Three-Staged Predictive Model for NCAA Men's Basketball

Henry Huang

Abstract

A modeling stack for NCAA basketball that (i) converts box scores to tempo-free metrics, (ii) estimates team strength via three engines (elo, Bradley–Terry, and Massey least squares), (iii) solves a two-way league system for schedule-adjusted offense/defense, and (iv) maps multi-source signals to calibrated win probabilities using a logistic regression. We evaluate with LogLoss, Brier score, calibration error, and ROC/AUC, and also describe the adjusted OE–DE landscape and player usage/efficiency.

1 Notation

Symbol	Meaning
$\mathrm{FGA}, \mathrm{FGM}, \mathrm{3PM}, \mathrm{FTA}, \mathrm{OREB}, \mathrm{DREB}, \mathrm{TO}$	Box-score counts
Poss	Possessions: $\mathrm{FGA} + 0.475\,\mathrm{FTA} - \mathrm{OREB} + \mathrm{TO}$
$\mathrm{OE}_g, \mathrm{DE}_g$	Game efficiencies per 100 possessions
μ	League-average offensive efficiency per 100
R_i	Elo rating of team i
H	Home-court Elo offset in rating points (e.g., $H \approx 65$; neutral 0)
K_0, λ	Elo base step size and MOV scaling factor
m	Margin of victory (home points − away points)
θ	Logistic scaling $\ln(10)/400$ for Elo link
β_i	Bradley–Terry log-strength for team i
γ	BT home parameter in log-odds
r_i	Massey point-differential power rating for team i
o_i, d_i	Schedule-adjusted offense and defense adjustments for team i
$\mathrm{AdjOE}_i, \mathrm{AdjDE}_i$	Adjusted offense/defense: $\mu + o_i$ and $\mu - d_i$
Net_i	Adjusted net rating $= o_i + d_i$ (per 100)
x	Feature vector for a game (differentials, deltas, venue, Log5)
$\alpha, \boldsymbol{\beta}$	Intercept and coefficients of the matchup logistic regression
p	Predicted home win probability, $p = \sigma(\alpha + \boldsymbol{\beta}^\top x)$
w_g	Time-decay weight for game g, $w_g = \exp(-\ln 2 \cdot \Delta\mathrm{days}/h_{1/2})$
λ_{logit}	Tikhonov penalty for the logistic model

2 Features

Possessions:

$$\mathrm{Poss} = \mathrm{FGA} + 0.475\,\mathrm{FTA} - \mathrm{OREB} + \mathrm{TO}. \tag{1}$$

Efficiencies per 100:

$$\mathrm{OE}_g = 100 \cdot \frac{\mathrm{Points}_g}{\mathrm{Poss}_g}, \qquad \mathrm{DE}_g = 100 \cdot \frac{\mathrm{OppPoints}_g}{\mathrm{OppPoss}_g}. \tag{2}$$

Four-factor rates:

$$\mathrm{eFG\%} = \frac{\mathrm{FGM} + 0.5\,\mathrm{3PM}}{\mathrm{FGA}}, \quad \mathrm{TOV\%} = \frac{\mathrm{TO}}{\mathrm{Poss}}, \quad \mathrm{OREB\%} = \frac{\mathrm{OREB}}{\mathrm{OREB} + \mathrm{OppDREB}}, \quad \mathrm{FTR} = \frac{\mathrm{FTA}}{\mathrm{FGA}}. \tag{3}$$

Empirical-Bayes shrinkage (stabilizes small samples) for cumulative P points on N possessions with prior π and weight w:

$$\widetilde{\mathrm{OE}} = 100 \cdot \frac{P + \pi w}{N + w}. \tag{4}$$

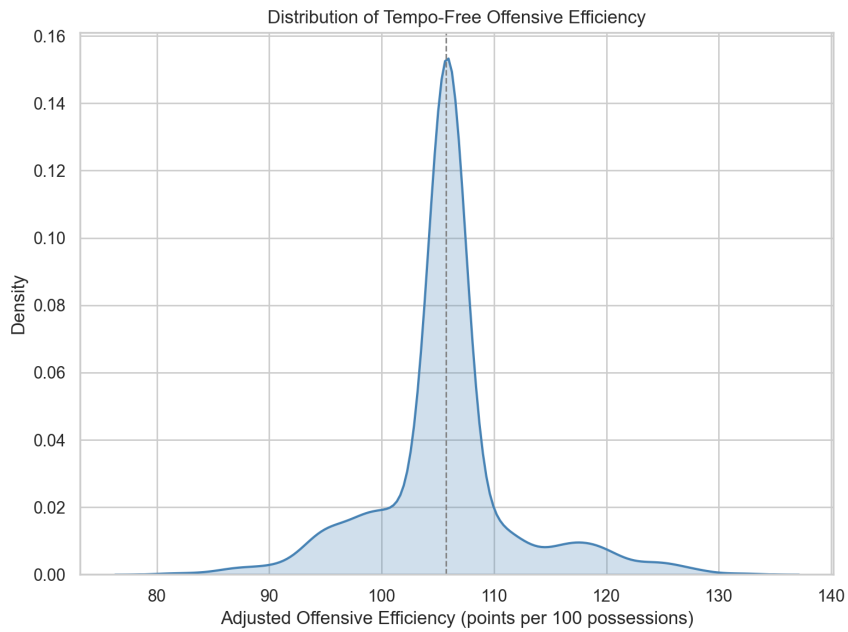

Figure 1: Tempo-free offensive efficiency distribution based on league mean μ

3 Elo, Bradley–Terry, & Massey Models

Elo is an win-probability tracker, Bradley–Terry (BT) is a paired-comparison model for odds on a neutral/venue-adjusted court, and Massey is a least-squares model for scoring margins.

