DSA 210: Introduction to Data Science

Fall 2025–2026

VAR Impact on Home Advantage in Football

Prepared by:

Ege Acar

Student ID: 29415

egeacar@sabanciuniv.edu

Submission Date: 9 January 2026

Table of Contents

1. Motivation

Football is a global phenomenon with deep cultural, social, and economic implications. Among the many factors that influence match outcomes, **home advantage** has long been recognized as a significant element. Teams playing at home tend to win more frequently, driven by psychological factors such as crowd support, familiarity with the playing surface, and reduced travel fatigue. Additionally, referee decisions may be subtly influenced by the home crowd, potentially favoring the home team in close calls.

The introduction of the **Video Assistant Referee (VAR)** system represents one of the most significant technological shifts in modern football. First implemented in various European leagues between 2017 and 2019, VAR was designed to minimize referee errors by providing video review capabilities for critical match decisions. The system aims to enhance the fairness and accuracy of officiating, potentially reducing the influence of subjective biases that may have historically favored home teams.

This project seeks to empirically investigate whether VAR's implementation has measurably altered home advantage patterns across 6 European football leagues spanning 11 seasons of data (2014-15 through 2024-25).

2. Research Questions

This project is guided by the following key research questions:

1. Has home advantage decreased since the introduction of VAR?
    a. Has there been a significant decline in home-win rates following the implementation of VAR?
    b. How has VAR affected performance indicators such as points, goals, and expected goals (xG) for home and away teams?
    c. Has the use of VAR influenced referee decisions in a way that reduces the traditional advantage of home teams?
2. Does the impact of VAR on home advantage vary across leagues?
3. How has the correlation between home points and final league ranking changed after VAR?

3. Data Sources

The datasets for this project will be obtained from the following publicly available sources:

3.1 Dataset 1: StatHead by Sports Reference

A comprehensive football statistics platform providing season, team and match-level data (accessed via subscription).

Source: https://stathead.com/fbref/

Key variables: Home and away match counts, wins/draws/losses, points per match, final league ranking, goals per 90 minutes, expected goals (xG), and penalties.

3.2 Dataset 2: Club Football Match Data (2000-2025) – Kaggle

An open dataset from Kaggle, containing match-level data across many leagues and seasons.

Source: https://www.kaggle.com/datasets/adamgbor/club-football-match-data-2000-2025

Key variables: Fouls, yellow cards, and red cards (home and away).

3.3 Temporal and Geographic Coverage

Time span: 2014–15 to 2024–25 seasons (11 seasons)

Leagues: English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, French Ligue 1, Turkish Süper Lig

VAR introduction timeline:

- Bundesliga: 2017–18
- Serie A: 2017–18
- La Liga: 2018–19
- Ligue 1: 2018–19
- Süper Lig: 2018–19
- Premier League: 2019–20

Final dataset size: ~1,281 rows (team-seasons) × 51 columns

4. Data Collection and Preparation

The quality and structure of the final dataset directly impact analysis validity. This section outlines the comprehensive process of transforming raw data into a clean, consistent, analysis-ready format.

4.1 Data Collection Process

Data was collected separately from two primary sources. Dataset 1 (StatHead) provided team and season-level data requiring merging. Dataset 2 (Kaggle) provided match-level data requiring aggregation to season level.

4.2 Data Cleaning, Standardization and Merging

Raw datasets underwent extensive cleaning: date standardization to consistent datetime format, column name alignment across datasets, numeric type conversions, complex merging of datasets and missing value treatment through external source retrieval when appropriate.

4.3 Output

The final dataset contains 1,281 team-season observations with 51 variables covering all six leagues across 11 seasons (2014–15 to 2024–25).

5. Exploratory Data Analysis

Exploratory Data Analysis revealed critical patterns and relationships within the dataset before applying formal statistical tests.

