Wharton Sports Analytics Journal

Multi-Model Consensus: an Ensemble Approach to Basketball Predictions

Shivam Gupta, '26
Lambert High School, GA, USA

Advisor: Stephanie Beaulieu
Lambert High School, GA, USA

Abstract

We present an ensemble approach to NCAA women's basketball prediction, combining four independent models to produce robust regional team rankings and win probabilities. Our primary method was an Elo rating system with basketball-specific considerations including home-court advantage, rest days, and travel distance. We verified its accuracy and reliability by cross-validating with three additional models: Dean Oliver's Four Factors, temporal logistic regression, and a network-based PageRank model. Using over 5,300 Division I games, our models provided a consensus, exhibiting the Elo rating system as the most conservative and reliable. Although our predicted rankings and win probabilities solely relied on our modified Elo rating model, the other three played a crucial role informing our final decision with data and statistics.

Introduction & Background

Basketball is a dynamic sport where outcomes are shaped not just by player skill, but by a combination of contextual and situational factors. Over recent years, the field of sports analytics has moved beyond basic win-loss records and point differentials, with advanced statistical models now capturing the influence of home-court advantage, rest, travel fatigue, and even event momentum.

Historically, many ranking systems relied on simplistic metrics like win-loss or margin of victory, which often overlooked elements like opponent strength or schedule. With the rise of data science in sports, the Elo rating system became a popular tool due to its adaptive updating and intuitive baseline structure. Elo's popularity in sports has been largely thanks to its success in chess and its subsequent adaptations for football, basketball, and other team sports.

However, Elo is not a panacea. While effective at capturing team strength evolution, traditional Elo rating systems can suffer from issues like long-term inflation or insensitivity to contextual factors of games. Other popular approaches, such as Dean Oliver's Four Factors, break down games into four values: shooting efficiency, turnovers, rebounding, and free-throw rate. These models offer explainability but may not fully capture team momentum or the impact of schedule difficulty.

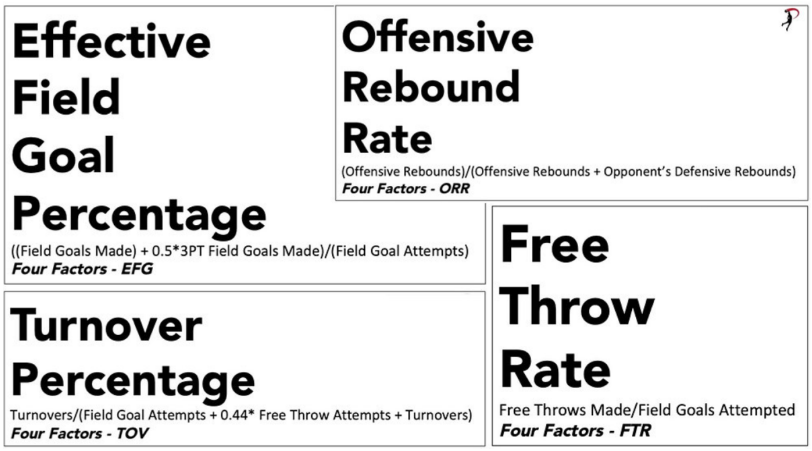

Figure 1: Dean Oliver's Four Factors

Recently, machine learning methods and network-based approaches like PageRank have entered the sports analytics domain. Temporal logistic regression can model how teams change over a season, leveraging rolling averages and recent performance trends. PageRank, borrowing ideas from network science and web search, can reflect the quality of a team's opponents and the interconnected structure of schedules.

Our goal was to develop a robust, justifiable method for ranking NCAA women's basketball teams and predicting tournament matchups by leveraging the strengths of these varied techniques. We designed and implemented four independent models-Elo, Four Factors, Temporal Logistic, and PageRank-then cross-validated our results to build consensus and confidence in our predictions. We found that our adapted Elo methodology, enhanced by contextual game factors and cross-validation, produced the most reliable results, but not in isolation: the interplay with our other models was crucial in shaping and validating our final rankings and probabilities.

Guiding Principles

Our research methodology was founded on three core principles that shaped every aspect of our analytical approach.

Explainability: We prioritized models and techniques that could provide clear, intuitive explanations for predictions. Each of our four models was chosen specifically for its ability to offer transparent insights. The Elo system provides straightforward rating updates based on game outcomes, Four Factors analysis breaks down performance into fundamental basketball components, logistic regression offers interpretable coefficients, and PageRank reveals network-based team strength relationships. This focus on explainability ensures that our predictions come with understandable reasoning.