3.1 Elo with Margin-of-Victory Scaling

Let $R_i \in \mathbb{R}$ be team i's Elo. For a home team h vs away a,

$$E_h \;=\; \frac{1}{1+10^{-(R_h+H-R_a)/400}} \;=\; \sigma\big(\theta[(R_h+H)-R_a]\big), \quad \theta = \ln(10)/400. \tag{5}$$

Here H is a home-court offset (0 on neutral). The logistic link arises from a constant-variance normal additivity assumption on latent strengths.

Given outcome $S_h \in \{0,1\}$ and margin m (home points minus away points),

$$K(m) = K_0\Big(1+\lambda\ln(1+|m|)\Big), \tag{6}$$

$$R_h \leftarrow R_h + K(m)\,(S_h - E_h), \quad R_a \leftarrow R_a + K(m)\,\big((1-S_h)-(1-E_h)\big). \tag{7}$$

The $\ln(1+|m|)$ term increases learning for decisive games but avoids runaway updates on blowouts.

Figure 2 shows seasonal trajectories, and the widening gaps imply that persistent performance differences across different teams exist. Figure 3 plots final Elo vs. empirical win%, which should be monotone and near-logistic.

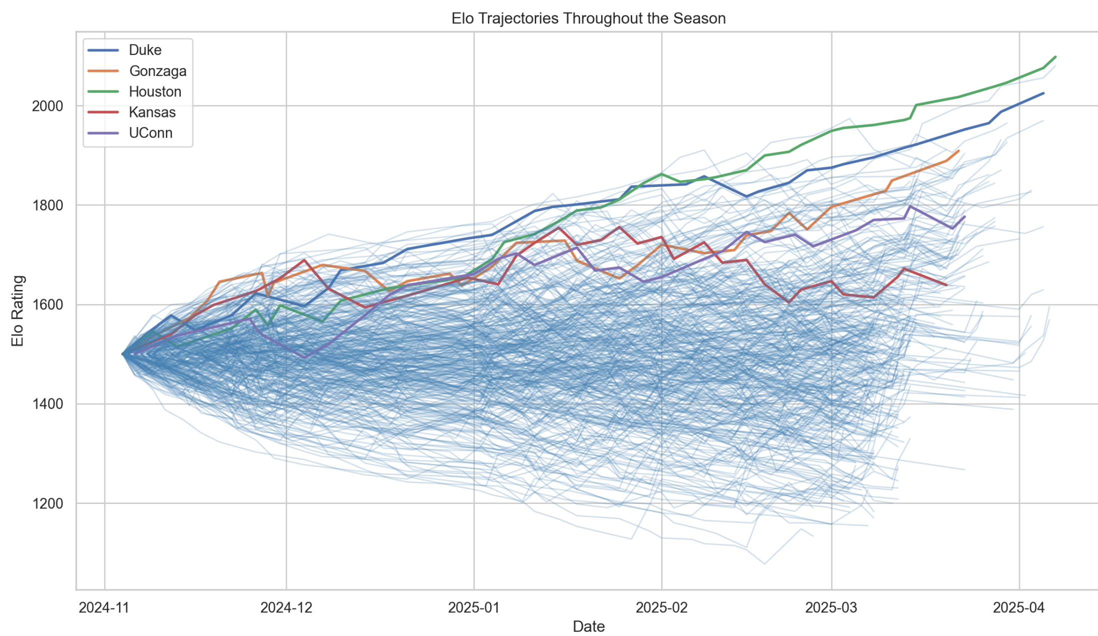

Figure 2: Elo rating trajectories of NCAAM teams

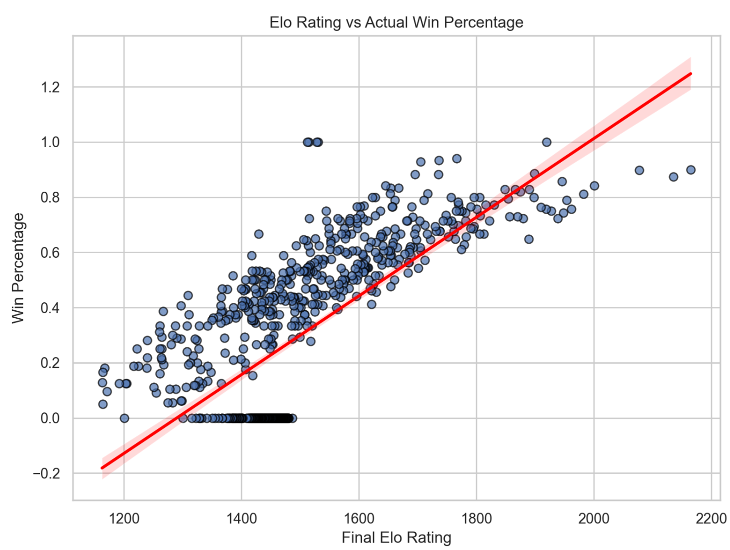

Figure 3: Elo vs. actual win percentage

3.2 Bradley–Terry Model

Bradley–Terry posits that the log-odds of i beating j is a difference of team abilities plus a venue term:

$$\log \frac{p_{ij}}{1 - p_{ij}} \;=\; \beta_i - \beta_j + \gamma \cdot \mathrm{Home}_i, \qquad \sum_i \beta_i = 0. \tag{8}$$

