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

ALI MOHAMMADI

Abstract. This paper outlines implementation of 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 presents a model for match outcome prediction based on the transitive property of sports performance, quantified through the calculation of the *projected score difference (PSD)*. The PSD is defined as the average difference in score differentials between two competing teams (or players) against recent common opponents. It captures the relative strength of the teams by evaluating how they have performed in similar matchups. That is, the model essentially relies on the assumption that if team A performed better than X (by α) and B performed better than X (by β), then A should perform better than B (by $\alpha - \beta$). Of course, to improve the practicality of such an approach, one also needs to ensure that the three events take place under similar circumstances, within a short period of time, and repeat this analysis for a large number of such X.

The model incorporates PSD statistics to train a machine learning framework that predicts match outcomes and assigns win probabilities to both teams. A disciplined betting strategy, based on the Kelly criterion [1], is then applied to evaluate the model's profitability.

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.

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 key contributions of this work include:

- The introduction of the PSD metric as a robust and interpretable feature for match outcome prediction.

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

- 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 a calibration curve.

2. Methodology

We outline the two key steps of the proposed method.

2.1. **Projected score difference calculation.** The PSD captures the relative performance of two teams against recent common opponents, accounting for scenarios where they compete as either home or away teams. 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 teams and their recent common opponents. These matches are further categorised based on whether the competing teams played as home or away teams.
2. Calculation of projected score difference. For each common opponent X, consider matches, within the predefined search window, where both teams 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 team A against X, and S_{XA} is the score of team X in that match. Similarly, S_{BX} and S_{XB} are the corresponding values for team B. The usefulness of PSD relies on the competing teams 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 teams, respectively.

In the case where there were multiple matches between either of the competing teams 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 team A in their i-th match against opponent X, and $S_{BX}^{(j)}$ denotes the score of team B in their j-th match against X.
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 teams.
- `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 team win and 0 for an away team 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 have assumed to be fixed and equal to one dollar.

3.2. Implementation.

For each match, we compute the optimal bet sizes for both outcomes using the Kelly criterion, which requires the offered odds and our predicted probabilities of both outcomes. We then 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 backtesting[1] the model on historical data.

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[3]. 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.

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

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

[3] bookmakers used in this study: *10bet, 10x10bet, 1xBet, 22Bet, 888sport, Alphabet, bet-at-home, bet365, Betfair, BetInAsia, Betsafe, Betsson, Betway, Coins.game, GGBET, NordicBet, Pinnacle, Sportium.es, Unibet, VOBET, Vulkan Bet, William Hill.*

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 Consensus Accuracy: 57.17%
- Home Win Percentage (Baseline Accuracy): 52.06%

Although the model's accuracy is slightly lower than that of the bookmakers[4], 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
- Total Bets Placed: \$374.21
- Return on Investment (ROI): 1.95%
- Total Matches Considered: 4431
- Number of Bets: 3632
- Number of Profitable Bets: 1667
- Percentage of Profitable Bets: 45.90%
- Average Bet Size: \$0.103

Note that for 799 of the 4431 available matches, the model did not place a bet on either outcome. This happens exactly when bookmakers' consensus aligns tightly with the model's prediction and thus Kelly's criterion does not find value in betting on either side (largely also because of the "house edge").

Further note that it is not only possible, but also very common for the model to pick a team as a potential winner (determined when a team is assigned a win probability strictly

[4]It is widely accepted that the bookmaker consensus offers the most accurate predictions for competitive sports. This is not solely due to bookmakers' superior modeling techniques but also because of the 'wisdom of crowds' effect, where market forces and financial incentives drive odds toward optimal probabilities.

higher than 0.5) but place a bet on the other team (determined by the value of the bet as dictated by Kelly's criterion). This also highlights significance of metrics like Brier score and log-loss in evaluating a model's efficacy, rather than just the usual notion of accuracy.

