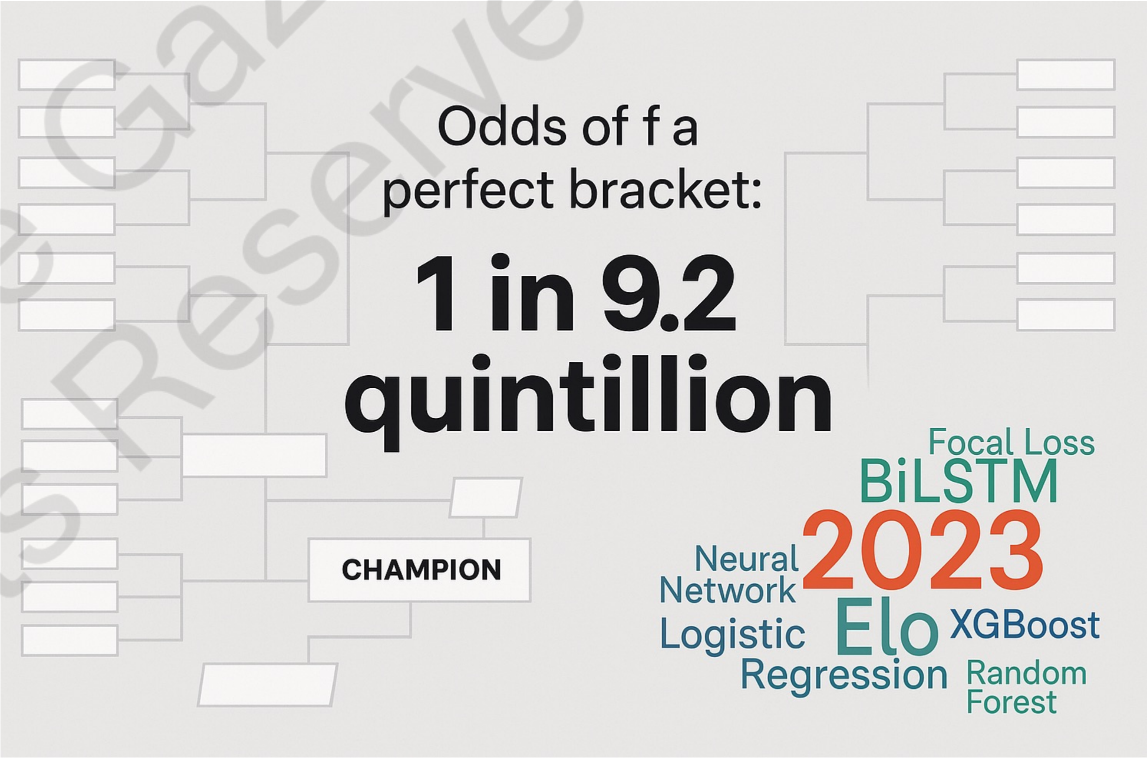
Competition: Kaggle March Madness 2025
Focus: Predict NCAA outcomes using multiple modeling strategies
Approach: Mix of deep learning & traditional ML

Fae Gaze

Machine Learning Researcher in Bioinformatics

Agendas

- **Competition & Data** – Overview of the task and dataset
- **EDA & Features** – Key exploratory findings and engineered features
- **Modeling Approaches** – Elo rating system, BiLSTM deep learning model, ensemble of tree models
- **Tournament Simulation** – Using model predictions to simulate bracket outcomes
- **Results & Evaluation** – Model performance, Brier Score metric, and insights
- **Interactive Polls/Quizzes** – Engaging questions throughout (marked accordingly)
- **Conclusion** – Lessons learned, future work, and competition results



Competition Overview

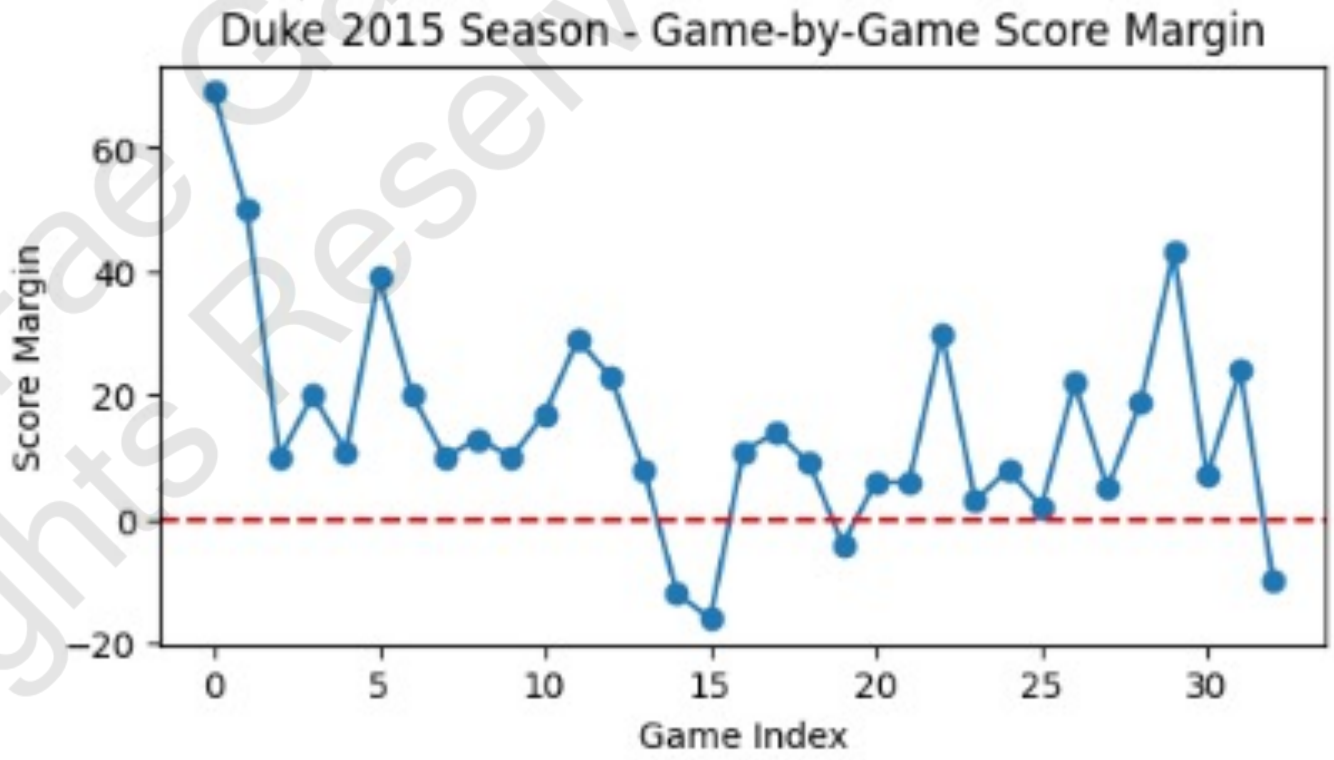
- **Objective:** Predict the probability of each team winning each possible matchup in March Madness
- **Format:** 64-team single-elimination tournaments (Men’s & Women’s); 63 games each
- **Data Provided:** Historical game results (regular season and tournament) and team seeds: https://www.kaggle.com/competitions/march-machine-learning-mania-2025/data
- **Evaluation:** Brier Score (mean squared error of predicted win probabilities) on actual tournament outcomes

Quiz: How Many Possible Brackets?