We solve a Tikhonov (or ridge-) penalized and time-decayed MLE:

$$\max_{\boldsymbol{\beta},\gamma} \sum_g w_g \Big[y_g \log p_g + (1 - y_g)\log(1 - p_g) \Big] \;-\; \frac{\alpha}{2} \|\boldsymbol{\beta}\|_2^2, \tag{9}$$

with $w_g = \exp(-\ln 2 \cdot \Delta\text{days}_g / h_{1/2})$ (half-life $h_{1/2}$) and $p_g = \sigma(\beta_{i(g)} - \beta_{j(g)} + \gamma \text{Home}_{i(g)})$. Ridge improves reliability when schedules are nearly disconnected.

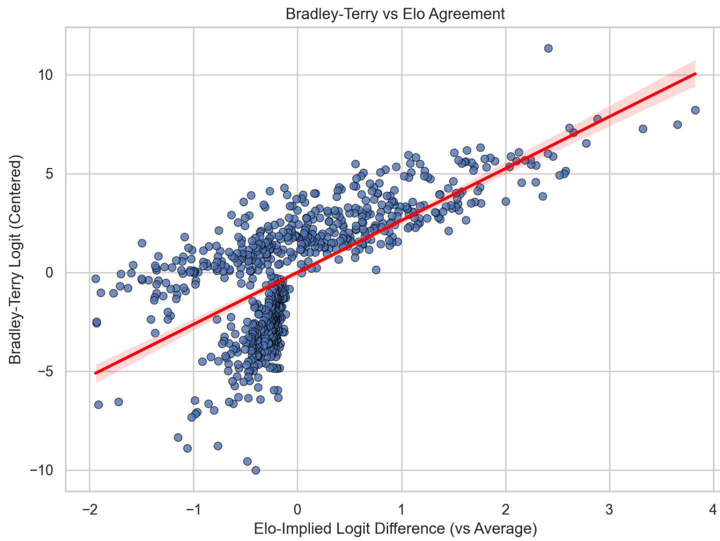

Figure 4: Bradley–Terry vs Elo-implied log-odds

3.3 Massey Least-Squares

Let M_{ij} be the observed scoring margin (home minus away). The Massey model assumes

$$M_{ij} \;=\; r_i - r_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \qquad \sum_i r_i = 0, \tag{10}$$

so each game contributes a row to $Ar = b$ with a $+1$ at the home team, -1 at the away team. We solve $\min_r \|Ar - b\|_2^2$ subject to $\sum_i r_i = 0$ (or equivalently, replace one equation by the constraint).

Margins carry more information than binary outcomes in balanced matchups and produce ratings on a point scale, which is directly useful for spread-like interpretations and for the schedule-adjusted system in §4. Figure 5 shows residuals are bell-shaped and centered which supports the assumption that it's linear-Gaussian. Large positive r_i shows the teams that consistently outscore opponents even when facing difficult matchups/schedules.

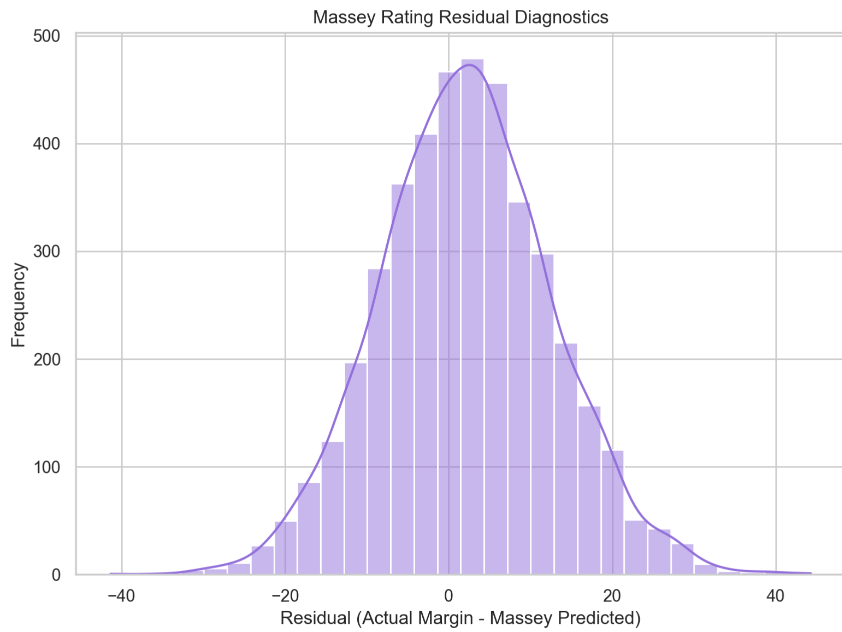

Figure 5: Massey residuals

4 Two-Way Schedule Adjustment for Offense & Defense

League mean:

$$\mu = 100 \cdot \frac{\sum_g \mathrm{Points}_g}{\sum_g \mathrm{Poss}_g}. \tag{11}$$