Reproducibility: We wanted to ensure that our results and analytical process could be used again in the future and verified to be accurate. We carefully documented our data preprocessing steps, feature engineering decisions, and model parameters. All hyperparameters were explicitly stated, and our validation methodology was designed to prevent data leakage. This systematic approach allows our work to be replicated and our findings to be validated.

Robustness: Rather than relying on a single modeling approach, we deliberately constructed an ensemble of diverse methodologies to ensure robust predictions across different scenarios. This multi-model approach guards against the systematic biases inherent in any individual technique while providing confidence intervals through model agreement analysis. Our cross-validation framework systematically identifies games where models diverge, allowing us to flag uncertain predictions and understand the sources of disagreement.

Methods

We were provided a comprehensive dataset of 5,327 NCAA Division I women's basketball games from the 2021–22 season. Each row records a single team's box-score statistics and contextual features: field goals (2- and 3-point), free throws, rebounds (offensive and defensive), assists, turnovers, personal fouls, plus metadata for game date, home/away status, rest days, and travel distance. Region assignments for all teams and a separate file listing East region tournament matchups were merged in to support our two-phase analysis.

Data Preparation

To ensure data quality and consistency, we first filtered out any non-Division I contests using the provided notD1_incomplete flag. We then imputed missing rest-day values with the regional median (3 days) and zero-filled rare events (e.g., technical fouls). All games were sorted chronologically by game_date to maintain temporal integrity for models that rely on sequential updates.

We created several derived variables essential for game-contextual modeling:

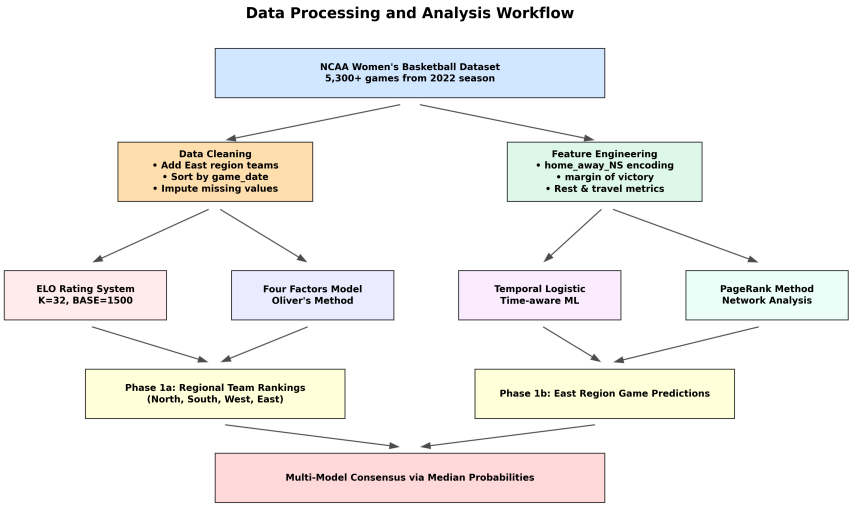

Figure 2: Data Preparation Workflow

- **Home-court indicator**: +1 for home, –1 for away, 0 for neutral venues.
- **Modified margin-of-victory multiplier**:

$$\text{MOV}_{\text{mult}} = \frac{(|\text{point\_diff}|)^{0.8}}{7.5 + 0.006|\text{elo\_diff}|}$$

  This formula dampens the effect of blowouts while still rewarding dominant wins.

- **Rest differential**:

$$\text{rest\_days\_Home} - \text{rest\_days\_Away}$$

- **Travel fatigue penalty**:

$$\log_{1p}(\text{travel\_dist\_Away})/10$$

  scaled to match Elo units.

Software and Tools

Our analysis was implemented in Python 3.9. We used **pandas** for data cleaning and transformation, **NumPy** for efficient numerical operations, **SciPy** for linear-algebra routines in our PageRank model, **scikit-learn** for temporal logistic regression and model validation, and **Matplotlib**/**Seaborn** for all visualizations. ChatGPT assisted with code debugging and repetitive formatting; however, all methodological decisions and modeling procedures were entirely devised by our team.