5.1 Overview and Summary Statistics

All core outcome variables such as home/away wins, draws, losses, goals, and points per match are fully observed with no missing values. Advanced metrics such as expected goals (xG) and disciplinary statistics contain missing values, primarily concentrated in earlier.

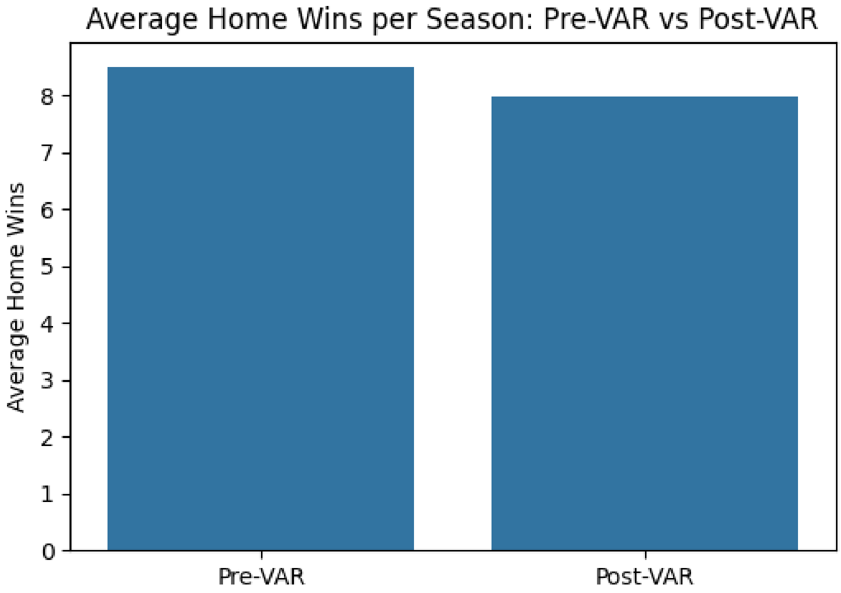
Home win rate ranges from 0.12 to 0.59, with a mean of approximately 0.44. Home points per match vary between 0.47 and 2.65, reflecting large differences in team strength.

5.2 VAR Era Classification

A binary indicator variable (VAR_era) was constructed based on league-specific VAR introduction seasons. Using this definition, 446 observations fall into the pre-VAR period and 835 observations into the post-VAR period.

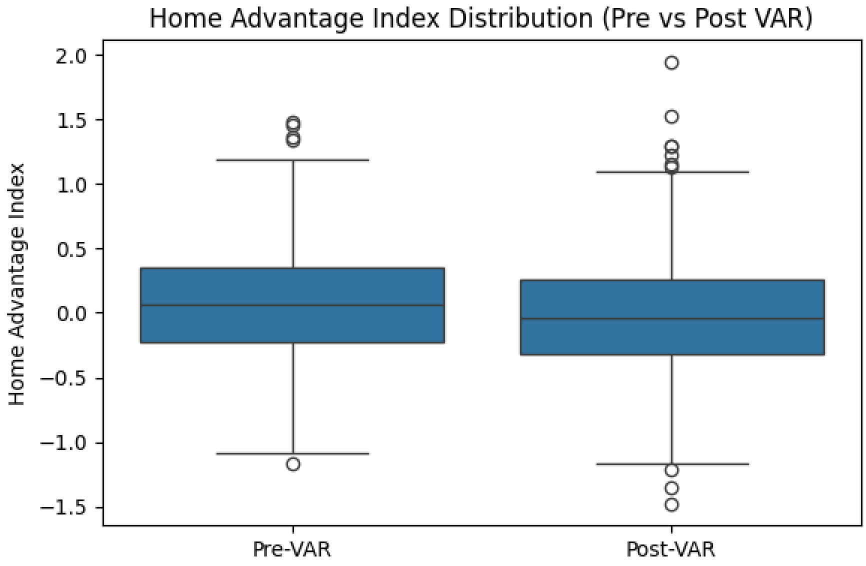
5.3 Home Wins Before and After VAR

Average home wins per season are lower in the post-VAR period compared to the pre-VAR period. The difference is small in magnitude and varies across leagues.



