Predicting NBA Game outcomes using Elo ratings and machine learning

Juan Luis Toledo Gómez

Home court advantage is a well-documented phenomenon in professional basketball, with home teams winning approximately 60% of games historically. Here we investigate whether simple machine learning models can predict NBA game outcomes using team strength ratings (Elo scores) as the primary predictor. Using a dataset of 126,314 games ranging from 1946 to 2015, we train a logistic regression model on historical data and evaluate its performance on held-out test sets. Our results show that a pure Elo-based model achieves 66% accuracy on modern games (2010-2015), beating a naive baseline of 59% by 7%.

1. Introduction

1.1 Background

Home court advantage is a well-established phenomenon across professional sports. In the National Basketball Association (NBA), home teams have historically won approximately 60% of regular season games. This advantage has been attributed to factors including crowd support, travel fatigue, familiarity with local conditions, and referee bias.

The ability to predict game outcomes has practical applications in sports analytics, betting markets, and team management. Several approaches have been proposed, from simple heuristics to complex machine learning models. Among these, Elo rating systems (originally developed for chess) have proven surprisingly effective for predicting sports outcomes.

1.2 Research Question

This project investigates the following question: Can we predict NBA game outcomes using only team strength ratings (Elo scores) as input to a machine learning classifier?

1.3 Contributions

The primary contributions of this paper are:

- A comprehensive evaluation of logistic regression on historical NBA game data over 70 years.
- Analysis of how generalization from past eras affects model performance (full dataset vs. modern era).
- Quantification of the home court advantage effect in prediction accuracy.
- Discussion of data leakage issues when using previously computed predictions as features.

2. Methods

2.1 Data Source and Description

Dataset: FiveThirtyEight NBA Elo Ratings

The data used in this analysis comes from the public domain FiveThirtyEight dataset maintained on GitHub. This dataset contains comprehensive game-by-game records for all NBA games from November 1, 1946 to the 2014-2015 NBA season.

Dataset Statistics:

- Total observations: 126,314 games.
- Observations per team per game: 2 (each game appears as two rows, one for each team).
- Actual unique games: 63,157.
- Time span: 1946-2015 (69 seasons).
- Variables: 23 per observation.

Key Variables Used:

Variable	Description
date_game	Date of the game
team_id	Team identifier (3-letter abbreviation)
pts	Points scored by team
opp_pts	Points scored by opponent
elo_i	Team's Elo rating before game
opp_elo_i	Opponent's Elo rating before game
game_location	'H' = home, 'A' = away, 'N' = neutral
forecast	FiveThirtyEight's pre-game win probability for home team

2.2 Data Acquisition

Data was obtained via the FiveThirtyEight GitHub repository:

https://raw.githubusercontent.com/fivethirtyeight/data/master/nba-elo/nbaallelo.csv

This dataset is available under public domain (CC0) licensing, allowing free use for research and educational purposes. The download was performed programmatically using Python's requests library, and the random seed used was 42. This ensures total reproducibility.

2.3 Preprocessing and Feature Engineering

Raw data was processed through the following steps:

Step 1: Home/Away Identification

Each game appears twice in the dataset (once per team's perspective). We created a binary indicator:

$$\text{is\_home\_team} = \begin{cases} 1 & \text{if game\_location} =' H' \\ 0 & \text{otherwise} \end{cases}$$

Step 2: Team Strength Feature

The primary predictor was the Elo rating difference, representing the home team's strength advantage:

$$\text{elo\_diff} = \text{elo\_home} - \text{elo\_away}$$

where elo_home is the home team's pre-game rating and elo_away is the away team's rating.

Step 3: Target Variable

The target variable (home team win/loss) was constructed by:

- Computing whether the team in each row scored more points than its opponent.
- Adjusting for perspective—if the row represents an away game, inverting the result.

$$\text{home\_win} = \begin{cases} \text{team\_scored\_more} & \text{if is\_home\_team} = 1 \\ 1 - \text{team\_scored\_more} & \text{if is\_home\_team} = 0 \end{cases}$$

2.4 Data Split Strategy

We employed chronological splitting rather than random splitting, respecting the temporal nature of the data. This mirrors real-world deployment where a model trained on historical data must generalize to future games.

Split Design:

- Training set: First 80% of games (chronologically).
- Test set: Final 20% of games (chronologically).

This prevents "leakage" from future information into training, providing an honest estimate of out-of-sample performance.

2.5 Experimental Design

We conducted two complementary experiments:

Experiment 1: Full Historical Dataset (1946-2015)

- Training data: First 101,051 games (~1946-1997).
- Test data: Final 25,263 games (~1997-2015).
- Features: Elo difference + FiveThirtyEight's pre-game forecast.
- Purpose: Assess performance across 70 years of basketball history.

Experiment 2: Modern Era Only (2010+)

- Training data: First 11,468 games from 2010-2015 (80%).
- Test data: Final 2,868 games from 2010-2015 (20%).
- Features: Elo difference only (no pre-computed forecasts).
- Purpose: Evaluate pure Elo strength in modern competition without data leakage.

2.6 Models

2.6.1 Logistic Regression

Logistic regression models the probability of a binary outcome using the logistic function:

$$P(\text{home\_win} = 1 \mid \text{elo\_diff}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{elo\_diff})}}$$

We used scikit-learn's `LogisticRegression` with default settings (L2 regularization, max_iter=200). This simple linear model provides an interpretable baseline.