- Q: *How many possible ways can a 64-team March Madness bracket play out (i.e., distinct bracket outcomes)?*
  - A. ~9.2 **quintillion** (9.2×10^{18})
  - B. ~9.2 **trillion** (9.2×10^{12})
  - C. ~92 **billion** (9.2×10^{10})

Data Sources and Preparation

- **Data Sources & Preparation**
- **Historical Games:** All NCAA DI men's & women's results (1985–2024 seasons)– includes regular season and tournament games
- **Team Seeds:** Each team's seed in each tournament (e.g., "1" for top seed, "16" for lowest) provided in data
- **Regular Season Stats:** Derived metrics per team (wins, losses, average scores, etc.)
- **Data Merging:** Combined team stats with matchup data to create features for each game (e.g., seed difference, win percentage difference)



The above line plot shows an example of a single team's game-by-game performance, visualizing score margin variability.

Game Location Analysis (EDA)
Win Rate by Location **Slide Content:**

- Home: 59%
- Away: 30.7%
- Neutral: 10.2%
- Plot: Bar chart (H, A, N)

Seed Extraction & Tournament Logic

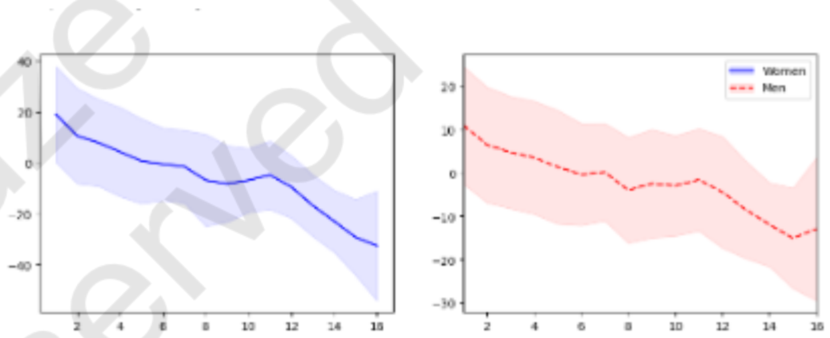
PARSING SEEDS: TOURNAMENT SEED STRINGS (E.G., "W01") WERE PARSED INTO NUMERIC SEEDS (1–16) AND REGIONS

BRACKET STRUCTURE: THE BRACKET IS FIXED – E.G., SEED 1 VS 16, 2 VS 15, ETC., IN THE FIRST ROUND. WE UTILIZED OFFICIAL BRACKET SLOT DATA TO KNOW WHICH WINNERS MEET IN LATER ROUNDS

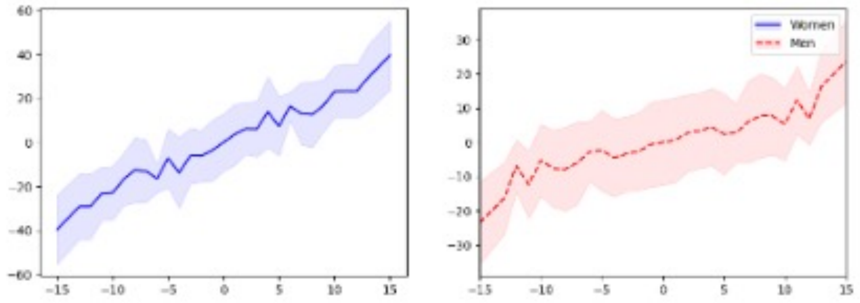
TOURNAMENT SLOTS: USED THE PROVIDED MAPPING OF WINNERS TO NEXT ROUND SLOTS (E.G., WINNER OF GAME X GOES TO SLOT Y) TO SIMULATE BRACKET PROGRESSION

DATA CONSISTENCY: ENSURED TEAM IDS, SEEDS, AND REGIONS ALIGN BETWEEN REGULAR SEASON AND TOURNAMENT DATA FOR CORRECT FEATURE MERGING

seed is predictive for predicting the point difference¶?



seed difference is predictive for predicting the point difference?¶



Data Sources & Preparation

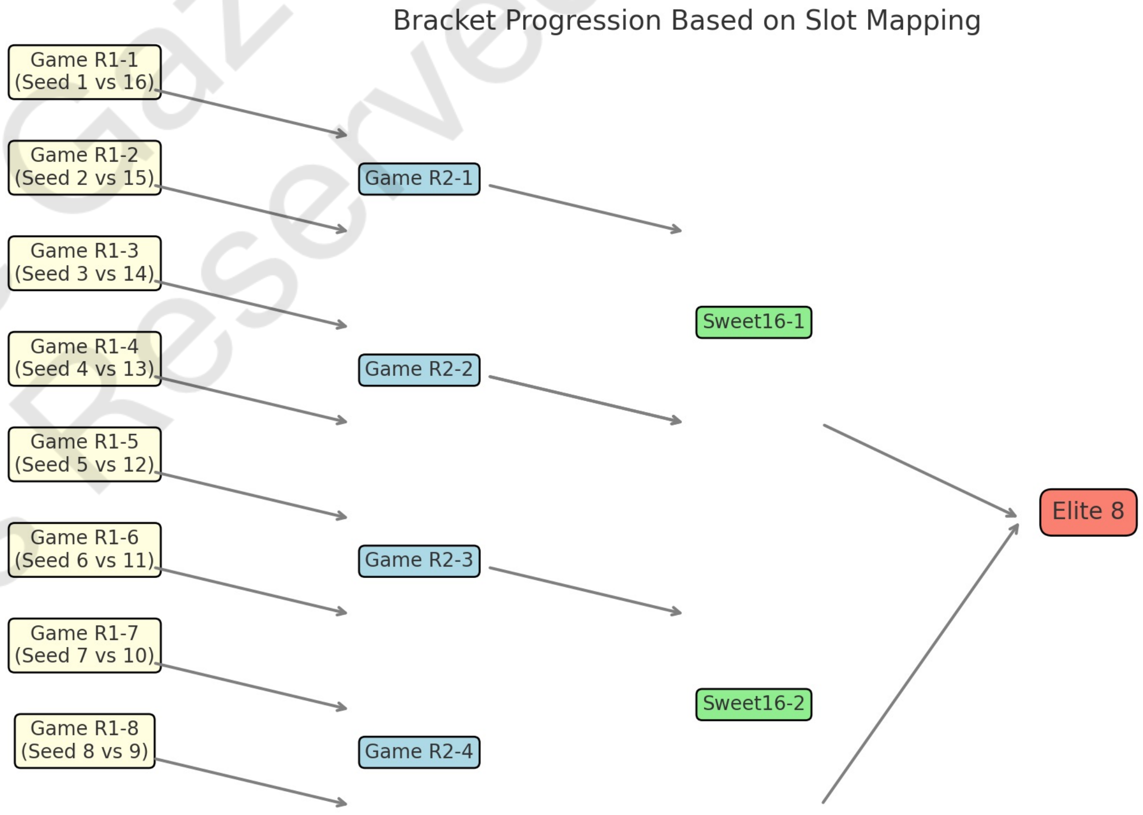
- **Historical Games:** All NCAA DI men's & women's results (1985–2024 seasons)

https://www.kaggle.com/competitions/march-machine-learning-mania-2025/data

- – includes regular season and tournament games
- **Team Seeds:** Each team's seed in each tournament (e.g., "1" for top seed, "16" for lowest) provided in data
- **Regular Season Stats:** Derived metrics per team (wins, losses, average scores, etc.)
- **Data Merging:** Combined team stats with matchup data to create features for each game (e.g., seed difference, win percentage difference)