5.4 Home Advantage Index

A composite Home Advantage Index was constructed using standardized home–away differences across wins, points, goals, xG, fouls, cards, and penalties. The distribution of the index is similar before and after VAR, with comparable variability and closely aligned median values.



5.5 Relationship with Final League Ranking

The correlation between the Home Advantage Index and final league position is weak and negative in both periods.

6. Hypothesis Testing

Formal hypothesis testing was conducted to statistically validate observed patterns.

6.1 Hypotheses

Null Hypothesis (H_0): The implementation of VAR is not significantly affecting home advantage.

Alternative Hypothesis (H_1): The implementation of VAR is significantly reducing home advantage.

6.2 Methodology

Test 1: Home Wins

- t-statistic: 2.699
- p-value: 0.007
- Mean Pre-VAR: 8.50 home wins per season
- Mean Post-VAR: 7.97 home wins per season

Statistically significant decrease ($p < 0.01$), allowing rejection of null hypothesis. The decline, while modest, indicates measurable reduction in home advantage.

Test 2: Home Win Rate

- t-statistic: 2.484
- p-value: 0.013
- Mean Pre-VAR: 0.461 (46.1%)
- Mean Post-VAR: 0.435 (43.5%)

Home win rate shows meaningful decrease ($p < 0.02$), aligning with Test 1 and supporting the hypothesis.

Test 3: Home Points per Match

- t-statistic: 2.404
- p-value: 0.016
- Mean Pre-VAR: 1.63 points per match
- Mean Post-VAR: 1.56 points per match

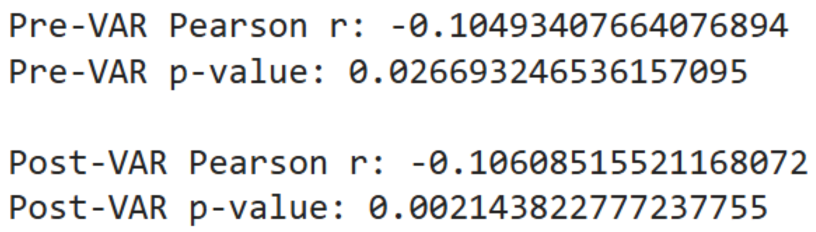
Test 4: Home Advantage Index

- t-statistic: 4.386
- p-value: 1.29×10^{-5} (< 0.001)
- Mean Pre-VAR: 0.083
- Mean Post-VAR: -0.033

The Home Advantage Index drops from positive to negative, indicating shift toward balanced home-away conditions. Highly significant (p < 0.001), providing strong evidence for VAR's association with home advantage reduction.

Test 5: Correlation Change Analysis

Pearson correlation between Home Advantage Index and Final League Position remained weak and negative in both eras (r ≈ -0.10), indicating VAR did not materially change how home advantage relates to final league performance.



```
Pre-VAR Pearson r: -0.10493407664076894
Pre-VAR p-value: 0.026693246536157095

Post-VAR Pearson r: -0.10608515521168072
Post-VAR p-value: 0.002143822777237755
```

6.3 Overall Summary

Across all major performance indicators, home advantage declines after VAR. Effects are statistically significant but modest in magnitude. Results provide consistent evidence supporting the alternative hypothesis: VAR contributes to a modest but measurable reduction in home advantage.

7. Machine Learning Methods

Supervised machine learning techniques were applied to model and predict aspects of home advantage through classification and regression tasks.

7.1 Classification Task

Binary target variable: home_advantage_win (1 if home win rate > away win rate, 0 otherwise). Distribution: 997 observations with home advantage, 284 without.