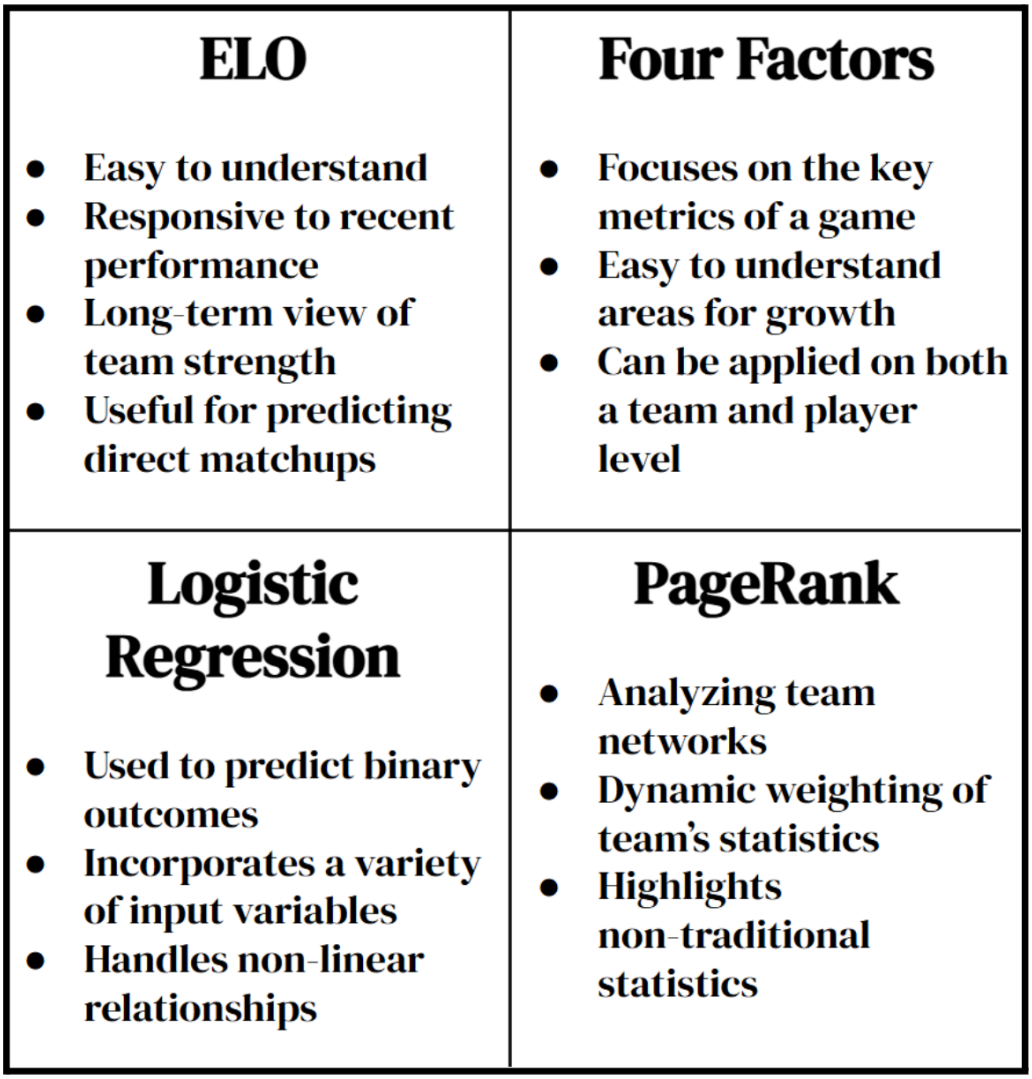

Figure 3: Our Ensemble Approach Comprised of Four Independent Models

Code Repository

Our complete codebase—including data preprocessing, model implementations, and visualization scripts—is publicly available on GitHub at `https://github.com/shivamCode0/whsdsc-2025`.

Modeling Approaches

We built four independent models, each capturing different aspects of team performance:

1. Elo Rating System (Primary)

The Elo rating system, originally developed for chess, provides a dynamic measure of team strength that updates after each game based on expected versus actual outcomes. Our

implementation extends the traditional Elo framework with basketball-specific modifications to account for margin of victory, contextual factors, and rating inflation.

Core Algorithm: Each team begins with an initial rating of 1500 points. For each game between teams A and B, we first compute home-adjusted ratings by adding a 70-point home court advantage:

$$R'_A = R_A + 70 \times \mathbb{I}_{\text{home A}}, \quad R'_B = R_B + 70 \times \mathbb{I}_{\text{home B}}$$

The expected win probability for team A follows the standard logistic formulation:

$$E_A = \frac{1}{1 + 10^{-(R'_A - R'_B)/400}}$$

After observing the actual outcome S_A (1 for win, 0 for loss), we calculate the rating update using a K-factor of 32, modified by our margin-of-victory multiplier:

$$\Delta R_A = K \times \text{MOV}_{\text{mult}} \times (S_A - E_A)$$

To prevent long-term rating inflation, we enforce a zero-sum constraint where $\Delta R_B = -\Delta R_A$.