Directional equations per game (home h, away a):

$$(o_h - d_a) \approx \mathrm{OE}_h - \mu, \qquad (o_a - d_h) \approx \mathrm{OE}_a - \mu, \tag{12}$$

with constraints $\sum_i o_i = \sum_i d_i = 0$. Stack $Ax \approx y$, $x = [o; d]$, ridge solution

$$\hat{x} = (A^\top A + \lambda I)^{-1} A^\top y, \quad \mathrm{AdjOE}_i = \mu + o_i,\ \mathrm{AdjDE}_i = \mu - d_i,\ \mathrm{Net}_i = o_i + d_i. \tag{13}$$

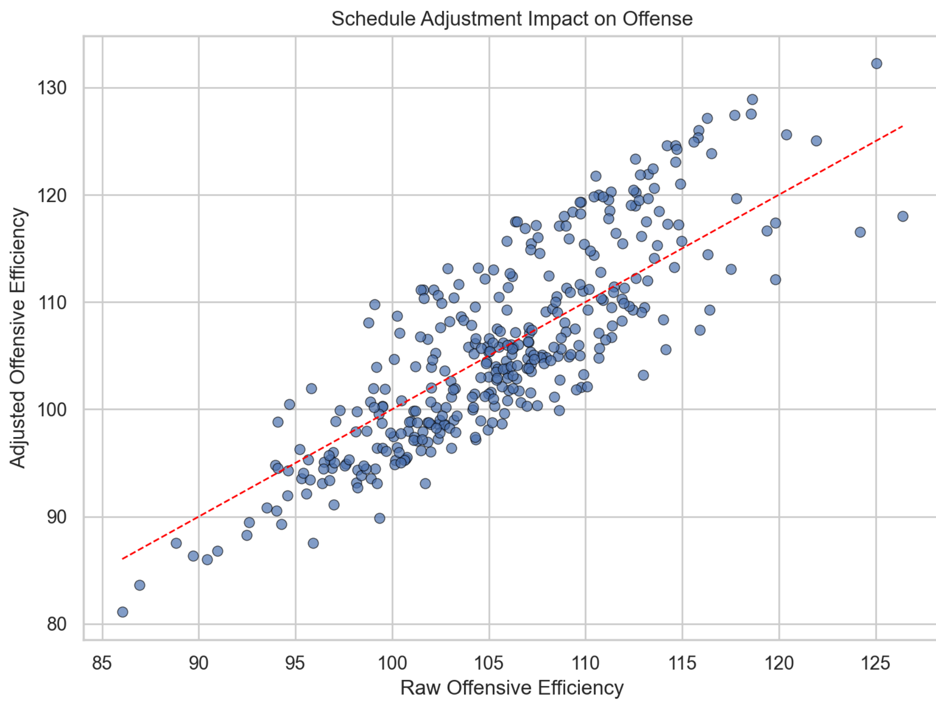

Figure 6: Adjusted vs raw OE. Above $y=x$: teams suppressed by strong schedules.

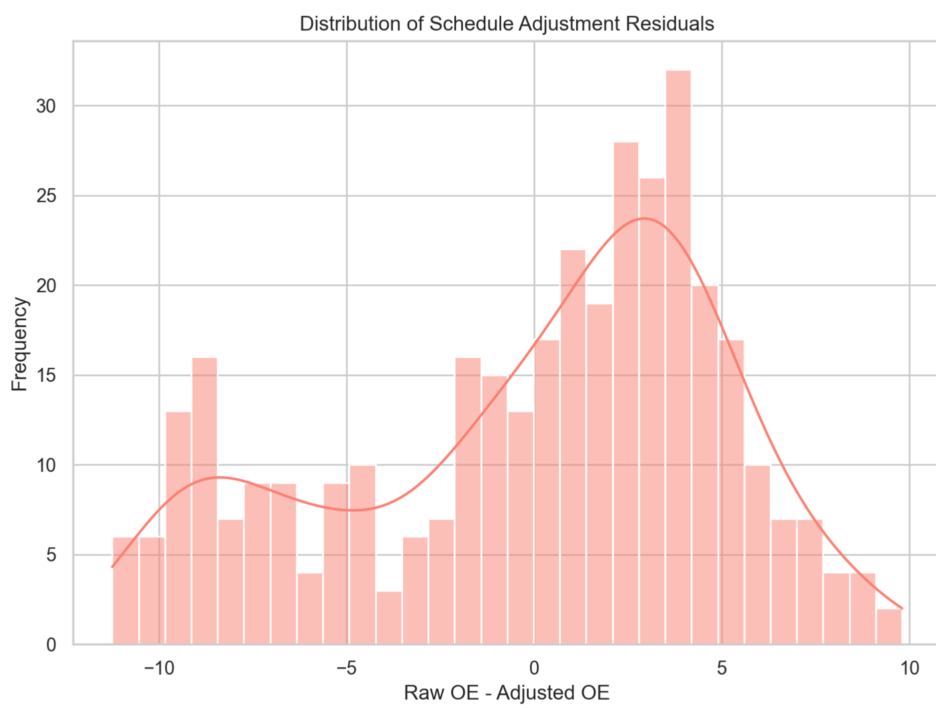

Figure 7: Adjustment residuals $Y - \hat{\mu} - \hat{o} + \hat{d}$: approximately normal.

