Introduction:

- Given a dataset with comprehensive game-level and team-level statistics from over 5,300 games with key metrics:
  - Create alternative women's basketball team rankings within regions,
  - Predict outcomes of matchups in the Eastern region.
- Basketball outcomes are **notoriously difficult to predict** due to varying team strengths, home-court advantages, and contextual factors
- Single-model predictions rarely consider a sufficient number of factors to make accurate predictions

Background:

- Traditional basketball analytics relied on simple win-loss records or point differentials, which neglected the influence of metrics like attendance, home-court advantage, and strength of schedule
- Modern approaches have evolved to include:
  - ELO ratings (popularized by FiveThirtyEight for sports predictions)
  - Advanced metrics like Dean Oliver Four Factors
  - Neural networks and time-series models for tournament predictions
- Despite these advances, **models rarely agree on exact win probabilities**, creating uncertainty
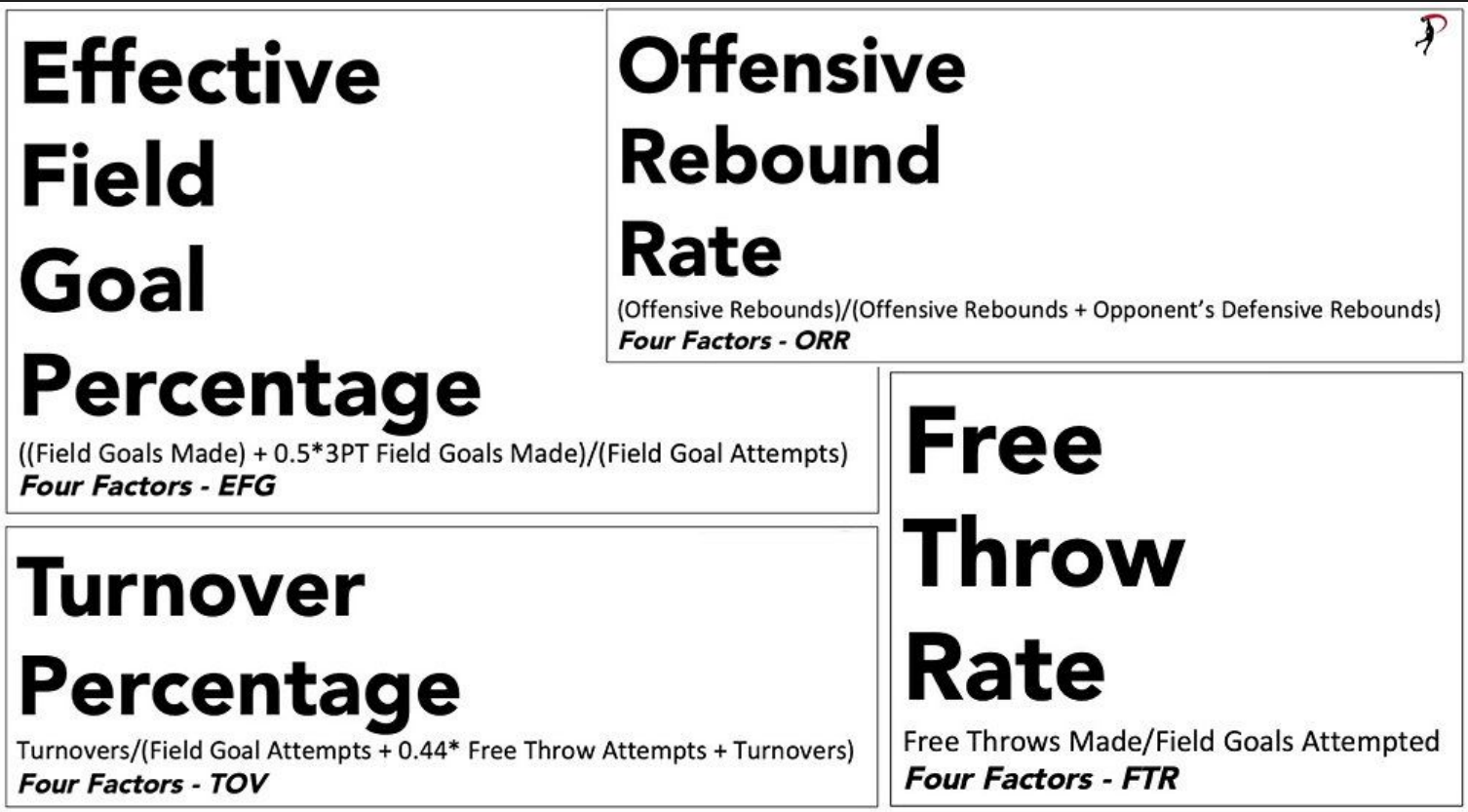
Effective Field Goal Percentage
((Field Goals Made) + 0.5*3PT Field Goals Made)/(Field Goal Attempts)
Four Factors - EFG

