city_spi is 0 and the game is an away match). The p-value of 0.0421 suggests that the intercept is statistically significant at the 5% level.

The Soccer Power Index (man_city_spi) has a positive coefficient (0.09712), meaning that as Manchester City's SPI increases, their odds of winning also increase. This relationship is statistically significant, with a p-value of 0.0318. The significance of this variable confirms that the SPI is a meaningful predictor of Manchester City's performance. Additionally, the is_home variable, which accounts for the home advantage, also shows a positive and highly significant relationship with the outcome. The coefficient (1.15870) indicates that playing at home significantly improves Manchester City's chances of winning, and the very low p-value (9.45e-05) underscores the strength of this effect. These findings align with general observations in sports, where teams tend to perform better at home due to familiar conditions and supportive crowds.

The model's performance metrics also show improvement. The reduction in deviance from the null deviance (312.61) to the residual deviance (291.84) indicates that including predictors like man_city_spi and is_home improves the model's explanatory power. Furthermore, the Akaike Information Criterion (AIC) value of 297.84 suggests a reasonable balance between model fit and complexity, though it's best used in comparison with other models.

To further evaluate the model, it is essential to assess its predictive performance using metrics like accuracy, precision, recall, and the Area Under the Curve (AUC). Techniques like ROC curves or confusion matrices can provide additional insights into how well the model performs on unseen data. If the model's performance needs enhancement, incorporating additional features such as the strength of opponents, team form, or historical trends could improve prediction accuracy.

Finally, for better stability and to address potential issues like overfitting or multicollinearity, applying regularization techniques like Lasso or Ridge regression could be considered. Moreover, interpreting the coefficients in terms of odds ratios reveals that a one-unit increase in SPI corresponds to approximately a 10% increase in the odds of Manchester City winning, highlighting the practical implications of the predictors.

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

There are limitations in the measurements, such as potential bias in the home team advantage and the inability of SPI to account for sudden changes in team dynamics (like player injuries or transfers), which could affect match outcomes. Despite these limitations, the dataset provides reliable data for predicting the 2024 Premier League champion based on historical trends and current team performance. The Soccer Power Index (SPI) is derived from past performance, which may not always reflect a team's true potential. Factors like injuries or player transfers, which could drastically affect team performance, are not always captured in the SPI.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC algorithm

References

Arel-Bundock, Vincent. 2022. “modelsummary: Data and Model Summaries in R.” *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

FiveThirtyEight. 2022. “Dataset Title (’Soccer Club Spi Matches’).” https://projects.fivethirtyeight.com/soccer-api/club/spi_matches.csv.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.