Tactical Profiling & Betting Strategy in the EPL

Barkın Akel - 34223

DSA 210 Introduction to Data Science (Fall 2025-2026)

Abstract

This project investigates the efficiency of the sports betting market, specifically focusing on the English Premier League corner kick market. By challenging the Efficient Market Hypothesis (EMH), I explored whether tactical data can be used to identify structural inefficiencies in bookmaker odds. Using a pipeline that combines K-Means clustering for tactical profiling, a weighted predictive model, and a Poisson-based probabilistic betting strategy, we demonstrate that a machine learning approach can generate statistically significant returns. The final model achieved a 55.3% win rate in "Under" markets and a 52.7% win rate in "Over" markets, suggesting that tactical matchups are not fully priced into standard markets.

1. Introduction

1.1 Motivation

The sports betting market offers a clear way to test the Efficient Market Hypothesis (EMH). The EMH suggests that all public information is reflected in prices (odds), making it impossible to consistently generate excess returns. However, while major markets like Match Result (1X2) are highly efficient due to massive liquidity, markets like **Corner Kicks** often receive less attention. This project hypothesizes that bookmakers rely heavily on historical averages and simple form, potentially overlooking complex tactical interactions.

1.2 Project Scope

The scope is limited to:

- **League:** English Premier League (EPL), chosen for the availability of advanced tactical metrics.
- **Market:** Total Corners (Over/Under), chosen for its volatility and potential for tactical exploitation.
- **Period:** 2019-2024 seasons.

2. Data & Methodology

2.1 Data Sources

The analysis integrates data from two primary sources:

- **fbref.com:** Provided detailed match statistics, including Expected Goals (xG), Possession, and Defensive Actions.

- **totalcorner.com:** Provided historical opening and closing odds for corner markets, enabling realistic backtesting.

2.2 Preprocessing & Feature Engineering

Raw match data was aggregated and cleaned to ensure consistency across seasons. Key feature engineering steps included:

- **Tactical Metrics:** Integration of advanced metrics such as Expected Goals (xG), Possession, and Touches in Attacking Penalty Area to capture team playing styles beyond simple result-based statistics.
- **Bias Correction:** Statistical adjustment for Home/Away performance variance to normalize team data for neutral comparison.

2.3 Tactical Profiling (Unsupervised Learning)

To move beyond simple table standings, **K-Means Clustering** was applied to categorize teams into distinct tactical archetypes. Using Silhouette Analysis, an optimal $K = 6$ was selected, identifying styles such as *High Pressing*, *Deep Block*, and *Box Siege*.

2.4 Predictive Modeling

A weighted predictive model was developed to predict the total number of corners in a match. The prediction P_{total} is a weighted sum of three components:

1. **Tactical Matchup History (80%):** How these specific tactical clusters interact historically.
2. **Season Averages (15%):** The baseline performance of the teams involved.
3. **Recent Form (5%):** Short-term trends from the last 5 matches.

Weights were optimized via Grid Search to minimize the Mean Absolute Error (MAE) and Brier Score.

3. Exploratory Data Analysis & Hypothesis Testing

Before modeling, I validated that tactical styles actually influence corner outcomes.