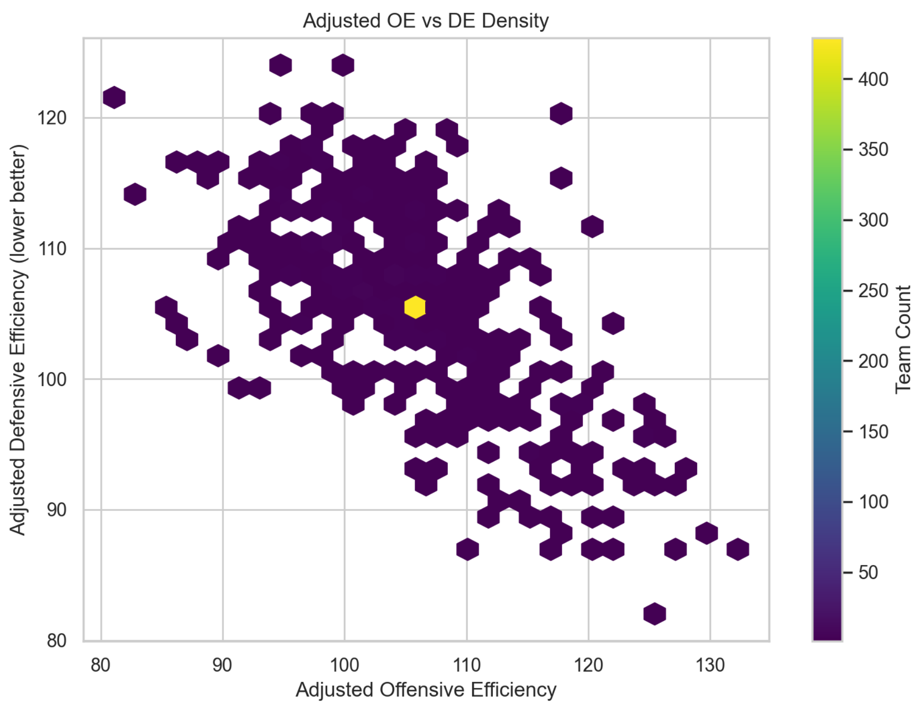

Figure 8: Adjusted (AdjOE, AdjDE) density: offense-heavy (right), defense-heavy (down).

5 Matchup Meta Model

5.1 Features

- Engine differentials: $\Delta R = (R_h + H) - R_a$, $\Delta\beta = \beta_h - \beta_a$, $\Delta r = r_h - r_a$.
- Adjusted efficiencies: ΔAdjOE, ΔAdjDE, ΔNet.
- Four-factor deltas (seasonal and rolling windows: last 3/7).
- Rest differential (days), venue flags (home/neutral), Log5 prior:

$$\text{Log5}(h{:}a) = \frac{w_h(1-w_a)}{w_h(1-w_a) + (1-w_h)w_a}.$$

All continuous features standardized using train-only mean/SD.

5.2 Objective, weighting, and inference

$$p_g = \sigma(\alpha + \boldsymbol{\beta}^\top x_g), \tag{14}$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[-\sum_g w_g \big(y_g \log p_g + (1-y_g)\log(1-p_g) \big) + \lambda_{\text{logit}} \|\boldsymbol{\beta}\|_2^2 \right], \tag{15}$$

$$w_g = \exp\left(-\ln 2 \cdot \frac{\Delta\text{days}_g}{h_{1/2}} \right). \tag{16}$$