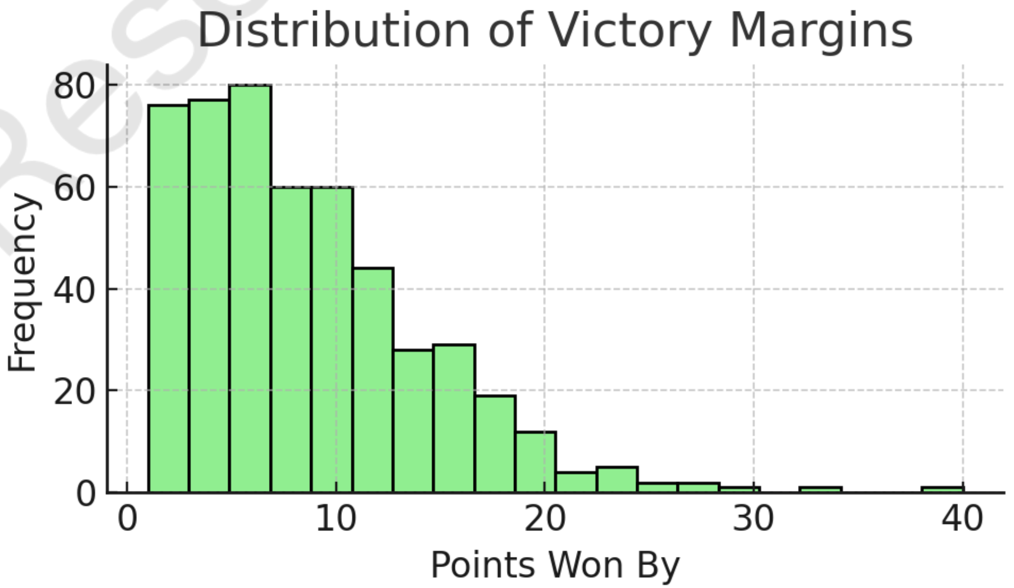
Seed Extraction & Tournament Logic

- **Parsing Seeds:** Tournament seed strings (e.g., "W01") were parsed into numeric seeds (1–16) and regions
- **Bracket Structure:** The bracket is fixed – e.g., seed 1 vs 16, 2 vs 15, etc., in the first round. We utilized official bracket slot data to know which winners meet in later rounds
- **Tournament Slots:** Used the provided mapping of winners to next round slots (e.g., winner of game X goes to slot Y) to simulate bracket progression
- **Data Consistency:** Ensured team IDs, seeds, and regions align between regular season and tournament data for correct feature merging



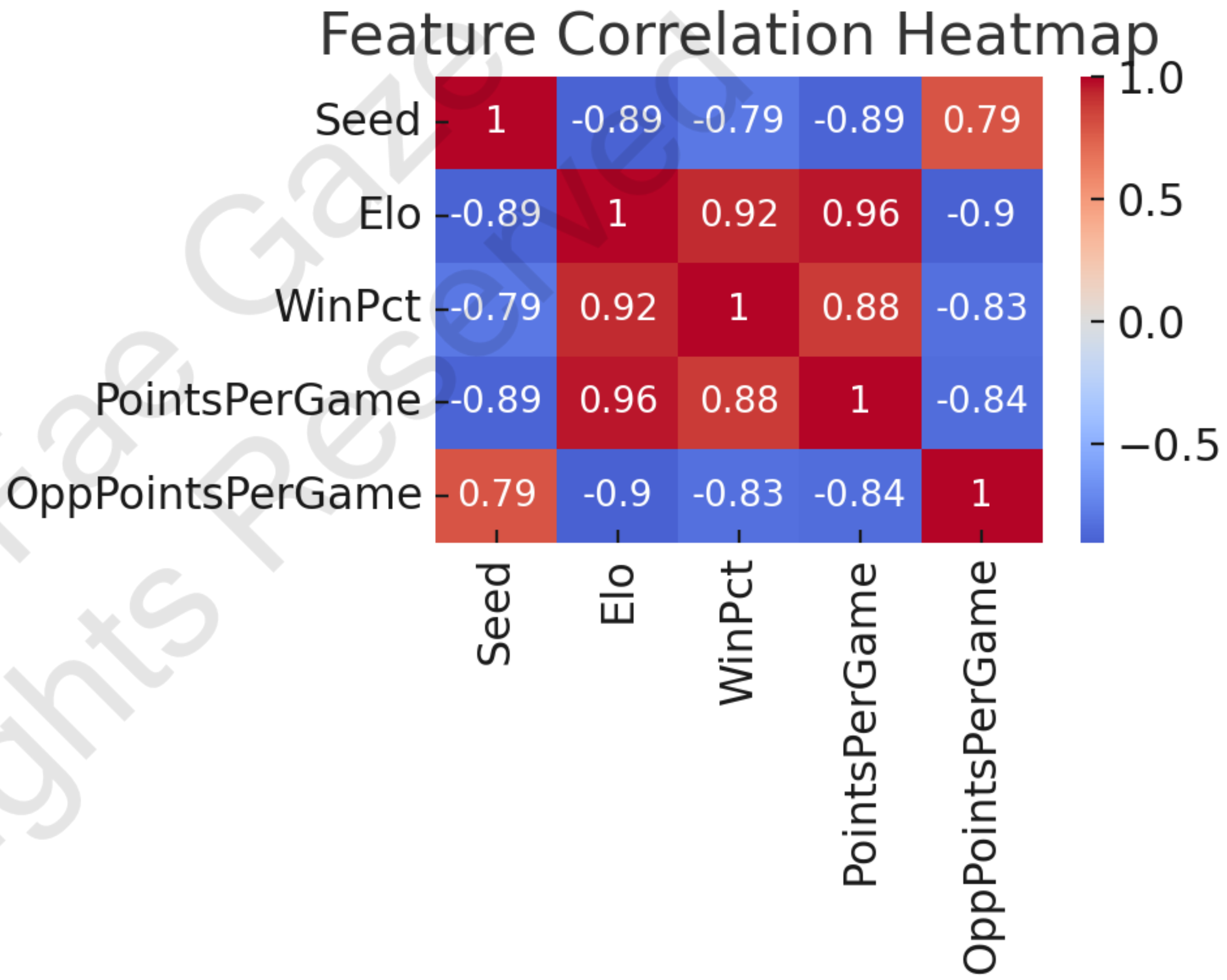
EDA: Historical Trends

- **Score Distributions:** Analyzed score totals and victory margins (blowouts vs close games)
- **Seed vs Outcome:** Higher seeds (1 = best teams) tend to win more often; documented frequency of upsets by seed difference
- **Tournament Upsets:** Identified patterns (e.g., the notorious 12 vs 5 seed upset rate ~35% historically)
- **Season Performance:** Teams with strong regular-season metrics (win%, rating) usually go deeper in tournament