Offensive Rebound Rate
(Offensive Rebounds)/(Offensive Rebounds + Opponent's Defensive Rebounds)
Four Factors - ORR

Turnover Percentage
Turnovers/(Field Goal Attempts + 0.44* Free Throw Attempts + Turnovers)
Four Factors - TOV

Free Throw Rate
Free Throws Made/Field Goals Attempted
Four Factors - FTR

Figure 1

Research Question: How does ranking teams based on win-loss records compare to ranking teams based on multiple metrics specific to the game?

Goals

- Create regional team rankings based on simulated women's basketball data

- Generate accurate win probabilities for regional tournament matchups

Approach

- **Primary Method:** Dynamic ELO rating system with game-specific adjustments

- **Cross-Validation Strategy:** Implemented three independent models:

  - Dean Oliver's Four Factors statistical framework

  - Time-aware Logistic Regression with rolling performance metrics

  - PageRank-inspired directed graph network analysis

- **Consensus Methodology:** Analyzed where models converged and to establish robust ranking and probability bounds

ELO	Four Factors
- Easy to understand - Responsive to recent performance - Long-term view of team strength - Useful for predicting direct matchups	- Focuses on the key metrics of a game - Easy to understand areas for growth - Can be applied on both a team and player level
Logistic Regression	**PageRank**
- Used to predict binary outcomes - Incorporates a variety of input variables - Handles non-linear relationships	- Analyzing team networks - Dynamic weighting of team's statistics - Highlights non-traditional statistics

Figure 2

Data Prep

- Considered East region teams for Phase 1a rankings

- Combined both rows per game into one

- Imputed missing values through:

  - Rest days (rest_days): replaced NAs with median (3 days)

  - Attendance (attendance): imputed with venue-specific averages

  - Technical fouls (F_tech): zero-filled (rare events)

- Ensured chronological integrity by sorting all games by game_date before processing

Additional Variables

- Home court advantage indicator

- Modified margin of victory formula accounting for point differentials and elo differentials