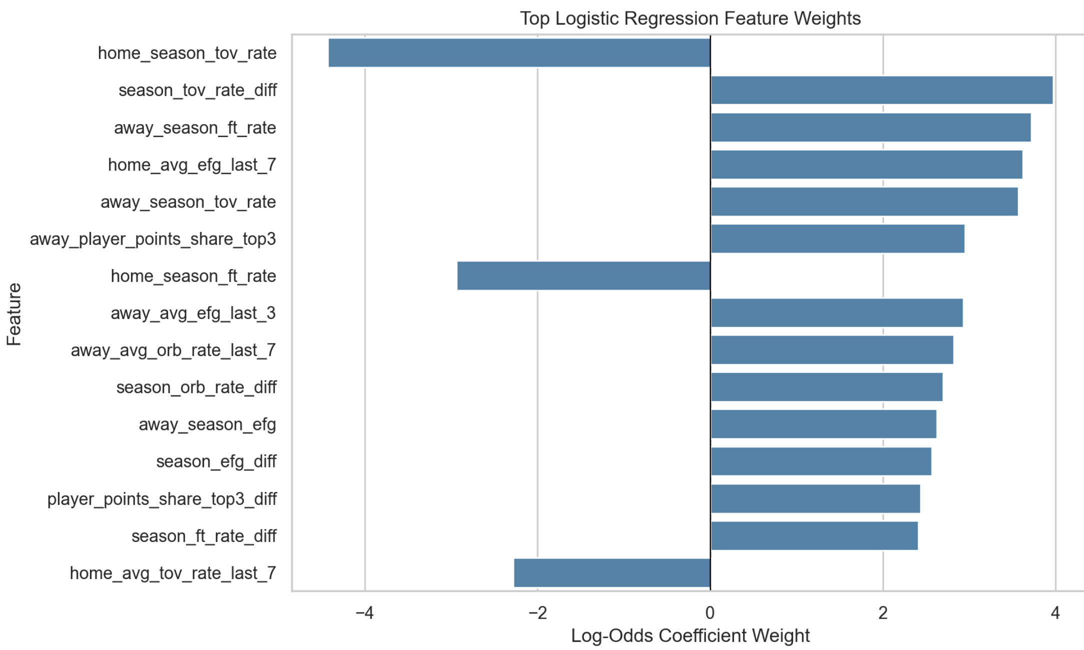

Figure 9: Meta regression coeffs

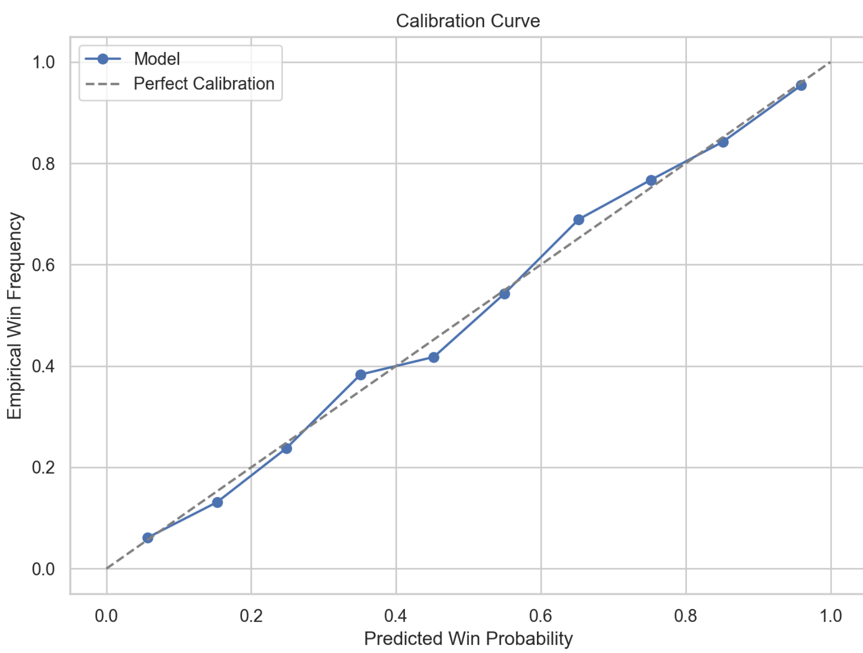

Figure 10: Calibration curve, predicted p vs empirical win frequency

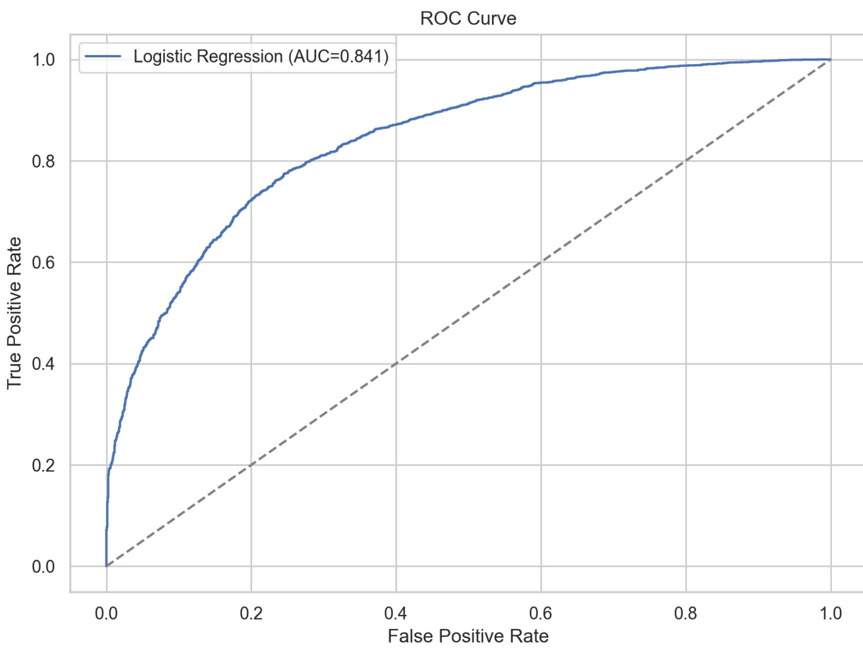

Figure 11: ROC curve. AUC summarizes ranking ability across thresholds