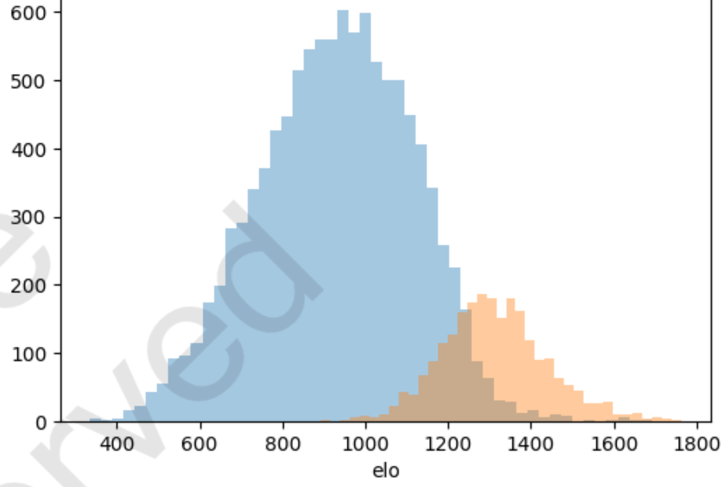
EDA: Feature Correlations

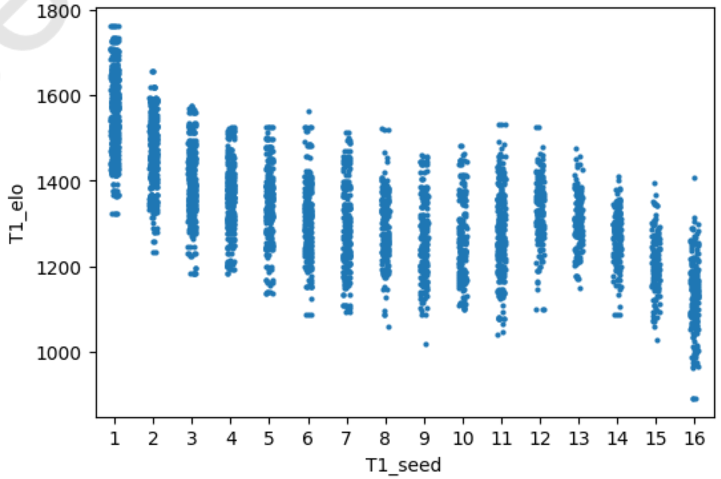
- Heatmap of key feature correlations. Notice the strong negative correlation between *Seed* and *Elo* (better teams have lower seeds and higher Elo ratings). *Elo* is highly positively correlated with Win% and Points Per Game (PPG), indicating that our Elo rating captures similar information as season win percentage. Opponents' PPG (defensive strength) is negatively correlated with Elo and Win%, as expected. **Seeds vs Elo:** Seed is inversely related to Elo rating (better-seeded teams have higher Elo)
- **Win% vs Elo:** Teams with higher Elo also have higher season win percentages (correlation >0.9)
- **Offense/Defense:** Teams that score more (PPG) and allow fewer points (Opp PPG) tend to have better Elo and lower seed numbers
- **Multicollinearity Note:** Many features are inter-related (we addressed this via model regularization and dimensionality reduction where needed)

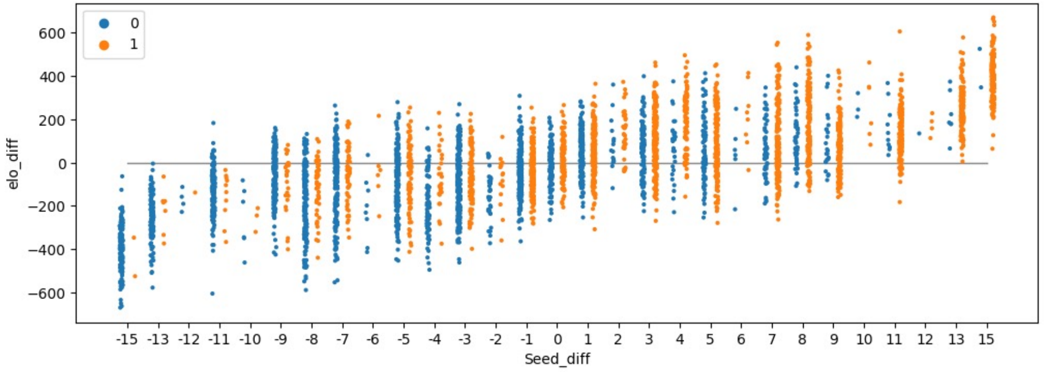


Feature Engineering

- **Elo Ratings:** Pre-computed a dynamic rating for each team based on game results (captures team strength and momentum)
- **Seed Difference:** Numerical difference in seeds between the two teams in a matchup (or seed rank vs rank)
- **Season Averages:** Offense vs defense stats (points scored minus points allowed, rebounding, etc.)
- **Recent Form:** Tournament-specific features like win streak entering tournament, conference champion indicator
- **Feature Aggregation:** Model inputs often use differences or ratios (Team A stat minus Team B stat) for each matchup