- Rest differential: rest_days_Home - rest_days_Away

- Travel-induced fatigue metric based on travel_dist

Tools Used

- **Python 3.9** — Runtime

- **Pandas** — Data Manipulation

- **NumPy** — Calculations

- **SciPy** — Linear Algebra
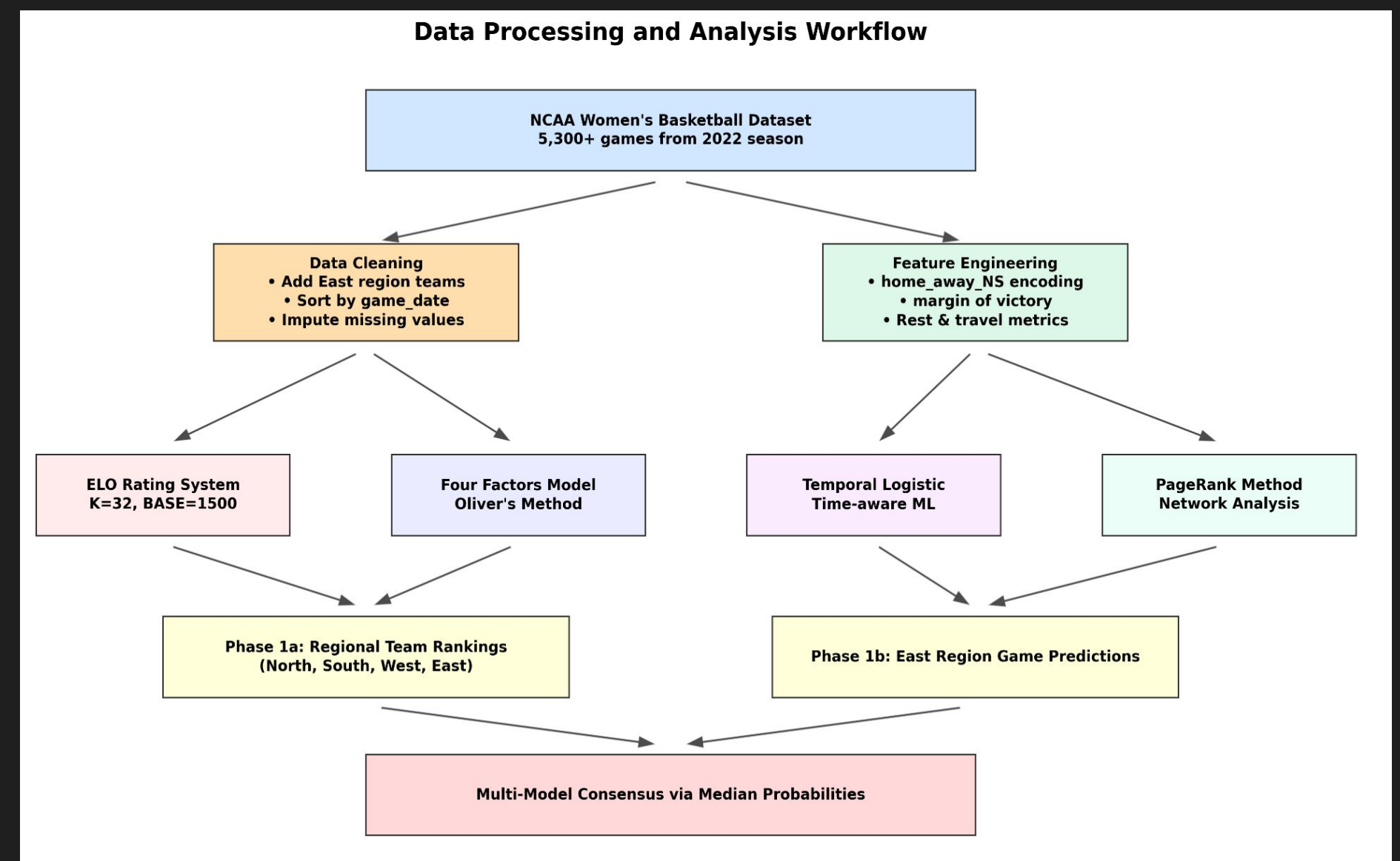
- **scikit-learn** — Temporal Modeling

Data Processing and Analysis Workflow

NCAA Women's Basketball Dataset
5,300+ games from 2022 season

Data Cleaning
• Add East region teams
• Sort by game_date
• Impute missing values

Feature Engineering
• home_away_NS encoding
• margin of victory
• Rest & travel metrics

ELO Rating System
K=32, BASE=1500

Four Factors Model
Oliver's Method

Temporal Logistic
Time-aware ML

PageRank Method
Network Analysis

Phase 1a: Regional Team Rankings
(North, South, West, East)

Phase 1b: East Region Game Predictions

Multi-Model Consensus via Median Probabilities

Figure 3

Method

Analysis

ELO Rating System (Primary):

- Initialized all teams at **1500 base rating**

- Applied Adaptations:

  - K-Factor = 32 (controls rating update magnitude)

  - Home court advantage = 70 points (~10% win probability)

  - Margin of victory multiplier = 1.1 (rewards dominant wins)

  - Rest Day Adjustment - +7.2 points/rest day

  - Travel Distance Adjustment = -2.8 points/300 miles traveled

Additional Validation Models:

- **Four Factors:** **Weighted combination** of shooting (35%), turnovers (30%), rebounding (25%), and free throws (10%)

- **Temporal Logistic:** Time-aware machine learning with **rolling team strength metrics** and L2 regularization

- **PageRank:** Directed network where wins create weighted edges between teams, with **eigenvector centrality** determining team strength

$$P(\text{win}) = \frac{1}{1+10^{-(\text{ELO}_A - \text{ELO}_B)/400}}$$

$$\text{MOV}_{\text{mult}} = \frac{(\text{point difference})^{0.8}}{7.5 + 0.006 \times |\text{ELO}_A - \text{ELO}_B|}$$