Features selected: home_goals_for_per_90, home_goals_against_per_90, away_goals_for_per_90, away_goals_against_per_90, home_xg_avg, away_xg_avg, home_penalties_won, away_penalties_won.

Logistic Regression Results:

- Accuracy: 82.1%

- Precision (class 0): 0.71, (class 1): 0.85 / Recall (class 0): 0.54, (class 1): 0.92 / F1-score (class 0): 0.61, (class 1): 0.88

Random Forest Results:

- Accuracy: 81.4%

- Precision (class 0): 0.83, (class 1): 0.81 / Recall (class 0): 0.37, (class 1): 0.97 / F1-score (class 0): 0.51, (class 1): 0.89

Both models demonstrate home advantage can be predicted from performance features with reasonable accuracy (~81-82%), though prediction remains moderate, aligning with findings that home advantage exists but is not overwhelmingly strong.

7.2 Regression Task

Linear regression modeled home_points_per_match magnitude using the same features.

- R^2 Score: 0.906 (explains 90.6% of variance)

- Mean Squared Error (MSE): 0.023

- Mean Absolute Error (MAE): 0.119

High R^2 indicates match performance indicators are strong predictors of home performance intensity. Low MAE (0.119) shows predictions typically accurate within approximately one-eighth point per match.

8. Key Findings

1. VAR is associated with a statistically significant reduction in several home-advantage metrics, including home wins, home win rate, home points per match, and the composite Home Advantage Index.

2. Average home wins per season declined from 8.50 to 7.97, and home win rates decreased from 46.1% to 43.5% after VAR introduction.

3. The Home Advantage Index shifted from positive (+0.083) in the pre-VAR period to negative (−0.033) post-VAR, with a highly significant difference ($p < 0.001$).

4. Penalty decisions increased for both home-favourable and home-against calls after VAR, with no clear directional shift favoring home teams.

5. League-specific variation in home advantage is observed, with changes differing across competitions.

6. A notable dip in home advantage is observed around the 2019–2021 period, coinciding with seasons played without spectators.

7. Machine learning models achieved 81–82% classification accuracy and a regression R^2 of approximately 0.91.

8. The correlation between home advantage and final league ranking remains weak ($r \approx -0.10$) and stable across both pre- and post-VAR periods.

9. Home advantage remains present after VAR implementation but at a reduced magnitude compared to the pre-VAR era.

9. Future Work

Incorporating additional domestic leagues or longer historical periods would allow a broader comparative perspective on how VAR interacts with different competitive environments.

The inclusion of additional performance and officiating metrics, provided that consistent data availability is ensured, could help refine the measurement of home advantage beyond the current indicators.

Repeating the analysis as new post-VAR seasons become available would allow for a more robust assessment of whether the observed patterns persist as leagues, referees, and teams further adapt to VAR over time.

10. Conclusion

This study examined the impact of the Video Assistant Referee (VAR) system on home advantage in football using team–season level data from six domestic leagues between the 2014–15 and 2024–25 seasons. Across multiple performance indicators, including home wins, home win rates, home points per match, and a composite Home Advantage Index, the results consistently indicate a **modest but statistically significant reduction in home advantage** following the introduction of VAR. While the magnitude of these changes is limited, their direction is stable across different metrics and analytical approaches.

Exploratory analysis shows that the decline in home advantage is not uniform across leagues and seasons, with noticeable variability driven by league-specific dynamics. Expected goals and referee-related metrics do not display strong or unidirectional shifts favoring home teams after VAR, suggesting that changes in officiating outcomes are more balanced rather than reversed. Additionally, the relationship between home advantage and final league ranking remains weak and largely unchanged between the pre- and post-VAR periods, indicating that VAR has not fundamentally altered how home performance translates into season outcomes.

