LEVERAGING THE TRANSITIVE PROPERTY IN SPORTS: A BASIC MODEL FOR PREDICTION AND BETTING

ALI MOHAMMADI

Abstract. This paper introduces a simple yet effective model for predicting the outcomes of sports matches by comparing the recent performance of competing teams against shared opponents. The model uses the resulting quantitative features, which depend only on past scores of teams, within a random forest machine learning framework to generate predictions based on historical data. Applied to 4431 Major League Baseball matches from the 2023 and 2024 seasons, the model achieved a 1.95% return on investment while maintaining probability calibration comparable to professional bookmakers, as evidenced by similar accuracy, log-loss and Brier scores. By combining these predictions with a disciplined betting strategy grounded in the Kelly criterion, the model demonstrates its potential for both accurate forecasting and profitable betting in a real-world context.

1. Introduction

Predicting the outcomes of sports matches has long been an area of interest for researchers, analysts, and betting enthusiasts alike. This paper proposes a model for match outcome prediction based on the transitive property of sports performance, quantified through the calculation of the *projected score difference (PSD)*. By analysing the relative performance of competing teams against common opponents, in terms of the difference in their corresponding score differences, the PSD provides a systematic measure of superiority, which serves as a foundation for both predictive modeling and betting strategies.

The methodology was tested on Major League Baseball data spanning the 2021–2024 seasons. Matches from 2021 were used exclusively to calculate PSD statistics for the 2022 season, ensuring that all predictions rely solely on prior information. The model incorporates PSD statistics to train a machine learning framework that predicts match outcomes and assigns win probabilities for both teams. A disciplined betting strategy, based on the Kelly criterion [1], is then applied to evaluate the model's profitability.

The key contributions of this work include:

- The introduction of the PSD metric as a robust and interpretable feature for match outcome prediction.
- The integration of PSD statistics into a machine learning framework for predictive modelling.
- A comprehensive evaluation of model performance, including accuracy, probability calibration, and profitability in a real-world betting context.
- Insights into the model's strengths and limitations through ROI analysis across odds ranges and calibration curves.

Correspondence: ali.mohammadi.np@gmail.com.

By demonstrating the profitability and competitive accuracy of the proposed approach, this work highlights its potential as a practical tool for forecasting and sports betting. The results validate the importance of incorporating interpretable features, such as the PSD, into machine learning models and underline the value of systematic betting strategies for maximising returns in the competitive domain of sports prediction. We note that a fundamentally similar approach was used in a predictive model for tennis matches in [2], although unlike ours, it relies heavily on domain-specific knowledge of the sport.

2. Methodology

The methodology employed in this study involves calculating the projected score difference (PSD), which serves as a key feature for predicting match outcomes. The PSD leverages the transitive property in sports: if player A performs better than player B against a common opponent, then player A is inferred to have a relative advantage over player B. It is thus quite natural to use the difference of the corresponding score differences to quantify the superiority of one player (or team) over another.

2.1. **Projected score difference calculation.** The PSD captures the relative performance of two players against recent common opponents, accounting for scenarios where they compete as either home or away players. The calculation proceeds as follows.
1. Identification of relevant matches. A predefined search window (e.g., twelve months prior to the match) is used to filter matches involving both competing players and their recent common opponents. These matches are further categorised based on whether the competing players played as home or away players.
2. Calculation of projected score difference. For each common opponent X, consider matches, within the predefined search window, where both players A and B played X and write

$$(S_{AX} - S_{XA}) - (S_{BX} - S_{XB}), \tag{1}$$

for the PSD of A and B (via X), where S_{AX} is the score of player A against X, and S_{XA} is the score of player X in that match. Similarly, S_{BX} and S_{XB} are the corresponding values for player B. The usefulness of PSD relies on the competing players performing under similar conditions, thus we distinguish between the quantities PSD_{home} and PSD_{away} by calculating instances of (1) where both A and B played as home and away players, respectively.

In the case where there were multiple matches between either of the competing players and a common opponent within the search window, let N_A represent the number of matches between A and X, and N_B the number of matches between B and X. We then calculate the PSD by

$$N_A^{-1} \cdot \sum_{i=1}^{N_A} \left(S_{AX}^{(i)} - S_{XA}^{(i)} \right) - N_B^{-1} \cdot \sum_{j=1}^{N_B} \left(S_{BX}^{(j)} - S_{XB}^{(j)} \right),$$

where $S_{AX}^{(i)}$ denotes the score of player A in their i-th match against opponent X, and $S_{BX}^{(j)}$ denotes the score of player B in their j-th match against X. This adjustment accounts for the differing number of matches each player may have had with the common opponent, ensuring a balanced and accurate measure of relative performance.

3. Aggregation across common opponents. As key features of the model, we calculate the means of the values of PSD_{home} and PSD_{away}, as well as their standard deviation and count, to assess variability and reliability.

2.2. **Integration into machine learning framework.** The PSD statistics (mean and standard deviation) are integrated as features into a machine learning framework, specifically a *random forest classifier*, to predict match outcomes. The process involves:
1. Feature engineering. For each match, the following features are extracted:

- `psd_home_mean`, `psd_away_mean`: Measures of projected superiority as home and away players.
- `psd_home_std`, `psd_away_std`: Variability in projected superiority.