Contextual Enhancements: For tournament predictions, we incorporate additional contextual adjustments that reflect the unique circumstances of postseason play:

- **Rest advantage**: +7.2 points per day of rest differential
- **Travel penalty**: -2.8 points per 300 miles of travel distance
- **Neutral site adjustment**: Removes home court advantage when applicable

Algorithmic Pseudocode:

```
FOR each game in chronological order:
    home_elo = team_A_rating + (70 if home else 0)
    away_elo = team_B_rating + (70 if home else 0)

    expected_A = 1 / (1 + 10^(-(home_elo - away_elo)/400))
    actual_A = 1 if team_A_score > team_B_score else 0

    margin = |team_A_score - team_B_score|
    mov_mult = margin^0.8 / (7.5 + 0.006 * |home_elo - away_elo|)

    rating_change = 32 * mov_mult * (actual_A - expected_A)
    team_A_rating += rating_change
    team_B_rating -= rating_change
END FOR
```

2. Weighted Four Factors

Dean Oliver's Four Factors framework decomposes basketball performance into four fundamental components that drive team success. Our implementation calculates team-level efficiency metrics and combines them using empirically-derived weights that reflect their relative importance to winning.

Factor Calculations: For each team, we compute season-aggregated statistics across all games:

$$\text{eFG\%} = \frac{\text{FGM}_2 + 1.5 \times \text{FGM}_3}{\text{FGA}_2 + \text{FGA}_3}$$

$$\text{TOV\%} = \frac{\text{TOV}_{\text{team}}}{\text{possessions}}$$

$$\text{ORB\%} = \frac{\text{OREB}}{\text{OREB} + \text{DREB}}$$

$$\text{FTRate} = \frac{\text{FTA}}{\text{FGA}_2 + \text{FGA}_3}$$

Composite Scoring: We weight these factors based on Oliver's research, with slight modifications to emphasize shooting efficiency and ball security:

$$\text{Score} = 0.35 \cdot \text{eFG\%} + 0.30 \cdot (1 - \text{TOV\%}) + 0.25 \cdot \text{ORB\%} + 0.10 \cdot \text{FTRate}$$

Prediction Methodology: For head-to-head matchups, we calculate the differential in each factor between opposing teams, apply our weighting scheme, and transform the result to a probability space using a linear scaling around 0.5. Contextual adjustments for rest, travel, and venue are applied as multiplicative factors to the base probability.

Algorithmic Pseudocode:

```
FOR each team:
    Calculate season totals for FGM_2, FGM_3, FGA_2, FGA_3, TOV, OREB, DREB, FTA
    eFG = (FGM_2 + 1.5*FGM_3) / (FGA_2 + FGA_3)
    TOV_pct = TOV / possessions
    ORB_pct = OREB / (OREB + DREB)
    FTRate = FTA / (FGA_2 + FGA_3)

    composite_score = 0.35*eFG + 0.30*(1-TOV_pct) + 0.25*ORB_pct + 0.10*FTRate
END FOR
FOR each matchup:
    factor_differential = home_team_score - away_team_score
    base_probability = 0.5 + (factor_differential * 0.8)
    Apply contextual adjustments for rest, travel, venue
```

```
END FOR
```

3. Temporal Logistic Regression

Traditional season-long statistics fail to capture the dynamic nature of team performance as seasons progress. Our temporal logistic regression model addresses this limitation by engineering features that explicitly model team evolution, recent form, and contextual factors that vary throughout the season.

Feature Engineering: We construct five primary features designed to capture different temporal aspects of team performance:

$$\text{home\_strength} = \text{home\_indicator} \times \log(1 + \text{attendance})/10$$

$$\text{rest\_impact} = \text{clip}(\text{rest\_days}, 0, 7)/7 - 0.5$$

$$\text{off\_eff} = \frac{\text{cumulative\_FGM}_2 + 1.5 \times \text{cumulative\_FGM}_3}{\text{cumulative\_FGA}_2 + \text{cumulative\_FGA}_3}$$

$$\text{def\_eff} = \text{expanding\_mean}(\text{opponent\_scores})$$

$$\text{last3\_wins} = \text{mean}(\text{wins\_in\_last\_3\_games})$$

Model Training: We employ a time-aware training methodology to prevent data leakage. Using TimeSeriesSplit cross-validation with 5 folds, we train a logistic regression model with L2 regularization. The temporal splits ensure that training data always precedes validation data chronologically.

Probability Calibration: Raw model outputs undergo isotonic calibration using a holdout set comprising 20% of the data. This step corrects for systematic bias in probability estimates and ensures that predicted probabilities align with observed frequencies.