ELO Rating Evolution of West Region Teams

- stanford_cardinal (1st)
- texas_longhorns (4th)
- south_dakota_state_jackrabbits (7th)
- colorado_buffaloes (10th)
- arizona_wildcats (13th)
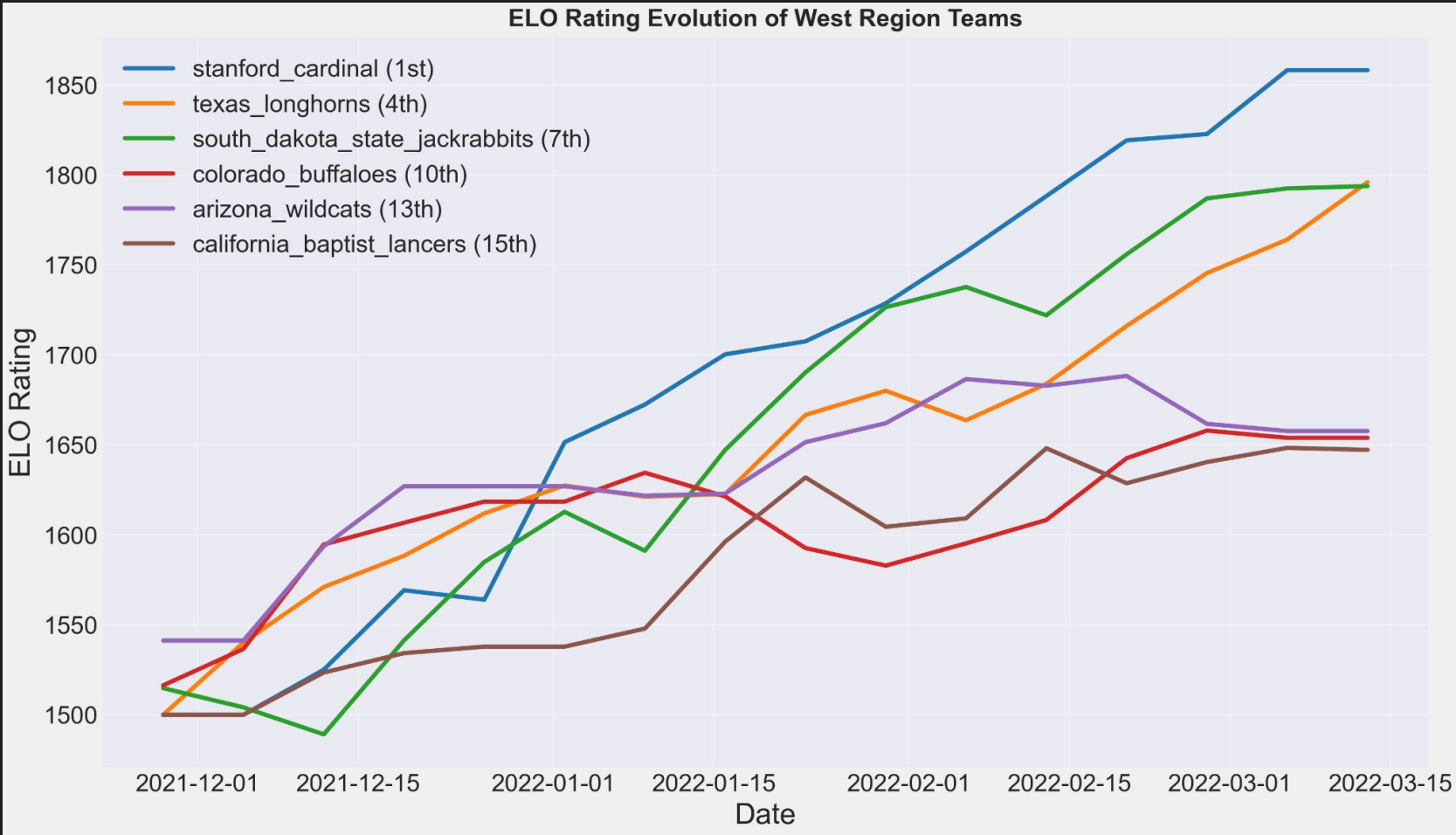- california_baptist_lancers (15th)

Figure 4

Result 1: Regional Ranks & Predictions

Regional Ranking Analysis:

- South Carolina Gamecocks (1907) emerged as the clear leader, with a 22-point gap to Florida Gulf Coast (1885)

- **Tightly clustered top teams** with Louisville Cardinals (1807) edging Iowa Hawkeyes (1798) by only 9 points

- Stanford Cardinal dominated with 1874 points, showing consistent performance against tough opposition