2. Target variable. The target variable, `result`, represents the actual match outcome, encoded as 1 for a home player win and 0 for an away player win.
3. Model training. The random forest classifier is trained on historical match data, using PSD features to predict match outcomes.
4. Probability estimation. The model outputs probabilities for home and away wins, which are subsequently used to inform betting strategies.

3. Betting Strategy

The probabilities outputted by the random forest classifier are combined with bookmaker odds to inform a systematic betting strategy. The Kelly criterion is used to determine optimal bet sizes, assuming a fixed capital of one dollar per match.

3.1. **Kelly criterion for betting.** The Kelly criterion provides a mathematical framework to maximise capital growth while minimising risk. For a given probability p and bookmaker odds b, the optimal bet size f is calculated by:

$$f = \max\left(0, p - \frac{1-p}{b-1}\right).$$

We recall that f here technically represents the fraction of one's capital to be wagered, which we assume to be fixed and equal to one dollar in our analysis.

3.2. **Implementation.** For each match, the following steps are undertaken:

(1) Use the predicted probabilities of the model for home and away wins.
(2) Compute the optimal bet sizes for both outcomes using the Kelly criterion.
(3) Calculate the profit (or loss) based on the result of the match.

4. Results and Discussion

This section evaluates the performance of the proposed model and betting strategy using various metrics and comparisons. The results are derived from the backtesting[1] of the model on historical data.

[1]source code for this analysis may be accessed at https://github.com/Ali-m89/Sports_Prediction_and_Betting_Model.

4.1. **Data preparation and experimental setup.** The analysis was conducted using data[2] from Major League Baseball matches for the seasons 2021–2024, containing information on the dates, scores and home/away win odds, representing market averages among a large number of bookmakers. The year 2021 was included only to calculate the projected score difference statistics for matches in 2022. We also note that since matches in 2020 were sparse due to the COVID-19 outbreak, it was natural to set the cut-off for our analysis at 2021. Thus, the core dataset comprises matches from 2022 to 2024, with projected score difference statistics calculated using a search length of 12 months.

The dataset was split into training and test sets based on a timewise split. Specifically, the most recent 60 percent of the matches (spanning most of 2023 and 2024) were used as the test set, while the earlier 40 percent were designated as the training set.

4.2. **Model performance.** The accuracy of the model was measured and compared with that of the bookmakers. The results are summarised as follows:

- Model Accuracy: 55.81%
- Bookmaker Accuracy: 57.17%
- Home Win Percentage (Baseline Accuracy): 52.06%

Although the model's accuracy is slightly lower than that of the bookmakers, it remains competitive and outperforms the baseline home win percentage.

4.3. **Probability calibration.** The *Brier score* and *log-loss* are key metrics used to evaluate the accuracy of predicted probabilities. The Brier score measures the mean squared difference between predicted probabilities and actual outcomes, defined as

$$\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2,$$

where N is the number of predictions, p_i is the predicted probability of the event occurring, and o_i is the actual outcome (1 for a win, 0 for a loss). A lower Brier score indicates better-calibrated predictions.

The *log-loss* quantifies the negative log-likelihood of the predicted probabilities, penalising overconfident incorrect predictions more heavily. It is defined as

$$-\frac{1}{N}\sum_{i=1}^{N}\left[o_i \log(p_i) + (1 - o_i)\log(1 - p_i)\right].$$

For the proposed model, the Brier score is 0.2452, compared to the bookmakers' score of 0.2419. Additionally, the model's log-loss is 0.6834, while the bookmakers' is 0.6766.

These metrics indicate that the model's probability predictions are well-calibrated and competitive with those of the bookmakers.

4.4. **Profitability and ROI.** The profitability of the betting strategy was evaluated, yielding the following key statistics:

- Total Profit: \$7.29
- Return on Investment (ROI): 1.95%
- Total Matches Analysed: 4431

[2]data was sourced from https://www.oddsportal.com.

- Total Bets Placed: 374.21
- Average Bet Size: $0.084
- Percentage of Profitable Bets: 37.62%

Figure 1 shows the cumulative profit of the model for the chronologically ordered matches.