3.1 Tactical Correlations

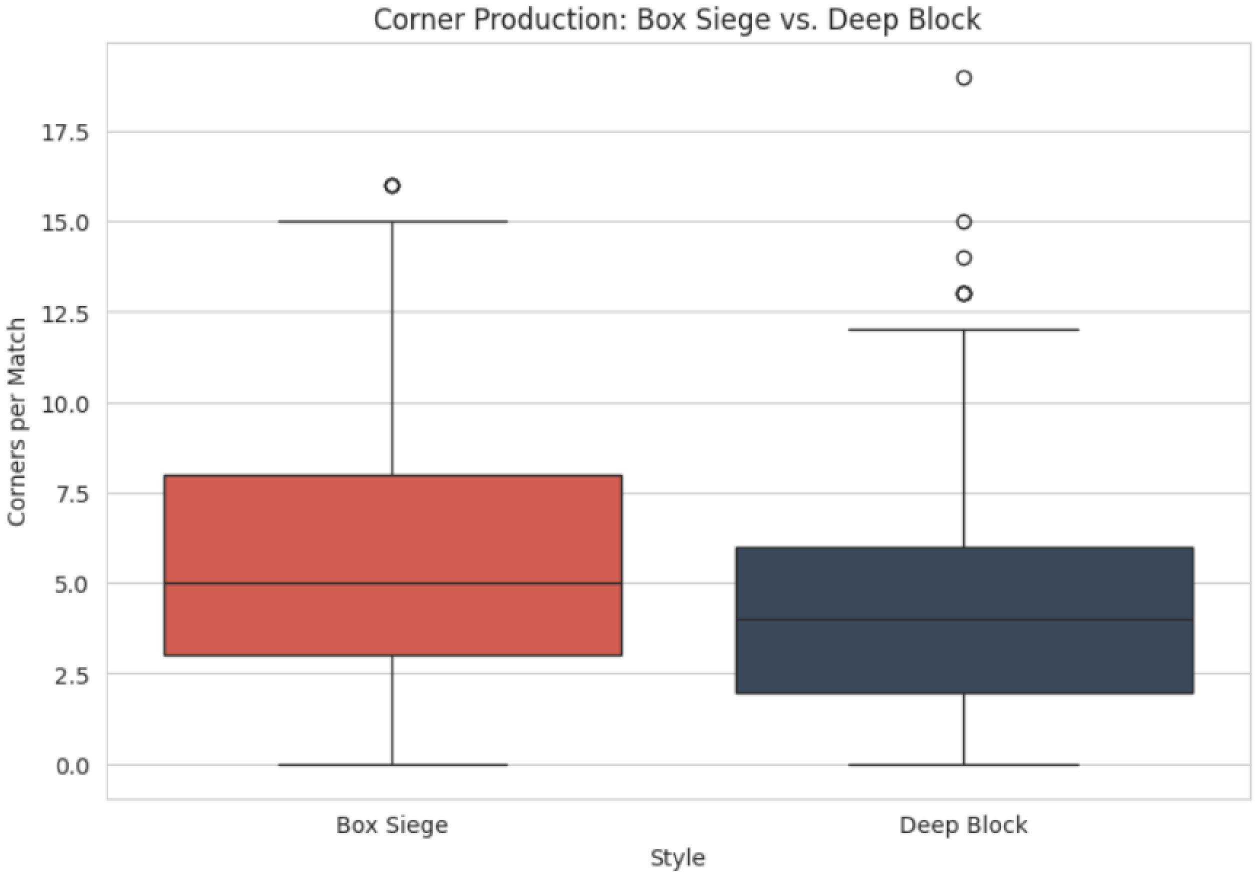

Figure 1: Distribution of corners per match for 'Box Siege' vs. 'Deep Block' teams.

A formal hypothesis test (T-Test) yielded conclusive results:

- **Box Siege Mean:** 5.83 Corners per match
- **Deep Block Mean:** 4.33 Corners per match
- **Significance:** The P-Value is **0.0000**, overwhelmingly rejecting the null hypothesis. This confirms that the "Box Siege" style structurally generates significantly more corners than the "Deep Block" style.