Algorithmic Pseudocode:

```
FOR each game in chronological order:
    Update cumulative statistics (FGM, FGA, scores)
    Calculate rolling 3-game win percentage
    Compute efficiency metrics using expanding windows
    Store feature vector with target outcome
END FOR
Split data temporally (80% train, 20% holdout)
FOR each TimeSeriesSplit fold:
    Train LogisticRegressionCV with L2 penalty
    Validate on subsequent time period
END FOR
Fit isotonic calibrator on holdout set
Apply calibration to final predictions
```

4. PageRank Network Model

The PageRank algorithm, originally developed for ranking web pages, provides a natural framework for evaluating team strength within the interconnected network of college basketball competition. Our adaptation treats teams as nodes and game outcomes as directed edges, with the resulting centrality scores reflecting both direct performance and strength of schedule.

Network Construction: We model the season as a directed graph where teams are vertices and games create weighted edges. For each game, a directed edge flows from the losing team to the winning team, with edge weights determined by margin of victory raised to the 0.8 power to provide diminishing returns for blowout victories.

Home Court Adjustment: Before determining winners and margins, we apply a symmetric home court adjustment of 35 points to each team's effective score, mirroring our Elo implementation.

Google Matrix Formulation: Following the standard PageRank methodology, we construct the Google matrix as:

$$G = dP + (1 - d)\frac{\mathbf{1}\mathbf{1}^T}{N}$$

where P is the column-normalized transition matrix derived from our adjacency matrix, $d = 0.95$ is the damping factor, and N is the number of teams. The high damping factor emphasizes direct competitive relationships over random transitions.

Eigenvector Computation: We compute the principal eigenvector of G^T using power iteration, which converges to the stationary distribution representing team rankings. The resulting PageRank scores are then scaled to an Elo-compatible range (base 1500, span approximately 500 points).

Sample Size Adjustment: Teams with fewer than 5 games undergo regression-to-the-mean adjustment to account for limited data, with the adjustment factor proportional to the difference between actual and minimum required games.

Algorithmic Pseudocode:

```
Initialize adjacency matrix A (n_teams × n_teams)
games_played = zeros(n_teams)
FOR each game:
    Apply home court adjustment to scores
    winner, loser = determine_outcome(adjusted_scores)
    margin = |adjusted_score_winner - adjusted_score_loser|
    weight = margin^0.8
    A[loser, winner] += weight
    games_played[winner] += 1
    games_played[loser] += 1
```

```
END FOR
# Normalize to create transition matrix
P = column_normalize(A)
# Construct Google matrix
G = damping_factor * P + (1 - damping_factor) * ones_matrix / n_teams
# Compute PageRank via power iteration
pagerank = compute_principal_eigenvector(G.T)
# Apply sample size adjustments
FOR each team with games_played[i] < 5:
    adjustment = (5 - games_played[i]) / 5
    pagerank[i] = pagerank[i] * (1 - adjustment) + mean(pagerank) * adjustment
END FOR
# Scale to Elo-like ratings
ratings = 1500 + 500 * (pagerank - min(pagerank)) / (max(pagerank) - min(pagerank))
```

Cross-Validation and Consensus

To assess reliability, we compared the predicted win probabilities of East region games. To do this, we constructed a model agreement heatmap of pairwise probability differences. Where models diverged, we traced discrepancies to contextual variables (e.g., extreme travel distances or lineup changes), which informed our confidence in the Elo median predictions.

Results & Discussion

Our primary Elo rankings identified the leading teams within each region, with South Carolina (North), Louisville (South), and Stanford (West) as clear leaders (see Fig. 4 & 5). In each region, the difference between the top and fifth-ranked teams was under 100 Elo points-showing the high level of competitiveness in women's basketball.

Tournament win probability predictions for the East region revealed both clear favorites and near-toss-ups. For instance, NC State's six-day rest advantage resulted in a predicted win probability of 78.2%, while matchups like Fairfield vs. Towson were virtual coin flips. The impact of rest and travel adjustments was especially evident in games involving long-distance travel or uneven rest advantages.

Cross-validation played a pivotal role in our methodology. Figure 6 compares predicted win probabilities from all four models. We found Elo was consistently the "median" model, with Four Factors and Temporal Logistic models producing the highest and lowest bounds in most matchups. Outlier games-like Stony Brook's 3400-mile journey-were the main source of model disagreement, and always for reasons that could be traced to contextual factors.

Agreement among models was high: 68% of top-10 teams were shared across Elo and Four Factors, as shown in Fig. 7. When model consensus was weak, investigation revealed

contextual causes for the variance-such as schedule quirks or recent team changes-validating our ensemble approach.