- Key Insight: Top 5 teams within each region separated by less than 100 ELO points (~14% win probability difference)

East Region Prediction Analysis:

- High Confidence Games: NC State vs. Rhode Island (78.2%), UConn vs. Campbell (72.0%)

- Contested Matchups: Five games had ELO probabilities between 43-47%, indicating **near coin-flips**

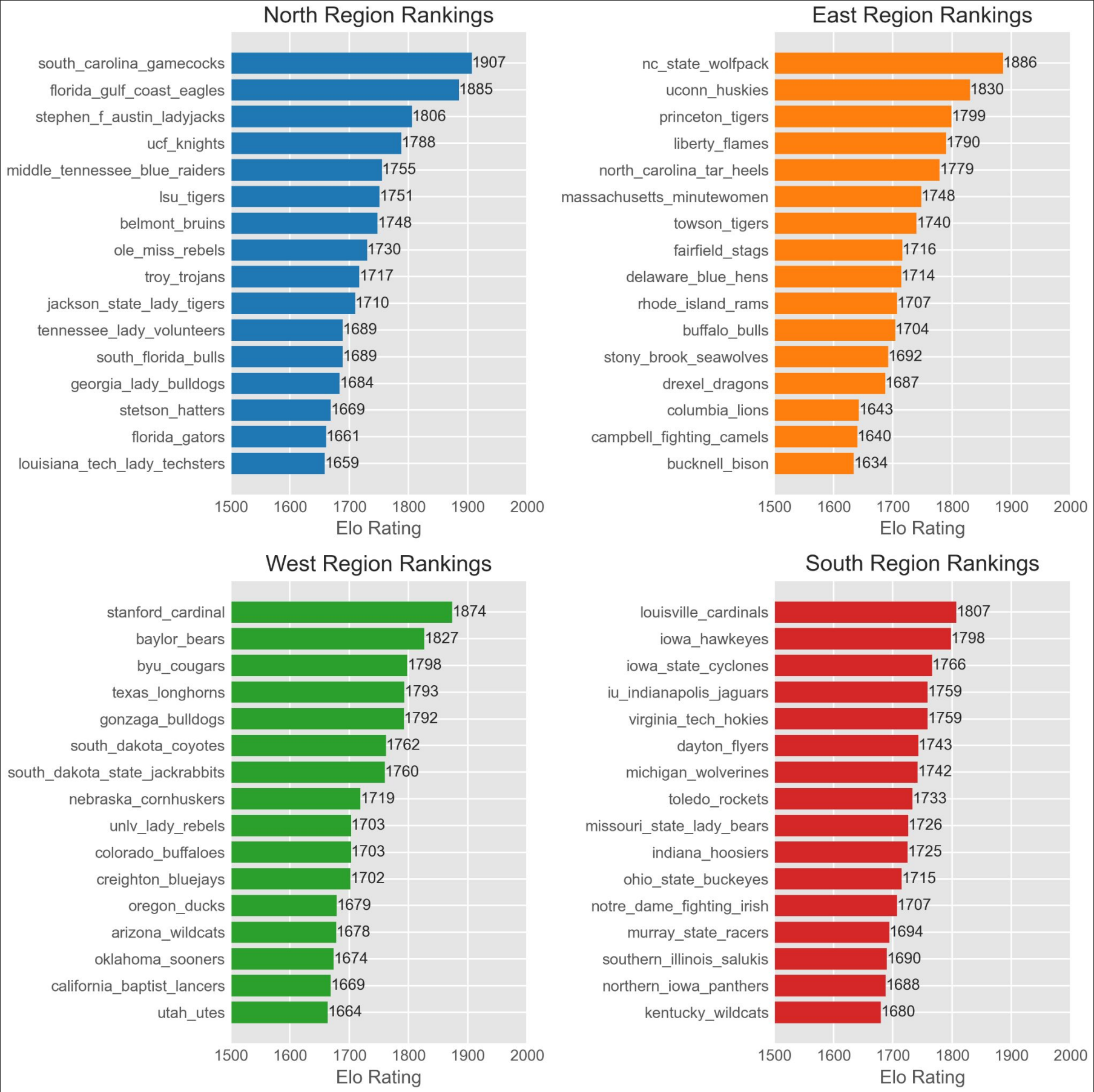- Liberty vs. Bucknell showed the **largest disagreement** (ELO: 68.1%, Temporal: 85.0%, Four Factors: 47.7%)
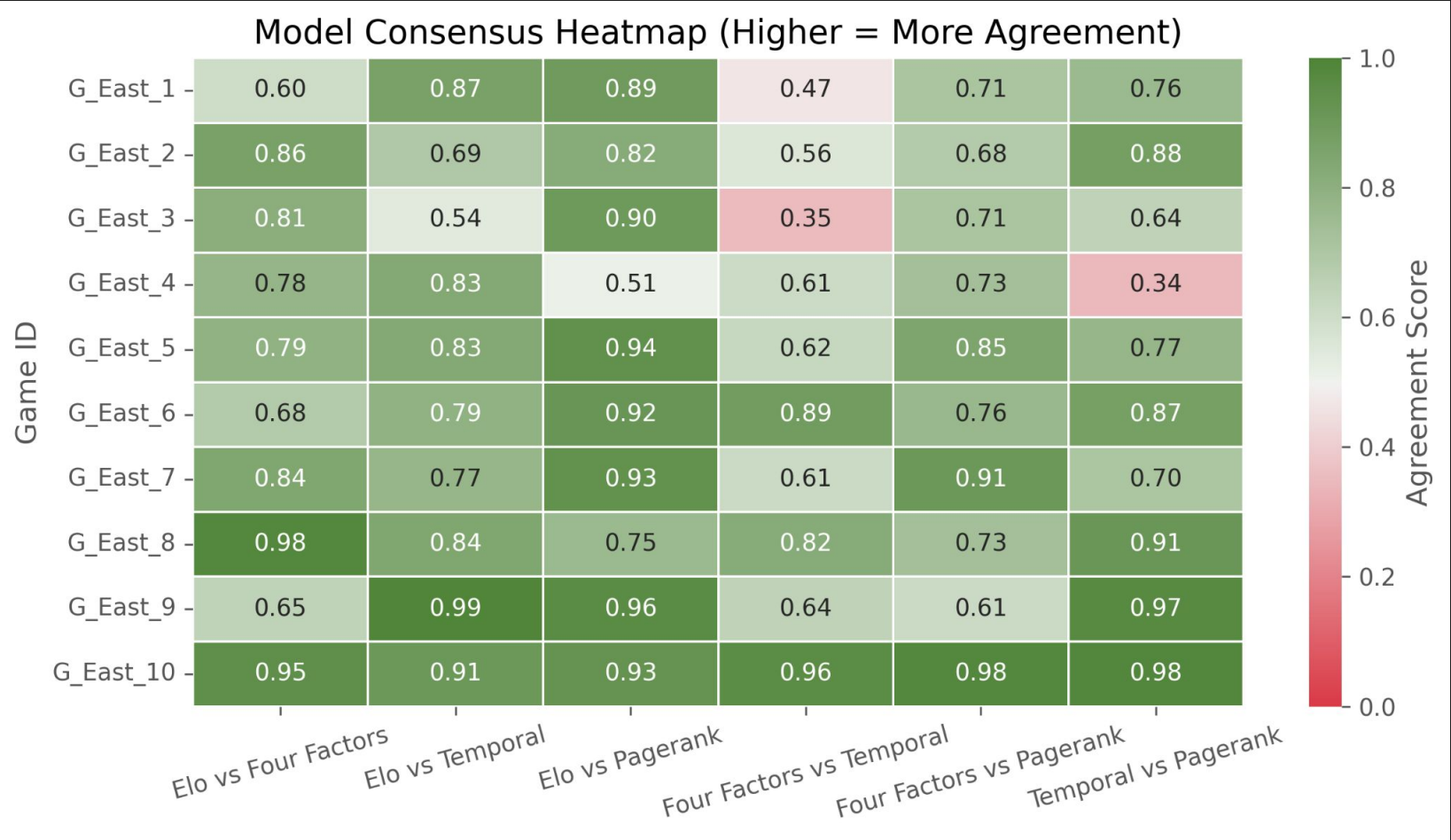