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

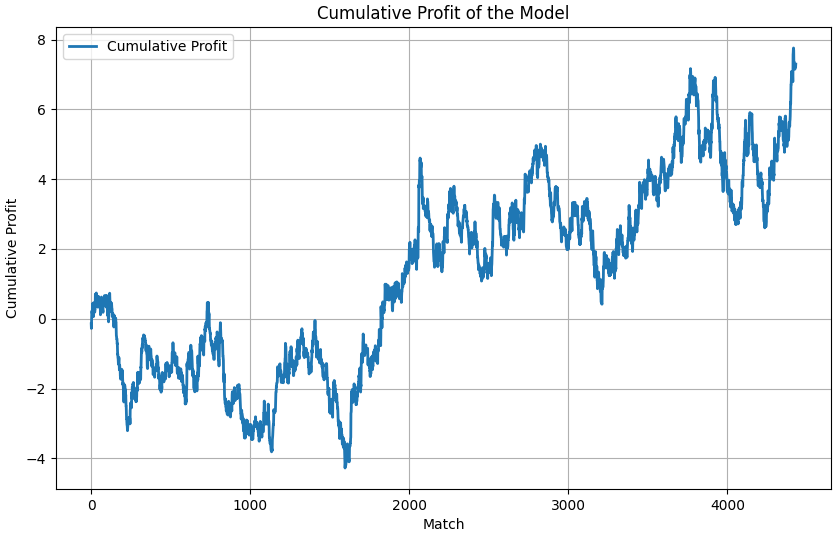

FIGURE 1. Cumulative profit of betting on the 4431 matches

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.

It is also important to point out that the ROI exhibited conspicuous variability, with large dips and an unpredictable pattern throughout the evaluation period. This highlights that, while the model demonstrated long-term profitability on average, its performance over extended periods can exhibit significant downturns, which is concerning. Such an unstable behaviour is of course to be expected from such an overly simple model.

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 improved with higher odds ranges. In the 1.0–1.5 range, the strategy incured a loss, possibly highlighting the model's inability to make accurate predictions for highly uneven matches. Conversely, and somewhat surprisingly, the model achieved its highest ROI (6.30%) in the 3.0–5.0 range. One likely needs to test the model on a much larger dataset to better understand such subtleties in its behaviour.

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.

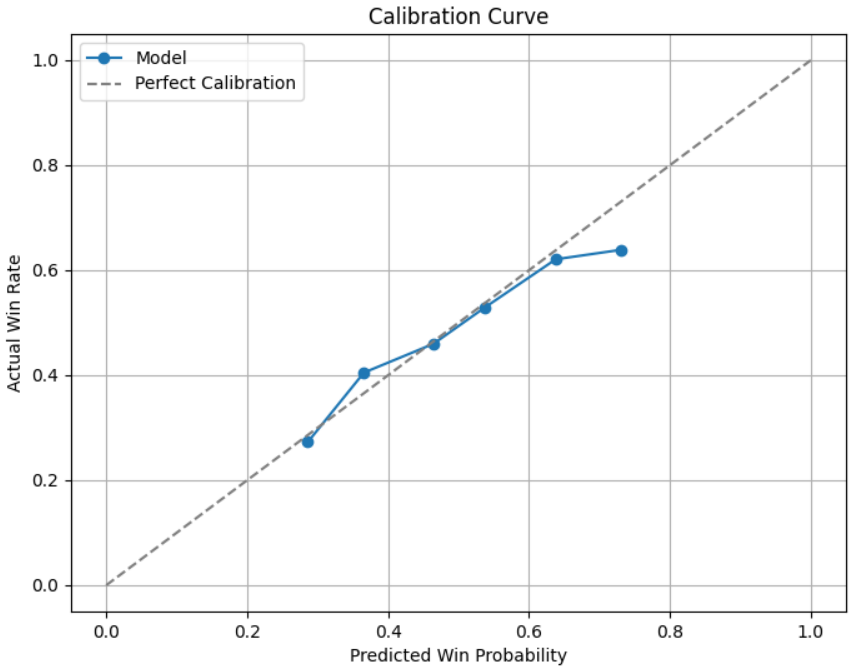

FIGURE 2. Calibration curve for model predictions

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.

Moreover, it is worth noting that extreme predictions occur in a domain where external factors—such as player fatigue, injuries, motivation, weather conditions, lineup changes, and even match attendance—become disproportionately influential. These factors introduce

significant variance, especially when the expected margin is large. Since our model does not account for such contextual variables, its predictions may become disproportionately less reliable in extreme cases.

This apparent inherent drop in the model's accuracy for predicting uneven matchups also warrants additional care in using the Kelly criterion. While the Kelly criterion provides an optimal staking strategy under correct probability estimates, it becomes highly sensitive to errors when the estimated probability of an outcome is extreme (i.e., very close to 0 or 1). For instance, overestimation of the winning probability of a heavy favorite leads to disproportionately large bets, increasing risk exposure.