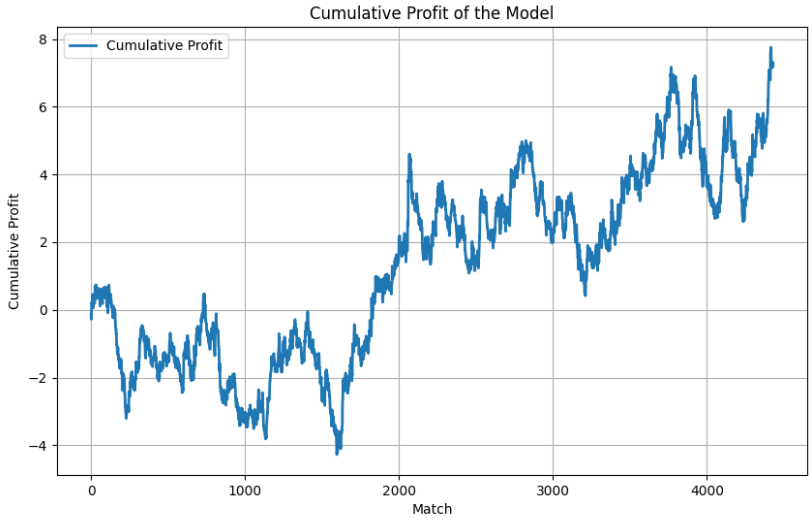

FIGURE 1. Cumulative profit of betting on the 4431 matches

The results demonstrate that the model is profitable over a large dataset, achieving a positive ROI even when evaluated against professional bookmaker odds. Notably, the profitability analysis in this study relied on average consensus odds provided by bookmakers, chosen to facilitate the simultaneous evaluation of their prediction accuracy. However, the model's ROI could likely be improved through odds shopping—selecting the best available odds across multiple bookmakers for each match. Future work could explore this approach to provide a more accurate reflection of the model's true profitability.

4.5. **ROI by odds range.** To understand profitability across different odds ranges, the ROI was analysed as follows

Odds Range	Total Bets	Total Profit	ROI (%)
1.0–1.5	49.04	-3.08	-6.27
1.5–2.0	209.88	5.52	2.63
2.0–3.0	108.46	4.42	4.08
3.0–5.0	6.84	0.43	6.30

TABLE 1. ROI by odds range

The results reveal that profitability improves with higher odds ranges. In the 1.0–1.5 range, the strategy incurs a loss, highlighting the difficulty of finding value in low-odds bets. Conversely, the model achieves its highest ROI (6.30%) in the 3.0–5.0 range, indicating that the model excels in identifying value in higher-risk scenarios.

4.6. **Calibration curve.** Figure 2 illustrates the calibration curve for the model, comparing predicted probabilities with actual outcomes. The curve demonstrates that the model is generally well-calibrated, with predictions aligning closely with observed win rates. Minor deviations are observed in extreme probability ranges, suggesting potential areas for improvement in probability estimation.

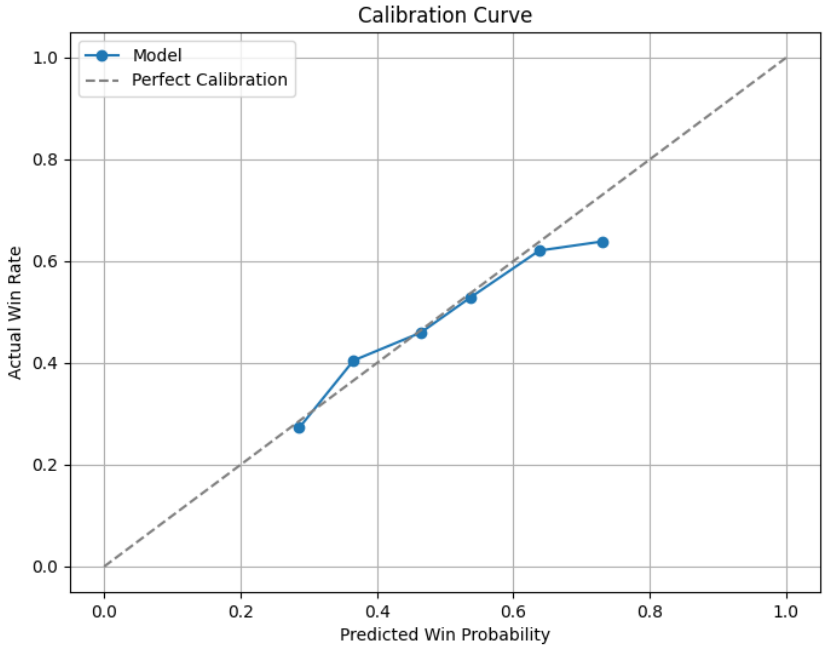

FIGURE 2. Calibration curve for model predictions

4.7. **Discussion.** The proposed model demonstrates profitability and robust probability estimation in a competitive environment. While its accuracy is slightly lower than that of bookmakers, the positive ROI suggests that it effectively identifies value bets. The ROI breakdown by odds range indicates that the model performs best in medium to high odds scenarios, where bookmakers may overestimate the risk of less favored outcomes.

The calibration curve highlights the reliability of the model's probability predictions, though minor improvements in extreme ranges could further enhance performance. Future work may focus on refining probability calibration and exploring alternative betting strategies to maximise ROI.

Overall, the results validate the efficacy of the model and its potential as a tool for both forecasting and profitable betting.

REFERENCES

[1] J. L. Kelly, A New Interpretation of Information Rate. *Bell System Technical Journal*, **35(4)** (1956), 917–926. https://doi.org/10.1002/j.1538-7305.1956.tb03809.x

[2] W. J. Knottenbelt, D. Spanias and A. M. Madurska, A common-opponent stochastic model for predicting the outcome of professional tennis matches, *Computers and Mathematics with Applications* **64(12)** (2012), 3820–3827. https://doi.org/10.1016/j.camwa.2012.03.005