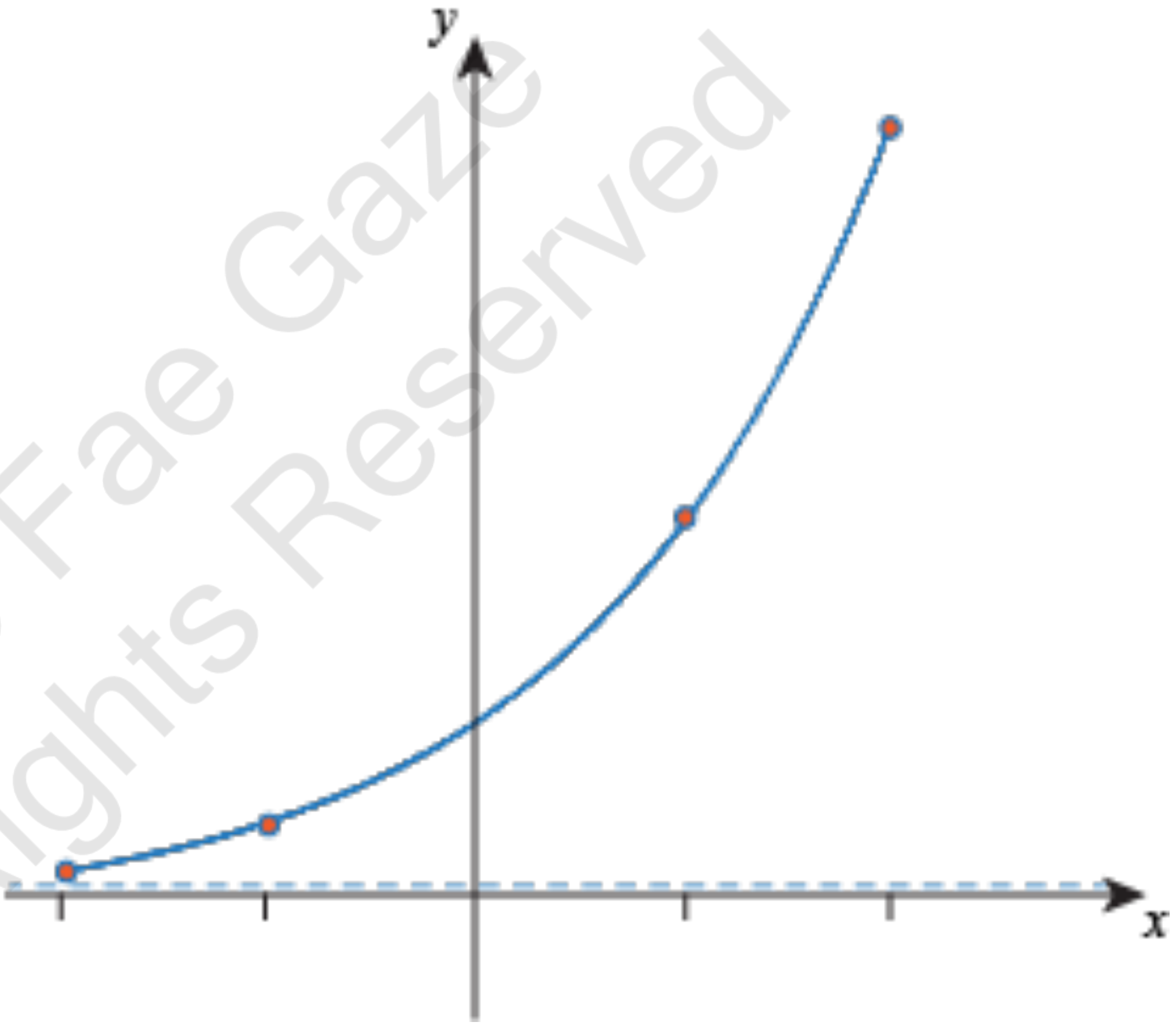






Advanced Feature Engineering / Hardest Features

- Modeled **team latent strength** using GLM (random effects on team IDs)
- Fitted nonlinear relationships between **predicted score margin** and **win probability**
- Applied splines to improve **probability calibration**

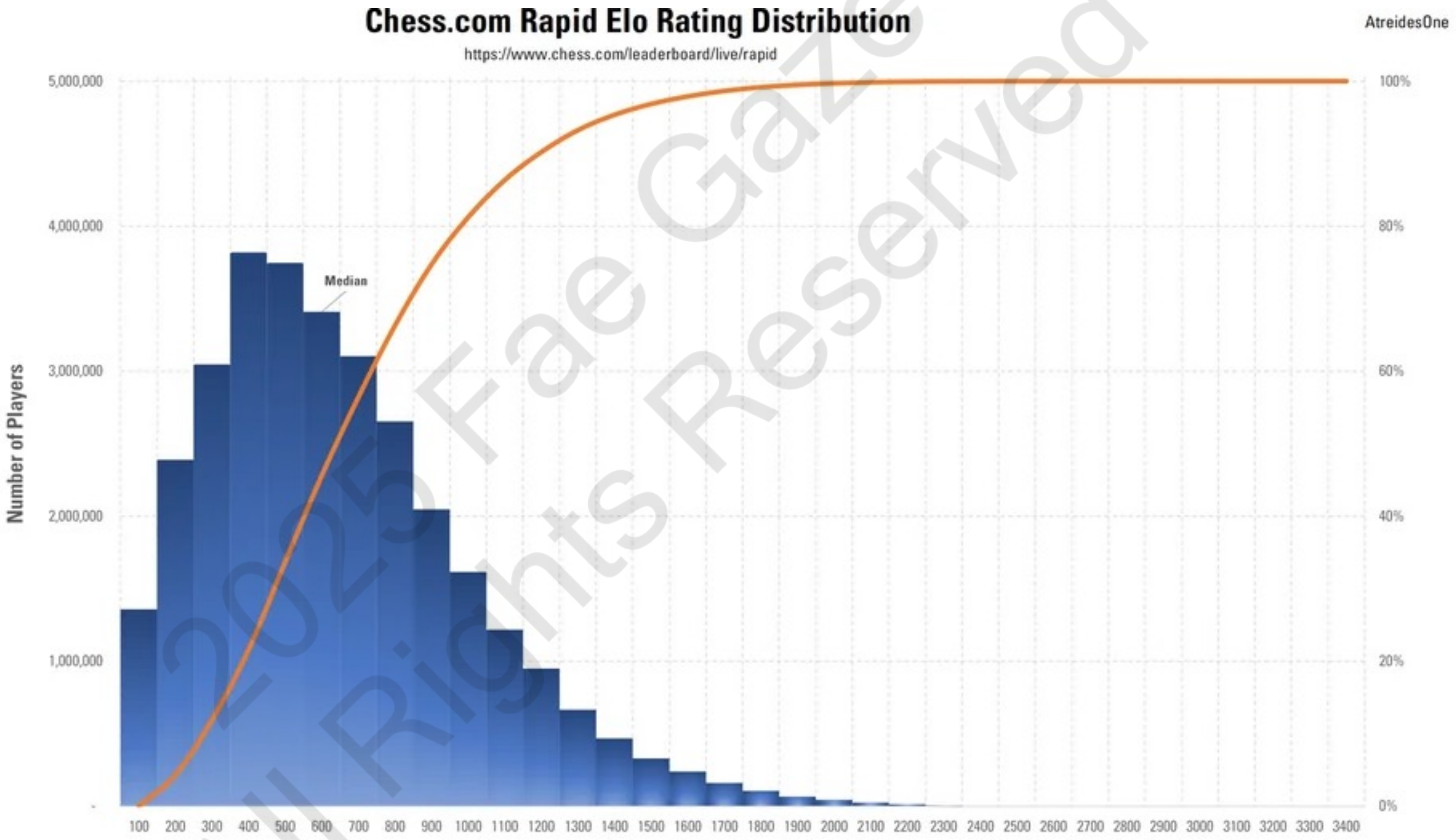


Elo Ratings Are Not Uniform

Data Date: 20/09/2023

Elo	Number of Players	Percentile
100	1,357,724	0.00%
200	2,389,510	4.31%
300	3,046,897	11.90%
400	3,820,148	21.58%
500	3,746,648	33.71%
600	3,409,892	45.61%
700	3,103,605	56.44%
800	2,654,090	66.30%
900	2,047,360	74.73%
1000	1,615,190	81.23%
1100	1,217,303	86.36%
1200	949,632	90.23%
1300	665,861	93.25%
1400	468,526	95.36%
1500	329,498	96.85%
1600	238,675	97.90%
1700	159,404	98.65%
1800	105,748	99.16%
1900	66,635	99.50%
2000	44,265	99.71%
2100	24,553	99.85%
2200	13,228	99.93%
2300	6,167	99.97%
2400	2,514	99.99%
2500	745	99.99643%
2600	241	99.99880%
2700	104	99.99956%
2800	26	99.99989%
2900	7	99.99997%
3000	-	99.999997%
3100	-	99.999997%
3200	-	99.999997%
3300	-	99.999997%
3400	1	99.999997%
Total	31,484,197	

Estimated Median: 591



*Best estimate of median for a histogram: L + ((n/2 – F) / f) * w, where: L= The lower limit of the median bin, n= The total number of observations, F= The cumulative frequency up to the median bin, f= The frequency of the median bin, w= The width of the bin. I am assuming that all the players listed as Elo 600 are in the range 550 - 650 (i.e. 600 is the bin centrepoint). From https://www.statology.org/histogram-mean-median/*

Poll: Can Simple Ratings Beat Complex Models?