Figure 5

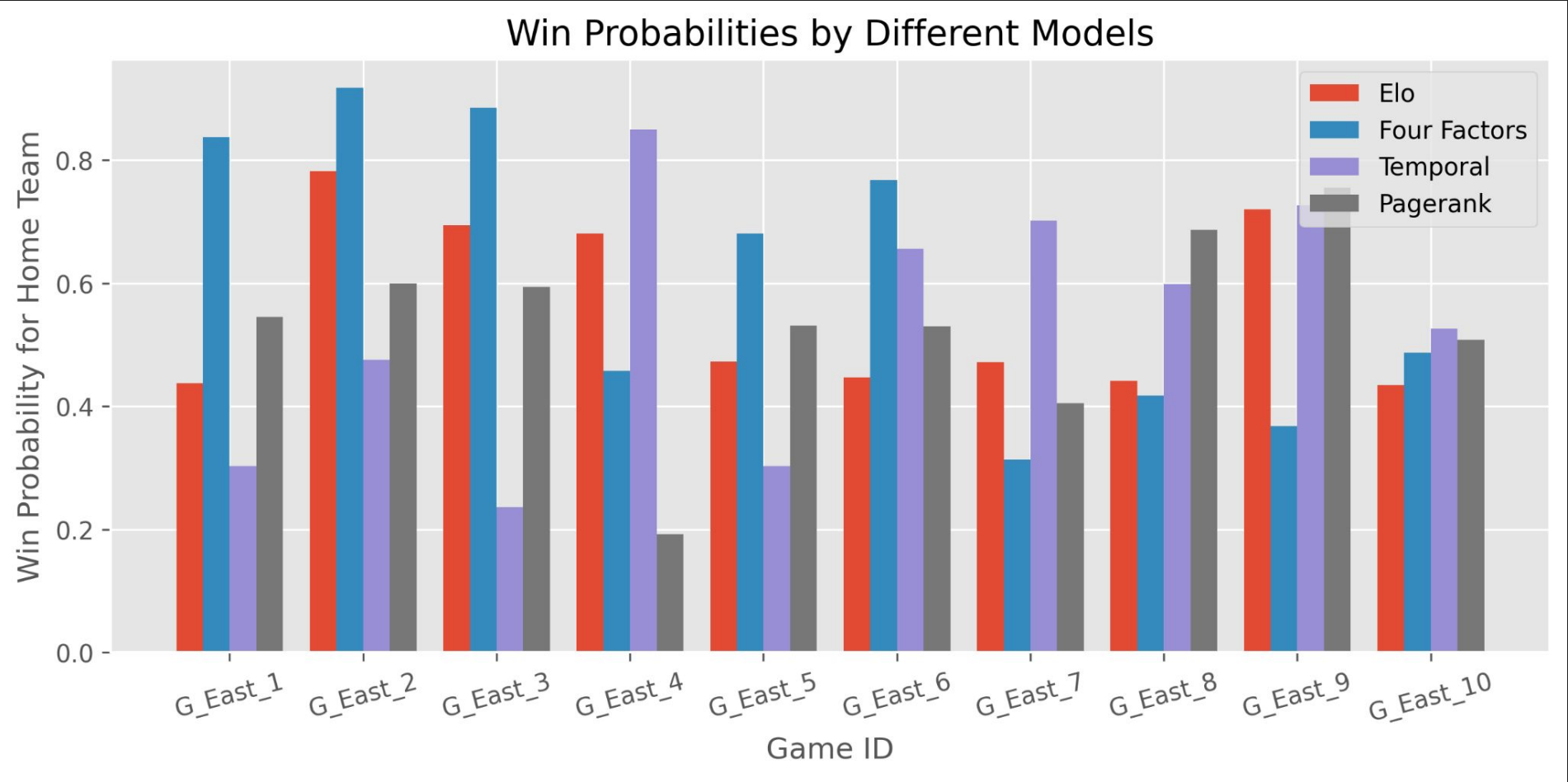# Result 2: Cross-Validation for ELO

Figure 6

Figure 7

Cross-Validation for Rankings:

- **68% agreement** on top-10 teams across regions

- Jackson State Tigers ranked significantly higher in Four Factors (#1 in North) than in ELO (#10)

- West Region showed highest consistency between models (80% overlap in top 10)

Cross-Validation for Probability:

- ELO provided the median probability in 7 of 10 games
  When models significantly disagreed, contextual factors explained differences:

  - Travel distance impact (Stony Brook's 3400-mile journey in Game 7)

  - Rest differential (NC State's 6-day advantage in Game 2)

Conclusion

General Findings

- **ELO provides reliable predictions** and typically fell between more extreme model outputs

- Multi-model consensus approach enabled confidence assessment: **70% probability agreement** within ±10% across models