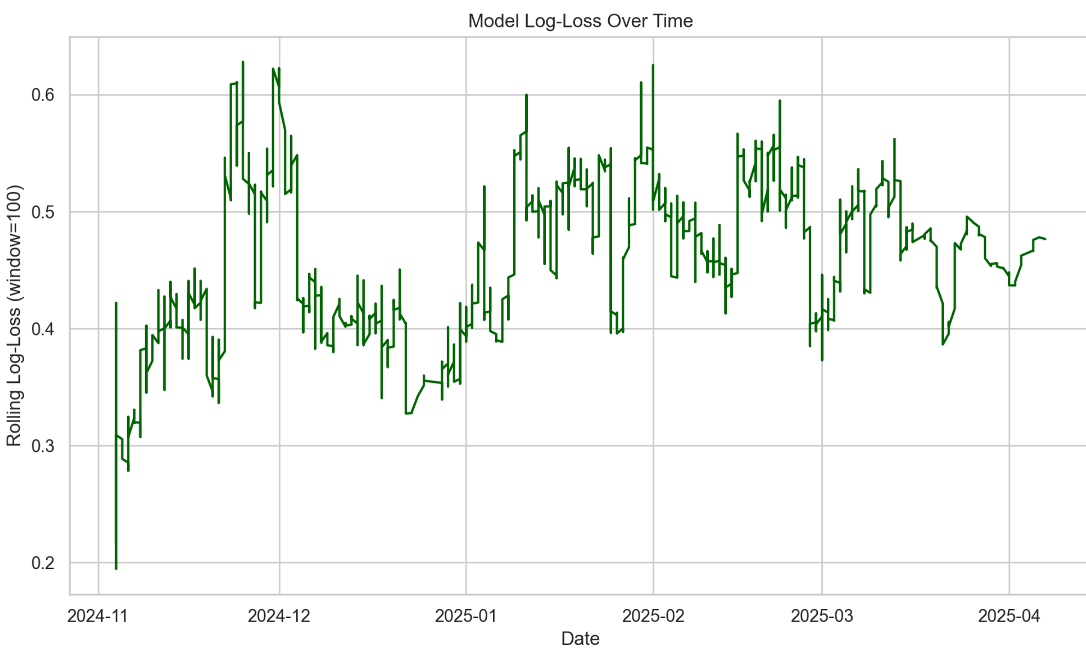

Figure 12: Rolling Logistic Loss

6 Player and Lineup Context

Per-40 scaling and usage proxy:

$$\text{Pts}/40 = 40 \cdot \frac{\text{pts}}{\text{min}}, \quad U = \frac{\text{FGA} + 0.44\,\text{FTA}}{40}, \quad \text{eFG\%} = \frac{\text{FGM} + 0.5\,\text{3PM}}{\text{FGA}}.$$