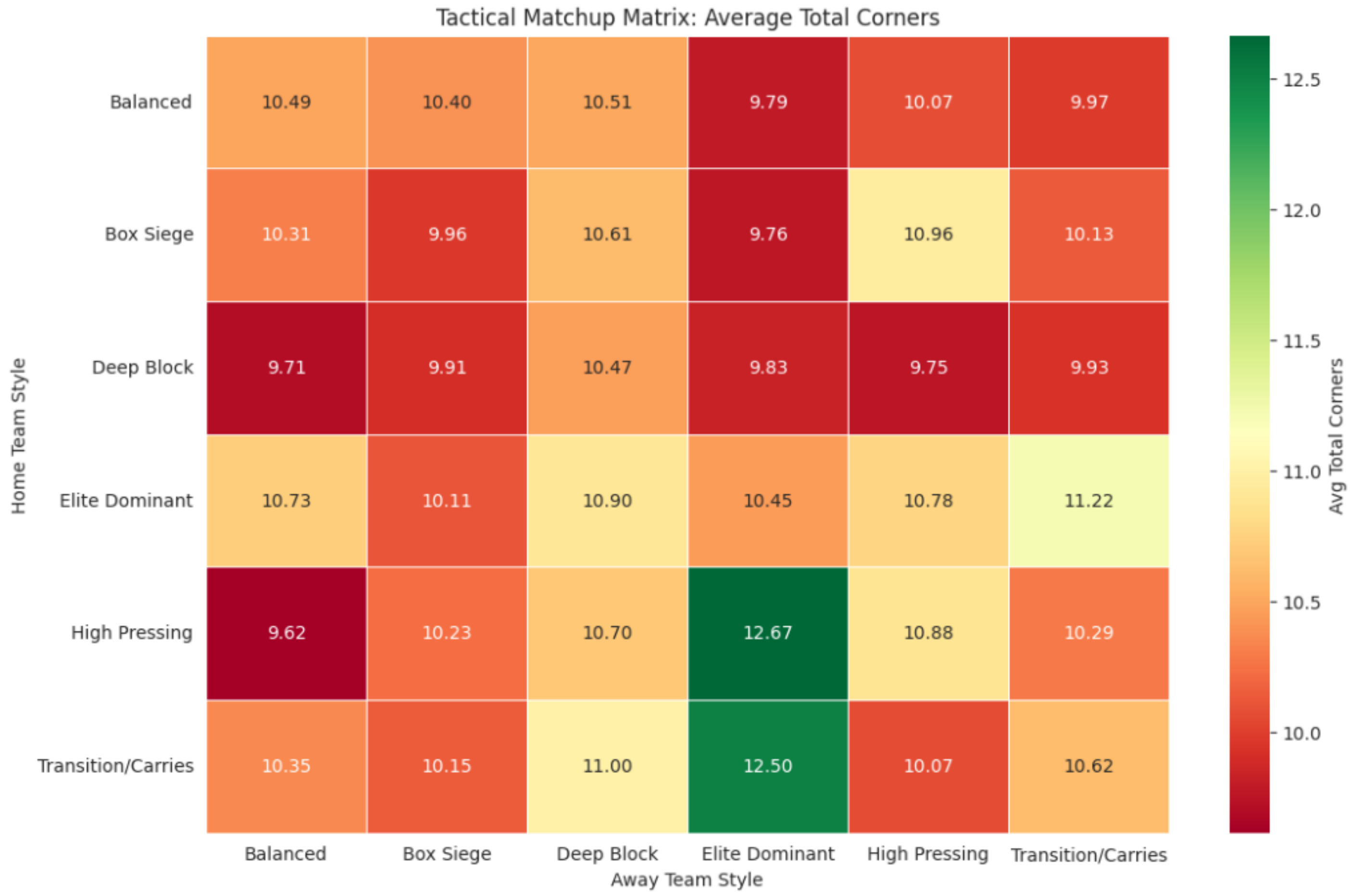

Figure 2: Tactical Matchup Matrix showing the average total corners for each Home vs. Away style combination.

As seen in **Figure 2**, the interaction between tactical styles produces distinct corner output patterns:

- The highest corner averages occur when **High Pressing** teams host **Elite Dominant** opponents (**12.67 avg**).
- The most frequent high-value matchup is **Elite Dominant** vs. **Deep Block**, which maintains a high average of **10.90 corners** across a sample of 84 matches which can be seen from the notebook.