- **Q:** Given the richness of historical data, do you think a simple rating system like Elo can match or outperform more complex ML models on this task?
  - A. *Yes – Elo is already very strong, complex models might overfit*
  - B. *No – A sophisticated model will find patterns Elo misses*
  - C. *Hybrid – Use Elo as a feature in a more complex model (ensemble)*

Focal Loss Implementation:

- $y \in \{0, 1\}$ — true label
- $\hat{p} \in (0, 1)$ — predicted probability of the true class
- $\gamma > 0$ — focusing parameter (usually $\gamma = 2$)

$$\text{Focal Loss} = -(1 - \hat{p}_t)^{\gamma} \cdot \log(\hat{p}_t)$$

Where:

- $\hat{p}_t = \begin{cases} \hat{p} & \text{if } y = 1 \\ 1 - \hat{p} & \text{if } y = 0 \end{cases}$

Intuition & Visual Explanation

If the model **predicts correctly** and is **confident** (e.g., $\hat{p}_t = 0.99$),

$$(1 - \hat{p}_t)^{\gamma} = (1 - 0.99)^2 = 0.0001$$

➤ The loss is tiny — we don't focus on easy examples.

If the model is **wrong** and confident (e.g., predicted 0.99 but true label was 0),

$$\hat{p}_t = 1 - 0.99 = 0.01, \quad (1 - \hat{p}_t)^{\gamma} = 0.99^2 = 0.9801$$

➤ The loss is **very high** — model learns from this mistake.

If the model is **unsure** (e.g., $\hat{p}_t = 0.5$),

$$(1 - 0.5)^2 = 0.25$$

➤ The loss is moderate — model still focuses on it.

Quiz: Why Use Focal Loss Loss in March Madness ?

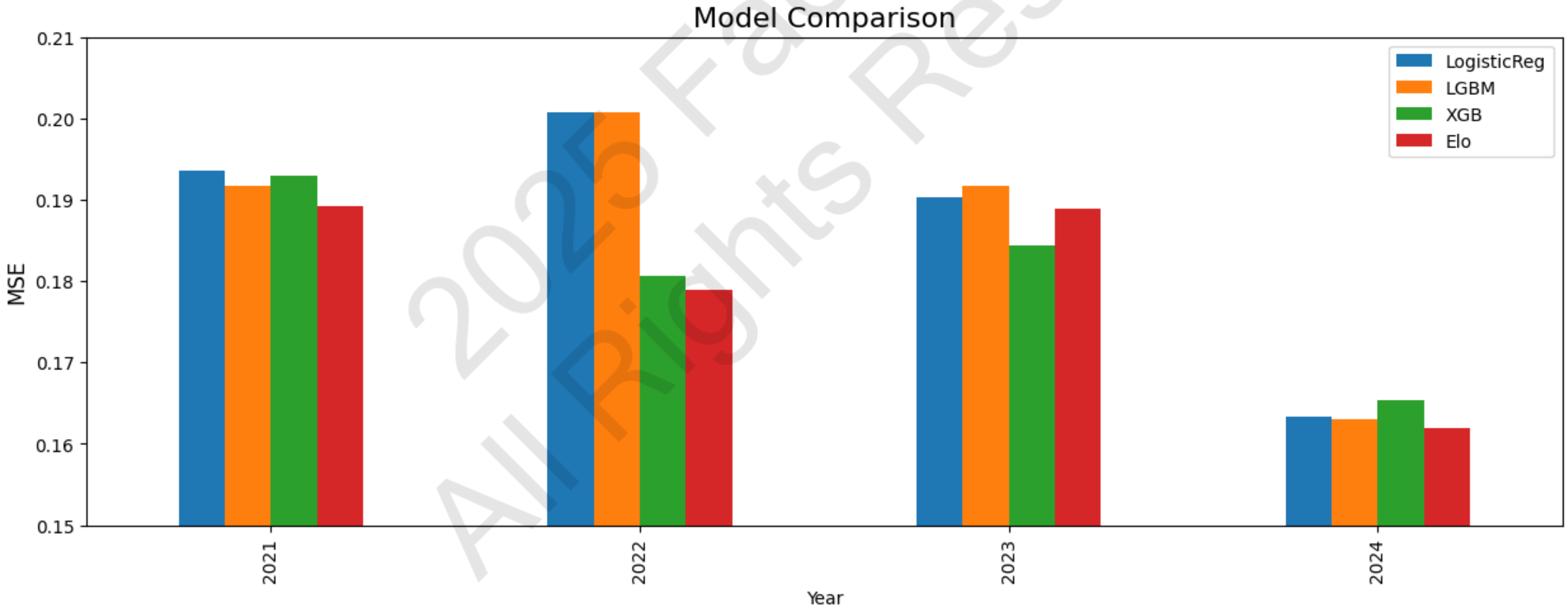
- **Why We Used Focal** March Madness games are **imbalanced**: most are easy (1-seed vs 16-seed), but some are true toss-ups or **upsets**.
- Focal loss **down-weights easy games** and emphasizes learning from **hard matchups** (e.g., 8 vs 9 or 12 vs 5).
- It **prevents the model from getting lazy** by just learning the dominant cases.
- In our BiLSTM, it helped learn from rare but important cases where lower seeds beat higher seeds.

Ensemble Modeling (Voting Regressor)

- **Ensemble Components:** We trained several classic ML models:
- Extra Trees Regressor (Extremely Randomized Trees)
- Random Forest
- Gradient Boosted Trees (XGBoost, LightGBM, CatBoost)
- **Meta-Model:** Used a **VotingRegressor** to average predictions from all models.
- Treated win probability as a regression target (0 to 1) and averaged outputs
- Assigned higher weights to models that performed better on validation
- **Tuning:** Performed hyperparameter optimization on each model (using Optuna) and selected the best ensemble combination via cross-validation
- **Rationale:** Ensembling leverages different models' strengths and reduces overfitting – improves stability of predictions

Model Performance Comparison: Elo vs ML Models

- Compare multiple model types used in the project
- Visualizes model accuracy across seasons
- MSE shown for each model: lower is better
- Models included:
- Logistic Regression
- LGBM
- XGBoost
- Elo Ratings (non-ML baseline)