Overall, the findings suggest that **home advantage persists in the VAR era but at a reduced intensity** compared to the pre-VAR period. The evidence supports the interpretation that VAR is associated with measurable adjustments in match outcomes at home, while broader structural factors influencing home advantage remain largely intact. These results contribute to a more nuanced understanding of VAR's role in modern football, emphasizing incremental change rather than a structural transformation of competitive balance.

11. Technology Stack and Tools

- Programming Language: Python
- pandas
- NumPy
- matplotlib
- seaborn
- scipy
- scikit-learn
- jupyter

12. Appendices

Appendix A: Project Timeline

- Project Proposal (31 October 2025)
- Data Collection & Exploratory Data Analysis (28 November 2025, extended to 30th)
- Machine Learning (2 January 2026)
- Final Submission (9 January 2026)

Appendix B: Data Dictionary

The final merged dataset contains 51 variables:

1. **id:** Unique team–season identifier.
2. **team:** Name of the football club.
3. **country:** Country in which the club competes.
4. **league:** Domestic league of the club.
5. **season_start:** Starting year of the football season.
6. **season_end:** Ending year of the football season.
7. **final_league_pos:** Final league standing of the team at the end of the season.
8. **home_matches:** Total number of matches played at home.
9. **home_wins:** Number of home matches won.
10. **home_draws:** Number of home matches drawn.

11. **home_losses:** Number of home matches lost.
12. **home_win_rate:** Proportion of home matches won.
13. **home_draw_rate:** Proportion of home matches drawn.
14. **home_loss_rate:** Proportion of home matches lost.
15. **home_points_per_match:** Average points earned per home match.
16. **home_goals_for:** Total goals scored at home.
17. **home_goals_for_per_90:** Average goals scored per 90 minutes at home.
18. **home_goals_against:** Total goals conceded at home.
19. **home_goals_against_per_90:** Average goals conceded per 90 minutes at home.
20. **home_xg:** Total expected goals (xG) generated at home.
21. **home_xg_avg:** Average expected goals per home match.
22. **home_penalties_won:** Number of penalties awarded in favor of the team at home.
23. **home_penalties_conceded:** Number of penalties conceded at home.
24. **home_fouls_won_avg:** Average fouls awarded in favor of the team at home per match.
25. **home_fouls_conceded_avg:** Average fouls conceded by the team at home per match.
26. **home_yellow_cards_won:** Average yellow cards received by opponents at home.
27. **home_yellow_cards_conceded:** Average yellow cards received by the team at home.
28. **home_red_cards_won:** Average red cards received by opponents at home.
29. **home_red_conceded:** Average red cards received by the team at home.
30. **away_matches:** Total number of matches played away.
31. **away_wins:** Number of away matches won.
32. **away_draws:** Number of away matches drawn.
33. **away_losses:** Number of away matches lost.
34. **away_win_rate:** Proportion of away matches won.
35. **away_draw_rate:** Proportion of away matches drawn.
36. **away_loss_rate:** Proportion of away matches lost.
37. **away_points_per_match:** Average points earned per away match.
38. **away_goals_for:** Total goals scored away.
39. **away_goals_for_per_90:** Average goals scored per 90 minutes away.
40. **away_goals_against:** Total goals conceded away.
41. **away_goals_against_per_90:** Average goals conceded per 90 minutes away.
42. **away_xg:** Total expected goals (xG) generated away.
43. **away_xg_avg:** Average expected goals per away match.
44. **away_penalties_won:** Number of penalties awarded in favor of the team away.
45. **away_penalties_conceded:** Number of penalties conceded away.
46. **away_fouls_won_avg:** Average fouls awarded in favor of the team away per match.

47. **away_fouls_conceded_avg:** Average fouls conceded by the team away per match.
48. **away_yellow_cards_won:** Average yellow cards received by opponents away.
49. **away_yellow_cards_conceded:** Average yellow cards received by the team away.
50. **away_red_cards_won:** Average red cards received by opponents away.
51. **away_red_cards_conceded:** Average red cards received by the team away.