- Game-specific factors (ex. home advantage) are significant

Considerations for Coaches

- Offensive index (FGA_2, FGM_2, FGA_3, FGM_3) found to be more significant than defensive index

- The specific environment & context of every game is crucial

Figure 8

Limitations

- Limited historical data for some East region teams

- Long-term inflation of ELO scores

- Models cannot account for "tournament psychology" (pressure, experience factors)

Improvements

- Incorporate neural networks to address highly non-linear variables

- Explore Bayesian updating for parameter optimization

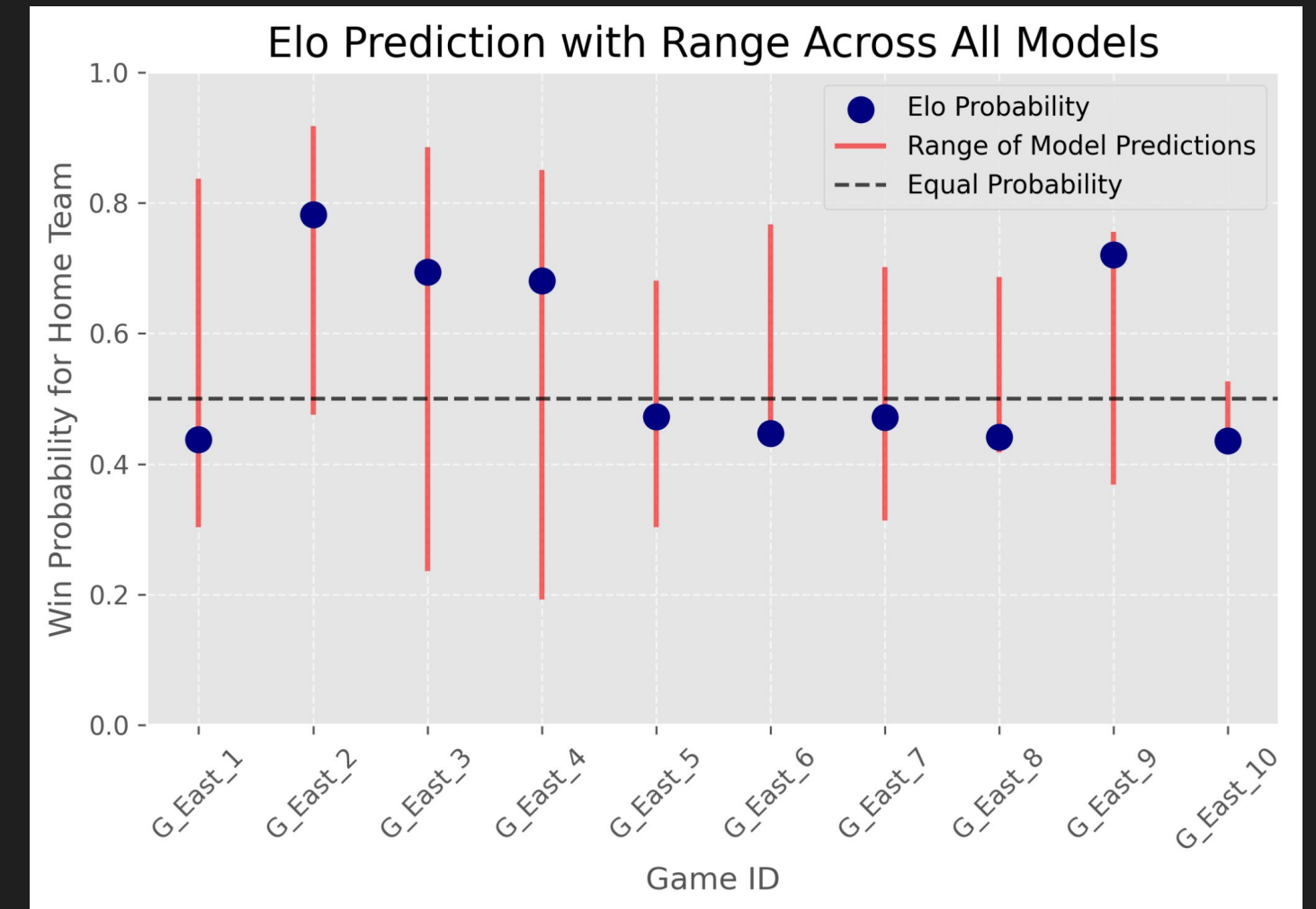- Finetune hyperparameters in existing four models