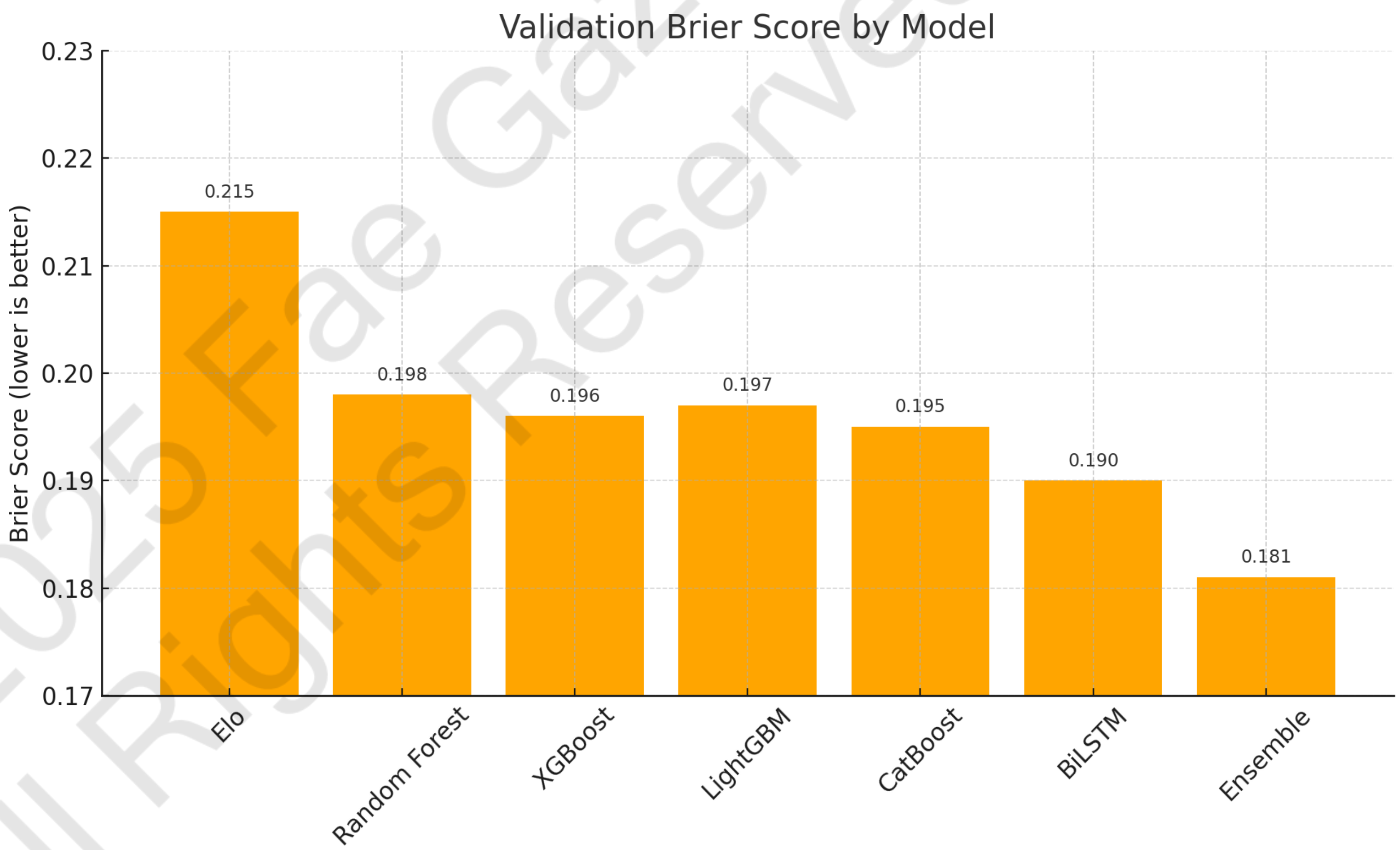


Poll: Which Model Performed Best?

- *Based on our validation tests, which approach do you think achieved the **lowest** (best) Brier Score?*
  - A. **Elo Rating** baseline model
  - B. **BiLSTM** deep learning model
  - C. **Ensemble** of all models
  - D. **XGBoost** (best single tree model)

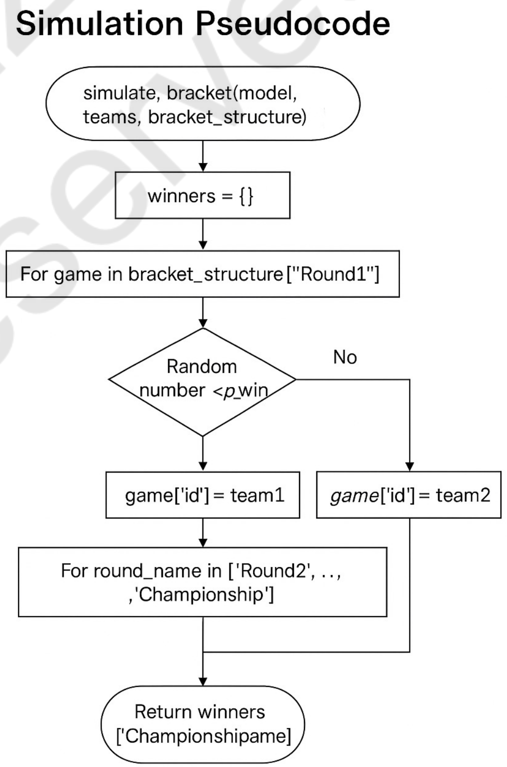
Model Performance Comparison

Elo: 0.215
Random Forest: 0.198
XGBoost: 0.196
LightGBM: 0.197
CatBoost: 0.195
BiLSTM: 0.190
Ensemble: **0.181**



Tournament Bracket Simulation (Monte Carlo)

Each game simulated based on predicted win probabilities.
Final outcomes aggregated over thousands of iterations



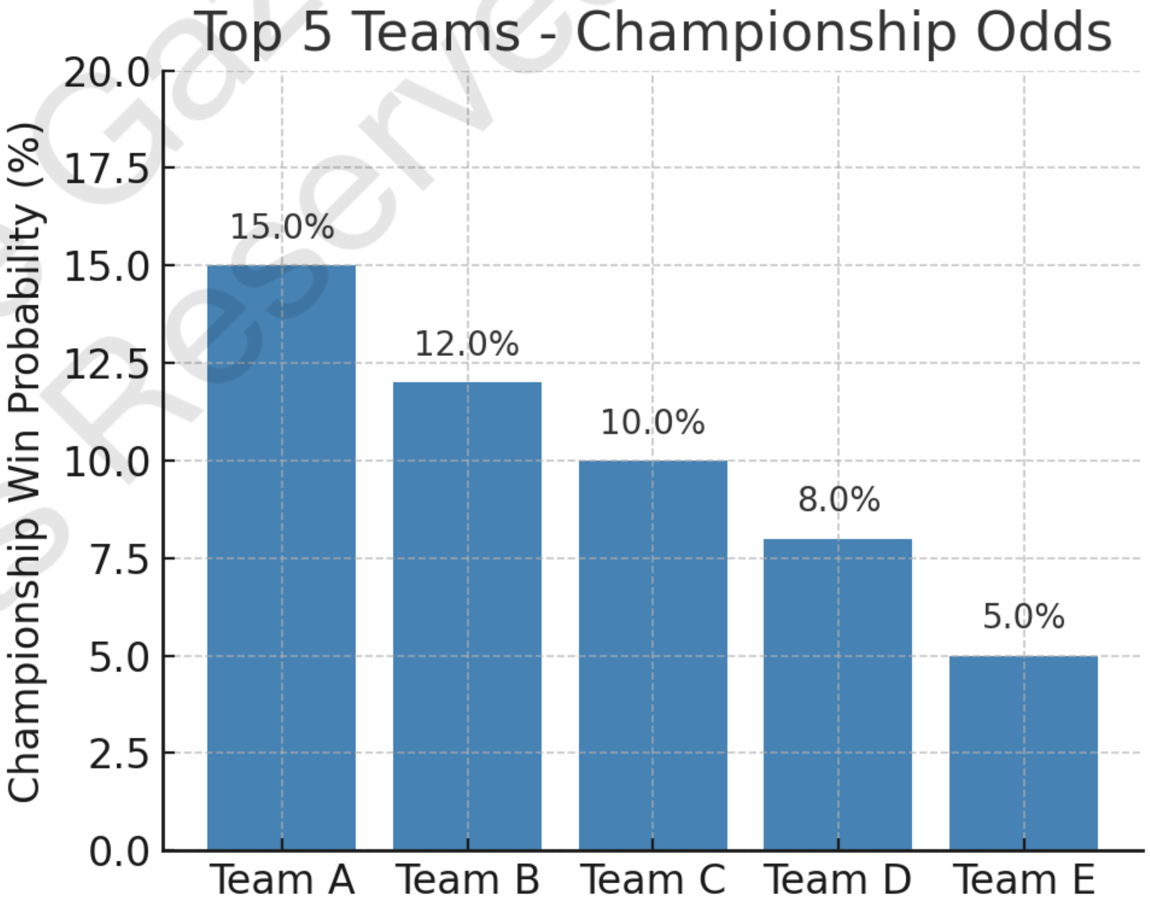
For simulator: iterate (poisson)
or
Repeat simulation N times to estimate probabilities