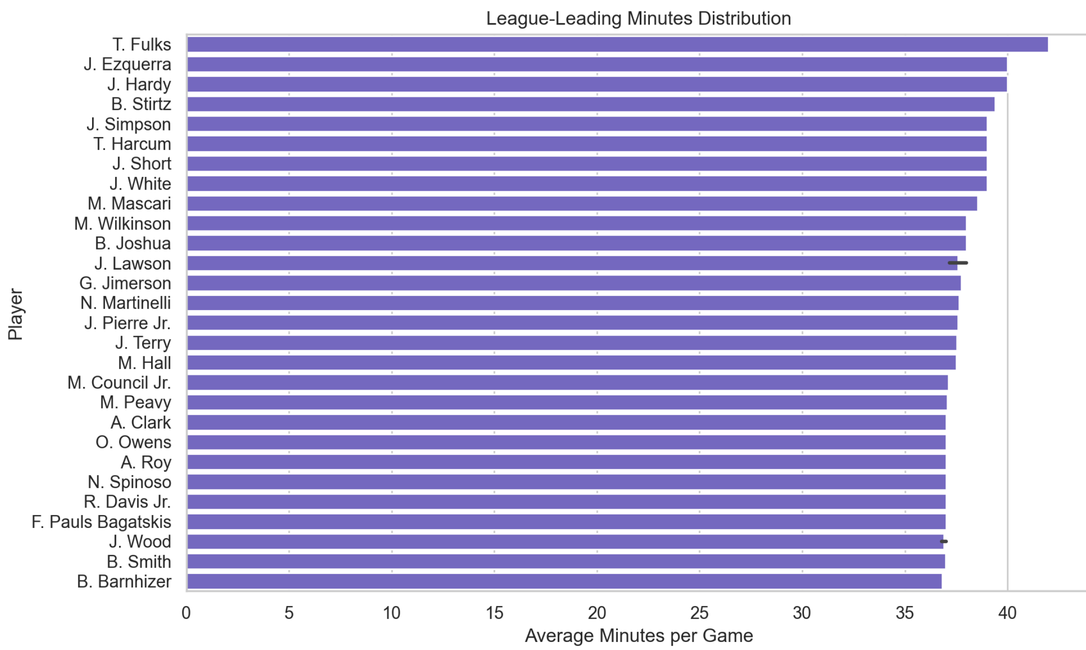

Figure 13: Minutes distribution, rotation leverage & depth

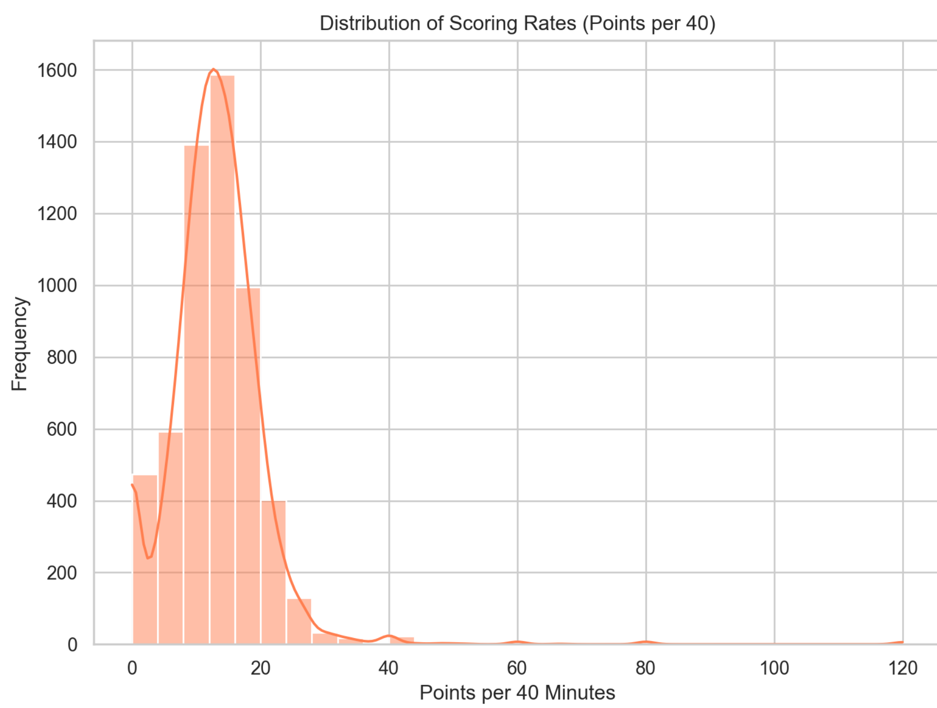

Figure 14: Points per 40, right-skewed

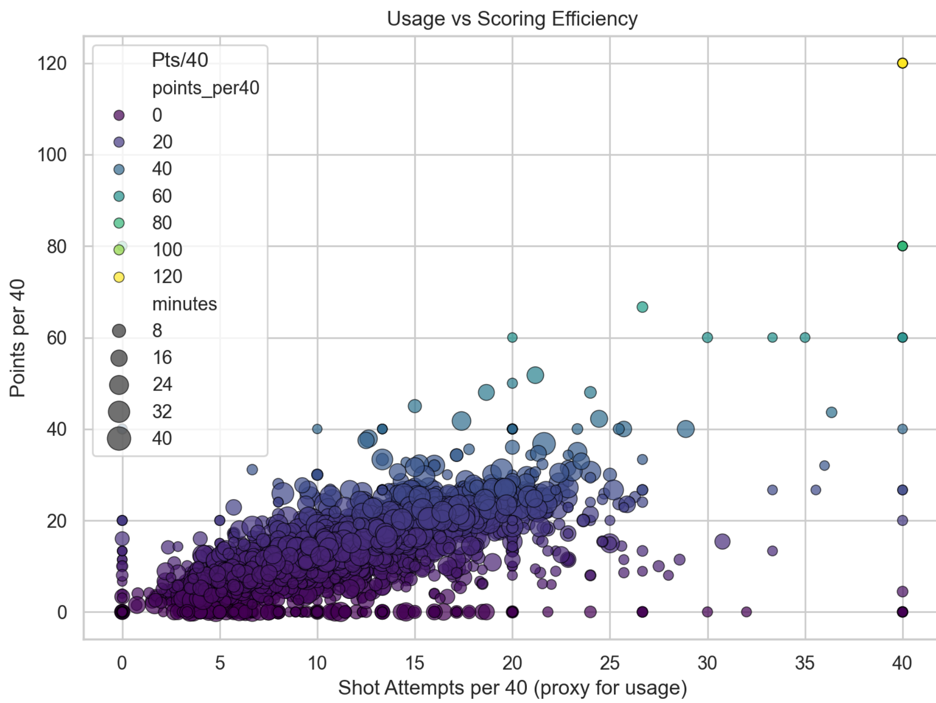

Figure 15: Usage vs points/40 (bubble=size minutes): diminishing returns at extreme usage.

7 Training Hyperparameters & Evaluation Metrics

Elo: $H \in [50, 75]$ grid; $K_0 \in [16, 32]$; $\lambda \in [0.1, 0.4]$ for MOV.
BT: Ridge $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}\}$; time-decay half-life $h_{1/2} \in \{90, 120, 150\}$ days.
Massey: Unregularized LS with zero-sum constraint; diagnostics on residuals.
Meta regression: λ_{logit} by val grid; $h_{1/2} \in \{90, 120, 150\}$; class weights for imbalance.
Logarithmic Loss: $-\mathbb{E}[y \log p + (1-y)\log(1-p)]$.
Brier score: $\mathbb{E}[(y-p)^2]$, decomposed into reliability (calibration), resolution, uncertainty.
Calibration: ECE/MCE over 10–20 equal-frequency bins.
Discrimination: ROC/AUC.

8 Biases & Data

We make five assumptions: 1) tempo invariance: points per possession provide an unbiased measure of scoring efficiency, 2) comparative strength: log odds of winning can be modeled as approximately linear in the difference in strength between two teams, 3) margin model: scoring differentials follow a linear relationship with point-differential strengths, 4) two-way separability: observed offensive efficiency can be decomposed additively into a team's offensive contrib and its opponent's defensive influence both centered around the mean efficiency of the entire NCAA μ, 5) non-stationarity: older games convey less power, down-weighted exponentially by time decay.

The data used are full regular-season box scores and play-by-play moves for all Division I NCAA men's basketball teams. The data preprocessing progresses as the following: (1) the dataset is split into training (Nov-Jan), validation (Feb), and tests (Mar); (2) all feature standardization parameters are trained exclusively on the training split and applied unchanged; (3) rolling statistics such as last 3 or last 7 game averages are computed using only games occurring before the prediction date.