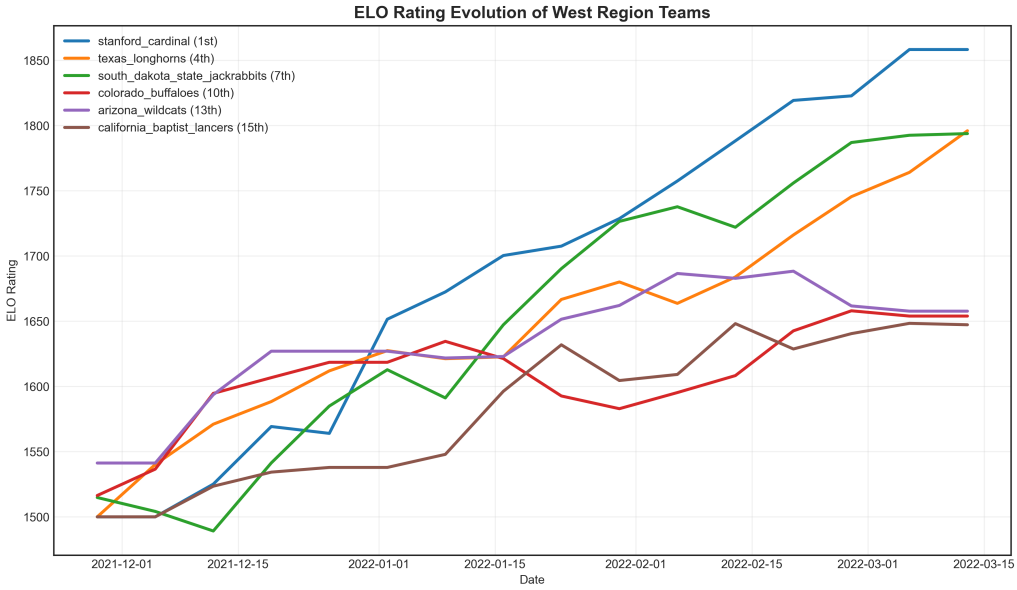

Figure 4: Elo Rating Evolution of West Region Teams

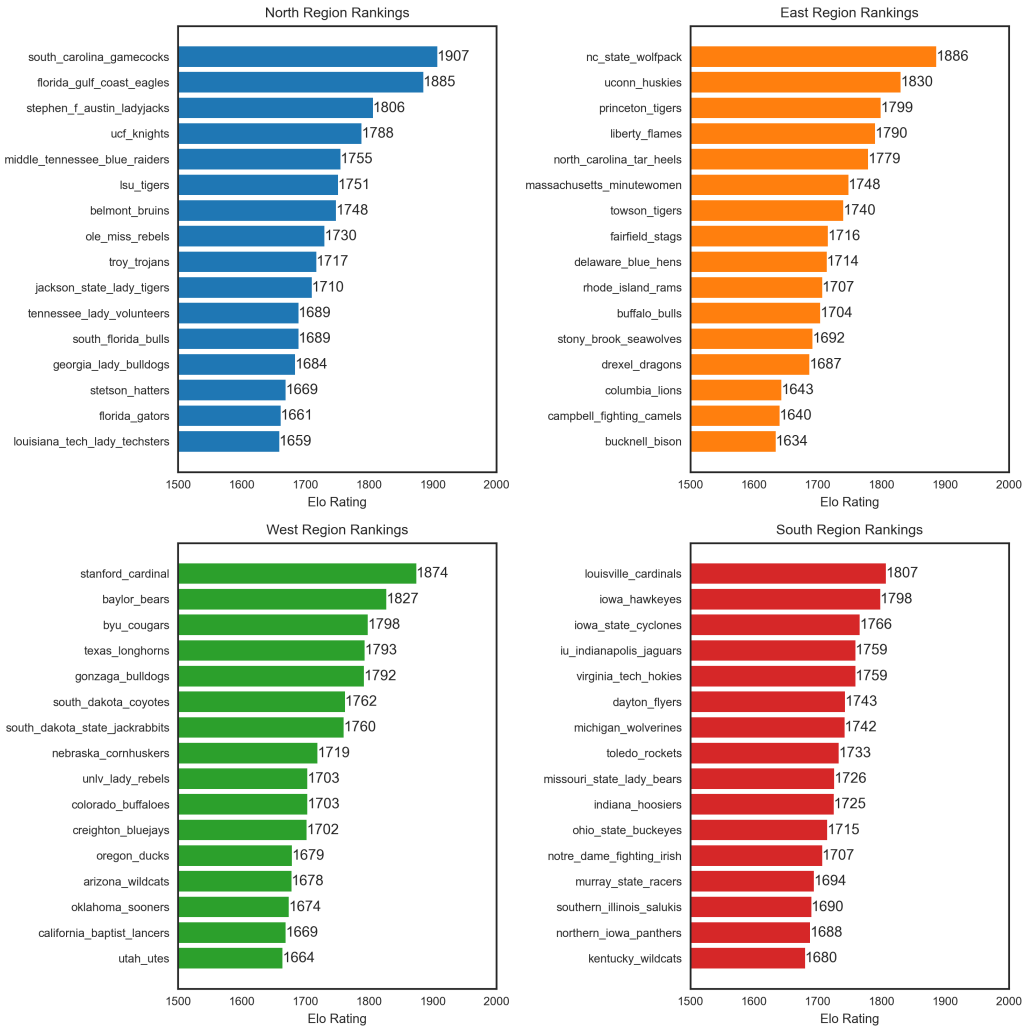

Figure 5: Regional Elo Rankings

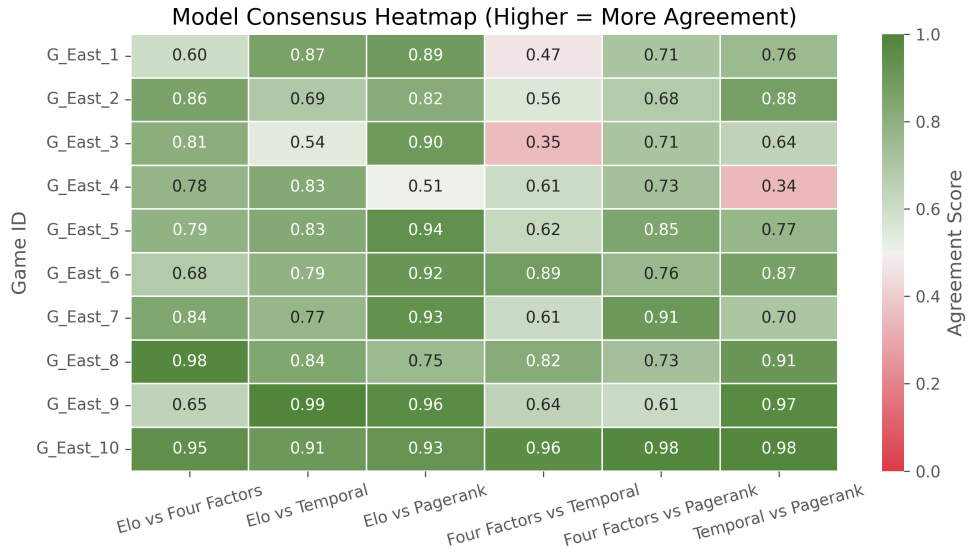

Figure 6: Model Consensus Heatmap

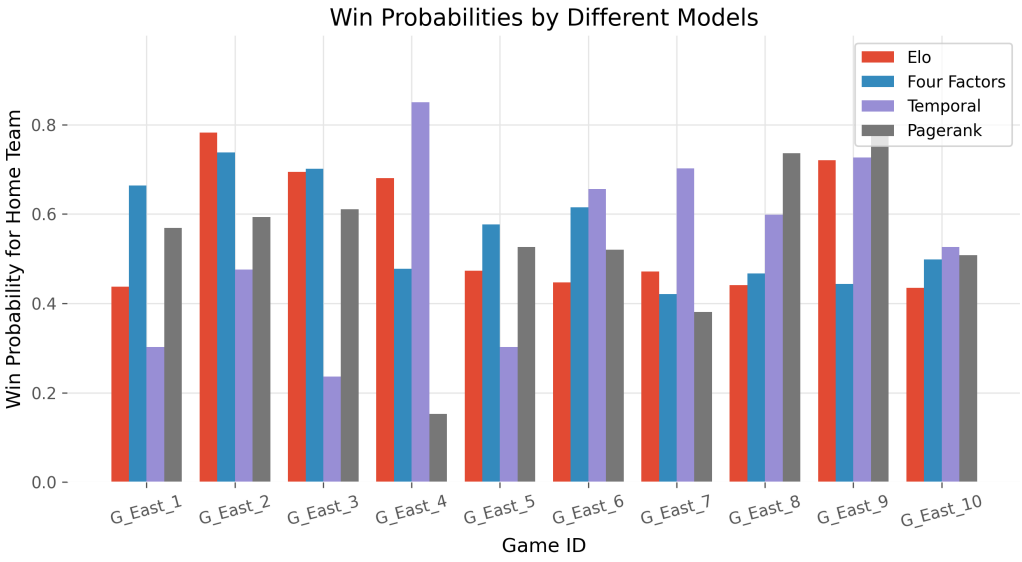

Figure 7: Predicted Win Probabilities of All Four Models

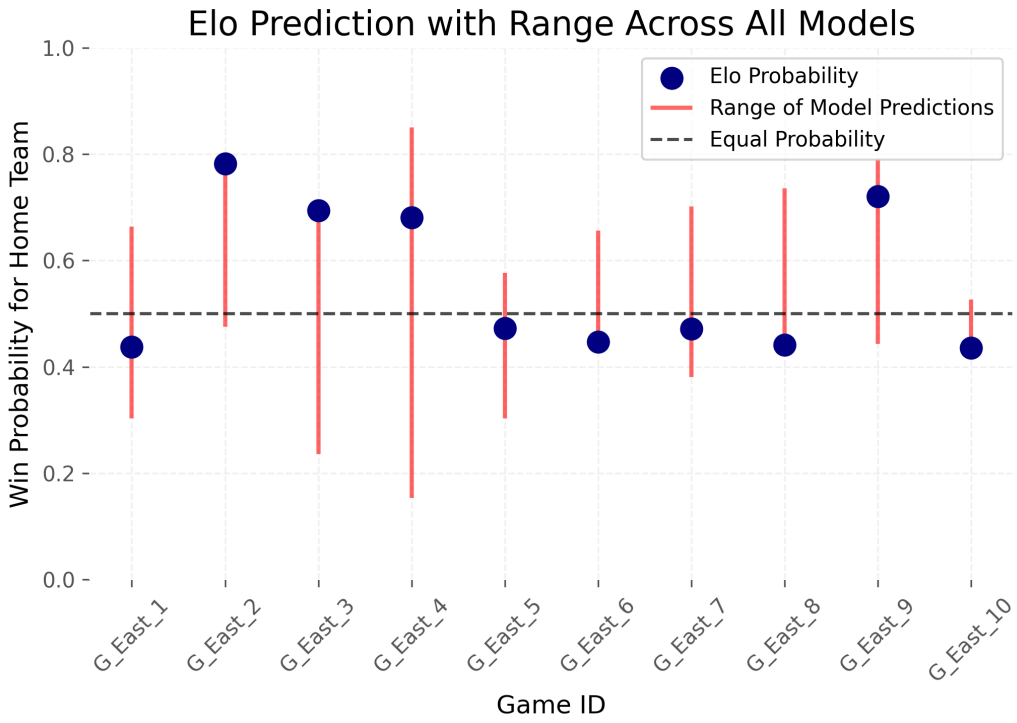

Figure 8: Elo Predictions and Disagreement

Conclusions

By combining four models-Elo, Four Factors, Temporal Logistic, and PageRank-we built a more robust and credible framework for basketball team rankings and win probability prediction than any single method could provide. Elo, enhanced with basketball-specific adjustments, gave us reliable and interpretable results. The other models were essential for detecting outlier games, highlighting uncertainties, and increasing our confidence in Elo's recommendations.