Tournament Simulation Utility

- **Purpose:** Translate predicted per-game win probabilities into full tournament outcomes
- **Monte Carlo Simulation:** We wrote a simulator to play out the entire bracket many times (e.g., 100,000 random brackets)
  - For each simulated bracket, each game winner is sampled according to our predicted win probability for that matchup
- **Outputs:** Estimated championship probabilities for each team, distribution of how far each team advances, etc.
- **Use Cases:**
  - Evaluate the uncertainty: which top seed has the highest chance to win it all?
  - Identify potential Cinderellas (lower seeds with non-negligible Final Four odds)
  - Engage users with bracket predictions (gamification)
- **Utility:** Provided insights beyond single-game predictions — e.g., probability that no #1 seed makes Final Four, etc.

Results: Championship Probabilities

•**Most Likely Champion:** *Team A* ~15% chance – highest among all teams, but still low (85% chance someone else wins)
•**Other Contenders:** *Team B* ~12%, *Team C* ~10%, *Team D* ~8%, etc. – several teams in the mix
•**Field vs Favorite:** Cumulatively, even the top 5 teams only sum to ~50%; the rest of the field had ~50% total championship probability
•**Insight:** The tournament is wide-open – even top seeds are far from guaranteed, aligning with historical Madness (upsets do happen)



Interactive: Bracket Prediction Challenge

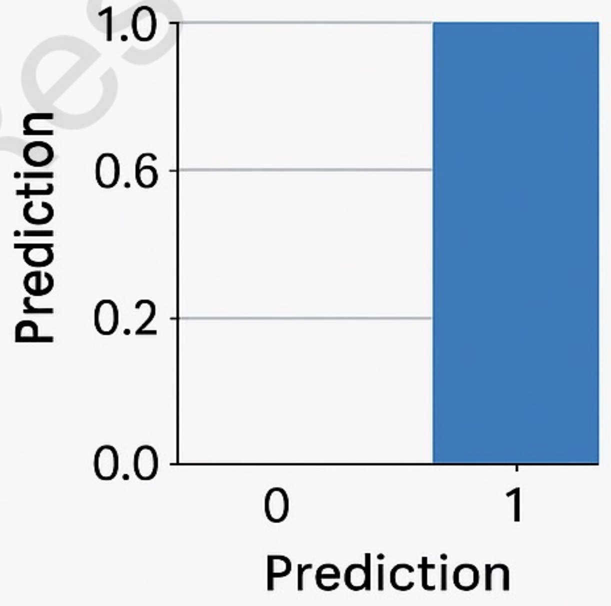
- **Audience Challenge:** Who would *you* pick as the champion?
  - Option 1: *Team A* (Pre-tournament #1 overall seed, model's top pick ~15%)
  - Option 2: *Team B* (Another high seed with strong stats)
  - Option 3: *Team C* (Historically strong program, moderate seed)
  - Option 4: *Field* (Any other team outside the top 3)
- **Discuss:** What factors influenced your pick? (Personal bias? Upset intuition? Model probabilities?)

Evaluation Metrics

Brier Score: The mean squared error of probability forecasts

- Example: if predicted win probability = 0.8 and actual outcome = 1 (win), contribution = $(0.8 - 1)^2 = 0.04$
- Range 0 to 1 (perfect = 0, worst = 1 if predictions were completely wrong with full confidence)
- Mean Squared Error (=]), [illegible]maily, measures squared error – in our case, Brier Score is a MSE on a 0/1 outcome

$$\text{Brier} = (p - o)^2$$



Key Lessons Learned

- **Elo & Simplicity:** Simple rating systems (Elo) provide a **strong baseline** for sports outcomes
- . Complex models must add real value to beat them on small datasets.
- **Feature Engineering Matters:** Domain-specific features (seeds, recent performance, etc.) were crucial. Properly capturing team strength in features often trumped fancy algorithms.
- **Ensemble Power:** Combining models was effective – different algorithms compensated for each others' biases. The ensemble performed better than any single model.
- **Deep Learning Challenges:** Our BiLSTM model was useful, but limited data means deep nets can easily overfit. Careful loss functions (focal loss) and architecture helped, but traditional models were competitive.
- **Calibration & Probability Focus:** We optimized for Brier Score, which taught us to focus on probabilistic calibration, not just accuracy. Our final model wasn't just choosing winners, but assigning sensible confidence levels to each prediction.

Future Work & Improvements

- **More Data / Features:** Incorporate player-level data or advanced metrics (e.g., KenPom ratings, injuries, head-to-head records) to further inform the model
- **Transfer Learning:** Use pre-training on many seasons for the LSTM or try Transformer-based models to capture team interactions across seasons
- **Dynamic Updates:** Update Elo and model probabilities round-by-round during the tournament (ingest actual results as they come in to improve predictions for remaining games)
- **Ensemble Diversity:** Include more diverse models (e.g., a logistic regression or Bayesian model for calibration) to complement current ensemble
- **Simulation-based Optimization:** Use the bracket simulator to optimize strategies (for example, maximizing expected points in a bracket pool by picking certain calculated upsets)
- **Generalization:** Apply the pipeline to other sports tournaments (e.g., soccer World Cup, which also has group and knockout stages) to test its versatility

Competition Results & Final Leaderboard

- **Kaggle Leaderboard Performance:** Final Brier Score = **0.18056** on 2025 tournament (lower is better)
- **Ranking:** Placed 42nd out of 821 teams (Top 5%) – secured a Silver Medal (if top 5%)
- **Comparison:** Top score on leaderboard was ~0.165; median was ~0.25 (our model far exceeded baseline)
- **Notebooks:** *(Links to solution code)* Our work was divided into:
    - *EDA & Elo Analysis Notebook* – data exploration, seed parsing, Elo rating computation
    - *BiLSTM Model Notebook* – deep learning model training with focal loss
    - *Ensemble & Simulation Notebook* – training tree models, creating ensemble, tournament simulations and analysis
- **Acknowledgments:** Kaggle community for insightful discussions, and NCAA for the rich historical data

Thank You!

- Q&A
- Link to slides + app
- Let’s simulate your picks!
- **Speaker Notes:** Wrap up. Offer demo or app link. Thank the audience and open the floor.