4. Betting Strategy & Optimization

4.1 Probabilistic Valuation

The model's predicted total was converted into probabilities using the **Poisson Distribution**. This allowed for the calculation of "Fair Odds" for any given line (e.g., Over 9.5 Corners).

4.2 The Safety Score

To manage risk, I introduced a Safety Score:

$$\text{Safety Score} = \frac{|\text{Line}-\text{Predicted Total}|}{\text{Historical MAE}}$$

This Z-score-like metric quantifies the confidence of a prediction relative to the model's historical error margin.

4.3 Staking Strategy

An optimized staking plan was simulated using three parameters, with the final selection maximizing the risk-adjusted return:

- **Edge Power ($B = 1.0$):** A linear scaling approach was chosen, increasing stakes directly in proportion to the calculated edge.
- **Confidence Threshold ($T = 0.68$):** A strict cutoff requiring a high Safety Score to deploy maximum capital.
- **Volatility Penalty ($C = 0.4$):** A moderate penalty applied to teams with high performance variance to protect the bankroll.

To translate these theoretical scores into actionable bets, a **1-10 Unit Staking System** was implemented using a Scaler of 2.0 (the detailed steps are in the notebook.).

5. Model Performance & Profitability

Backtesting on historical data revealed profitable thresholds for market entry. Our optimization analysis confirmed a direct correlation between the Safety Score and betting success. As we increased the confidence threshold, the win rate improved, validating the Safety Score as a reliable filter for value.

Key Performance Metrics:

- **Under Markets:** A Safety Score threshold of **> 0.11** yielded a **55.3% Win Rate** across 338 matches.
- **Over Markets:** A Safety Score threshold of **> 0.22** yielded a **52.7% Win Rate** across 307 matches.

Financial Performance: Applying the optimized staking strategy to these signals resulted in a Return on Investment (ROI) of 12.84% and a total Net Profit of $10,441 over the 4-year backtesting period, based on a hypothetical Starting Bankroll of $10,000.

Profitability Insight: While a 55.3% win rate might seem modest, in sports betting, any win rate above the break-even point (typically ~52.4% for standard -110 odds) implies long-term profitability. The model successfully identified "Value Bets" which represent situations where the bookmaker's implied probability was significantly lower than the true probability of the outcome. The higher win rate in "Under" markets suggests that the market tends to overestimate the likelihood of high-corner events in specific tactical matchups.

6. Discussion

6.1 Limitations

- **Game State:** The current model uses pre-match expectations and aggregate stats. It does not account for in-game dynamics, such as a favorite scoring early and subsequently reducing their attacking output (and thus corners).
- **Data Availability:** The reliance on advanced metrics limits the strategy's transferability to lower-tier leagues where such data is hard to find.

6.2 Future Work

Future iterations will focus on:

- **Market Diversification:** Expanding the framework to cover **Corner Handicap** markets and general **Over/Under** markets (e.g., Total Goals) to broaden the strategy's scope.
- **Cross-League Validation:** Testing the model's robustness on other major European leagues like the Bundesliga and La Liga once reliable historical data is sourced.

7. Conclusion

This project successfully demonstrates that the Premier League corner market is not fully efficient. By utilizing unsupervised learning to define tactical profiles and identifying a fair value, we constructed a betting strategy that outperforms random chance. The results highlight the value of specific feature engineering over generic statistical approaches in sports analytics.

8. References

1. **fbref**, "Premier League Stats," https://fbref.com.
2. **totalcorner**, "Corner Odds Database," https://www.totalcorner.com/.
3. **Python Libraries:** `pandas`, `numpy`, `scikit-learn`, `scipy`, `matplotlib`, `seaborn`.

9. AI Disclosure

AI Assistance: Generative AI tools were utilized extensively throughout the entire development process. This includes assistance with writing data scraping scripts, implementing complex logic, debugging code, optimizing model parameters, formatting. While AI accelerated the coding process, conceptual frameworks, the core ideas and the analytical decisions are entirely my own.