Recall, the Kelly fraction is given by

$$f = p - \frac{1-p}{b-1}.$$

To understand its sensitivity to errors in p, we differentiate f with respect to p

$$\frac{df}{dp} = 1 + \frac{1}{b-1}.$$

This expression shows that the bet size f grows faster with p when b is small (i.e., for heavy favorites with low odds). In other words, when b is large (betting on underdogs), the fraction $\frac{1}{b-1}$ is small, meaning f changes gradually with p. When b is small (betting on favorites), the fraction $\frac{1}{b-1}$ is large, meaning even small errors in p cause large swings in f.

To quantify this effect, consider a scenario where the true probability is $p = 0.9$ but is overestimated by $\Delta p = 0.05$. The change in bet size is

$$\Delta f \approx \left(1 + \frac{1}{b-1}\right) \Delta p.$$

For a strong favorite, at fair odds, with $b = 1.\bar{1}$, we get

$$\Delta f \approx \left(1 + \frac{1}{0.\bar{1}}\right) \times 0.05 \approx (1+9) \times 0.05 = 0.5.$$

That is, a mere 5% overestimation in p causes a 50% increase in bet size. If the true probability was only 50% ($p = 0.5$), the same 5% overestimation would cause a 10% increase at fair odds ($b = 2$).

This explains why Kelly betting on heavy favorites is disproportionately sensitive to probability misestimation—the closer p is to 1, and the lower b, the greater the impact of a small error in p. This problem is exacerbated in real-world settings where external factors introduce variance that the model does not account for, making these estimates inherently uncertain.

A practical mitigation strategy is to use fractional Kelly betting (e.g., half-Kelly) to reduce bet volatility in extreme probability cases. Alternatively, implementing uncertainty-aware probability models (e.g., Bayesian approaches) can help prevent overconfident staking in cases where estimates are likely to be noisy.

5. Future Work

The model presented in this paper offers a solid foundation for sports match prediction and betting strategies based on the projected score difference (PSD). However, several promising directions remain for further exploration and enhancement:

5.1. **PSD as a feature in more nuanced models.** While the PSD metric has proven useful in predicting match outcomes, it could be further refined and incorporated into more sophisticated models. For instance, machine learning approaches beyond random forests, such as deep learning or gradient boosting, could be explored to extract more intricate patterns from PSD statistics.

Moreover, the current model relies primarily on PSD as a single performance indicator. Future work could involve incorporating additional, more subtle features to capture nuances in team performance. These could include:

- Contextual factors such as fatigue, injuries, or lineup changes that influence match outcomes.
- Surface- or location-specific performance metrics (e.g., a team's strength on artificial turf vs. natural grass).
- Advanced statistical features such as momentum indicators, clustering of similar teams, or interaction effects between features.

By enriching the feature space with these elements, the model could achieve improved predictive accuracy and robustness.

5.2. **Transitive models in other sports.** The transitive approach to sports prediction, as demonstrated in this work, could be adapted and validated for other sports beyond Major League Baseball. Sports with different scoring structures, such as tennis, basketball, or football (soccer), may exhibit varying degrees of transitivity, influencing the effectiveness of the PSD-based methodology.

5.3. **Weighted PSD calculation.** One potential refinement of the model is to introduce a weighting scheme in the calculation of PSD. For instance, matches that resulted in significant deviations from the model's own predictions could be assigned lower weights in future calculations, thereby reducing the influence of highly uncharacteristic outcomes. This could improve overall model calibration and robustness.

5.4. **Optimising the search window for PSD.** Currently, the model utilises a fixed 12-month window to calculate PSD statistics. However, it remains an open question whether a shorter (e.g., 6-month) or longer (e.g., 24-month) window might offer improved predictive power. A systematic analysis could identify the optimal trade-off between recency and sample size, determining where predictive degradation begins.

5.5. **Applications beyond sports betting.** Although the primary focus of this study has been sports betting, the transitive approach and PSD metric could have broader applications. For example, this model could be adapted for evaluating the relative playing strengths of AI-driven game agents in board games or video games. For instance, in the same context as [3]. Instead of running exhaustive pairwise matchups, a PSD-based approach could efficiently infer relative strengths, reducing computational overhead.

These directions represent natural extensions of the current work and could significantly enhance both the accuracy and applicability of the model in various domains.

Acknowledgements

The author thanks Edwige Elysee for helpful comments.

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

[3] D. J.N.J. Soemers, G. Bams, M. Persoon, M. Rietjens, D. Sladić, S. Stefanov, K. Driessens, and M. H.M. Winands, *Towards a Characterisation of Monte-Carlo Tree Search Performance in Different Games*, *arXiv preprint*, arXiv:2406.09242, 2024. Available: https://arxiv.org/abs/2406.09242