NCAA Basketball Point Spread Prediction Model

Team CMMT: Caleb Han, Mason Mines, Mason Wang, Tony Wang

Model Architecture

We developed an ensemble model combining Ridge Regression (30%) and LightGBM (70%) to predict point spreads:

$$\text{Predicted Spread} = 0.3 \times \text{Ridge}_{\text{pred}} + 0.7 \times \text{LightGBM}_{\text{pred}}.$$

Ridge provides a stable, interpretable linear baseline, while LightGBM captures complex non-linear patterns. The ensemble balances model transparency with predictive accuracy.

Data Sources & Features

Our model uses **11 features** derived from two primary data sources:

- **Barttorvik efficiency ratings**[1]: Adjusted Offensive/Defensive Efficiency (AdjOE/AdjDE), measured as points per 100 possessions and adjusted for opponent strength. Efficiency Margin (AdjEM) = AdjOE−AdjDE.
- **Elo ratings**[2]: FiveThirtyEight-style Elo with K-factor = 38, home-court advantage = 4.0 points, and 64% season carryover to conference average.

Complete feature set (11):

- **Team efficiencies (6)**: home/away AdjOE, AdjDE, AdjEM.
- **Elo features (3)**: home/away Elo, Elo differential.
- **Derived features (2)**: efficiency differential, Elo-based spread.

Training Data

Trained on **8,850 NCAA games** from 2020–2025 (6 seasons). Games were processed chronologically for accurate Elo histories and to avoid data leakage. Includes all major conferences.

Model Evaluation

We used **5-fold time-series cross-validation** to respect temporal ordering, evaluating on future games to mimic real-world prediction.

Performance metrics:

Model	MAE (points)	RMSE (points)
Ridge	6.024 ± 0.183	7.92
LightGBM	4.82 ± 0.31	6.45
Ensemble	**5.535 ± 0.281**	**7.21**

The ensemble achieves MAE of **5.5 points** (typically within 5–6 points of actual margin) with lower variance, demonstrating improved robustness.

Key Decisions

- **Feature selection:** Advanced features (Four Factors, momentum) decreased performance (MAE 5.017 vs. 5.001). Retained 11 baseline features. AdjOE/AdjDE and Elo were strongest predictors.
- **Ensemble weighting:** LightGBM 70% (accuracy), Ridge 30% (stability). Optimized via validation.
- **Hyperparameters:** Ridge $\alpha = 1.0$; LightGBM: 100 trees, depth = 8, rate = 0.1. Tuned via CV.

Code: github.com/calebyhan/triangle-sports-analytics-26

[1] barttorvik.com

[2] fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/