The main limitations of our work are the lack of player-level or injury data in the input, and residual Elo inflation in certain conference contexts. Future work should add Bayesian updating, explore neural networks for non-linear effects, and collect more granular data to support even better contextual modeling. Our approach demonstrates that ensemble methods and cross-validation drive both practical and scientific progress in sports analytics.

Acknowledgments

I am grateful to my advisor, Stephanie Beaulieu, for her invaluable guidance throughout this project. I also wish to acknowledge my teammates Samhitha Kovi and Ethan Baek for their work on data analysis and model development. My thanks also go to the Wharton Sports Analytics team for organizing the competition and providing the comprehensive dataset.

Finally, I appreciate the pioneering methodology of Seung et al. (2024), which informed several of our modeling decisions.

References

Bradley, R. A., & Terry, M. E. (1952). *Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 39*, 324–345.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting, 26*(3), 460–470.

Oliver, D. (2004). *Basketball on paper: Rules and tools for performance analysis.* Potomac Books, Inc.

Seung, E., Xu, J., Katz, R., Wetzstein, M., & Barr, M. (2024). *Calculating Win Probabilities of Any Matchup of Soccer Teams: A Whole-History Rating Approach for the Wharton High School Data Science Competition. Wharton Sports Analytics Journal.*

Silver, N. (2015). *FiveThirtyEight's 2015 NCAA tournament predictions.* `https://fivethirtyeight.com/features/how-we-made-our-forecasts-for-the-womens-ncaa-tournament/`

Solmos. (2020, March 10). Elo rating system. Solmos. `https://solmos.netlify.app/post/2020-03-10-elo-rating-system/elo-rating-system/`