2.6.2 Random Forest

Random Forest is an ensemble method that combines multiple decision trees. We trained a forest with 50 trees using scikit-learn's `RandomForestClassifier`. This non-linear model can capture complex interactions between features.

2.7 Evaluation Metrics

We report the following metrics:

- Accuracy: Percentage of correct predictions (primary metric).
- Precision: Of predicted home wins, what fraction were correct.
- Recall: Of actual home wins, what fraction were correctly identified.
- Baseline: Naive accuracy of always predicting the majority class (home win).

3. Results

3.1 Descriptive Statistics

The baseline home team win rate across the entire dataset is 62.2%, reflecting the substantial home court advantage in professional basketball. In the modern era (2010-2015), this decreases slightly to 59.3%, suggesting teams have become more evenly matched or travel/scheduling factors have shifted.

3.2 Experiment 1: Full Historical Dataset (1946-2015)

Model	**Test Accuracy**	**vs. Baseline**	**Precision (Home Win)**
Logistic Regression	67.1%	+6.6%	0.74
Random Forest	59.9%	+0.4%	0.66
Naive Baseline	60.0%	-	-

Logistic regression achieved 67.1% accuracy, outperforming the baseline by 6.6 percentage points. This improvement is substantial for a binary classification task. Random Forest performed slightly worse at 59.9%, likely due to overfitting on the large historical dataset or imbalanced classes.

3.3 Experiment 2: Modern Era Only (2010-2015) - Pure Elo

Model	**Test Accuracy**	**vs. Baseline**	**Recall (Home Win)**
Logistic Regression	65.8%	+8.3%	0.81
Random Forest	57.0%	-0.5%	0.68
Naive Baseline	57.5%	-	-

Using only Elo ratings as a predictor, logistic regression achieved 65.8% accuracy on modern games, an 8.3% improvement over the baseline. This pure Elo model achieves substantially better generalization than when combining Elo with FiveThirtyEight's pre-computed forecast, suggesting that the forecast variable introduces noise or leakage.

3.4 Classification Metrics (Modern Era)

Class	Precision	Recall	F1-Score	Support
Home Loss (0)	0.64	0.45	0.53	1,218
Home Win (1)	0.67	0.81	0.73	1,650
Weighted Avg	0.65	0.66	0.65	2,868

The model shows asymmetric performance: it correctly identifies 81% of home wins but only 45% of home losses. This reflects the class imbalance (59% home wins) and the model's bias toward the majority class. The high recall for home wins suggests the model is conservative, it identifies true home victories well but may over-predict them.

3.5 Feature Importance (Full Dataset)

When using both Elo difference and FiveThirtyEight's forecast as features in the Random Forest, feature importance (mean decrease in impurity) reveals:

- Elo difference: 54.8%
- FiveThirtyEight forecast: 45.2%

This demonstrates that team strength ratings drive predictions, with the pre-computed forecast contributing but not dominating the decision process.

4. Discussion

4.1 Interpretation of Results

Our results demonstrate that Elo ratings, despite their simplicity, provide strong predictive power for NBA outcomes. A 65-67% accuracy on binary classification substantially exceeds the 50% random baseline and the 57-60% baseline of simply predicting the majority class. This finding aligns with prior work in sports analytics showing that Elo-based ratings outperform more complex models in many competitive domains.

4.2 Home Court Advantage

The baseline accuracy (57-62%) directly reflects home court advantage. Over 70 years, home teams win 62% of games, a 12% gap from 50-50. This confirms factors attributing advantage to crowd effects, travel fatigue, and officiating patterns.

4.3 Temporal Generalization

A notable finding is the performance difference between the full historical split and the modern era:

- Full dataset (1946-2015): 67% LR accuracy
- Modern era only (2010-2015): 66% LR accuracy (pure Elo)

The slight decrease suggests that the relationship between Elo and outcomes may weaken in modern professional sports, possibly due to increased player movement, salary cap mechanics, or the rise of advanced analytics in team management. However, the difference is very small (1%), indicating that Elo remains a good option.

4.4 Data Leakage Issue

An important methodological observation: including FiveThirtyEight's pre-computed forecast (Experiment 1) as a feature creates data leakage. The forecast is itself a probability derived from Elo ratings, so using it alongside Elo introduces redundant information.

Experiment 2 (modern era, pure Elo) avoids this issue and achieves 66% accuracy, nearly identical to the full dataset result. This demonstrates that Elo alone is sufficient and that additional engineered features did not improve generalization.

4.5 Model Comparison

Logistic regression consistently outperformed Random Forest:

Model	Full Data	Modern
Logistic	67.1%	65.8%
Random Forest	59.9%	57.0%

This is somewhat counterintuitive, as Random Forests typically excel with non-linear relationships. Possible explanations:

- Class imbalance: The 60-40 win/loss ratio may cause Random Forest to overfit to the majority class.
- Insufficient features: With only 1-2 features, ensemble methods have limited diversity to exploit.
- Tree depth: Default tree depth (unlimited) may cause overfitting on the large training set.

4.6 Limitations

Several limitations should be noted:

Data Limitation: The dataset ends in 2015. Current NBA (2016-2026) includes new factors (player load management, COVID disruptions, rule changes) not reflected in historical patterns.

Feature Limitation: We used only Elo ratings. A richer model might include:

- Individual player statistics (injuries, form).
- Recent team performance (momentum effects).
- Rest days and back-to-back schedules.
- Coaching changes.
- Home venue characteristics.

Class Imbalance: Home teams win 59-62% of games, creating imbalanced classes that bias models toward predicting home wins.

Temporal Non-stationarity: NBA team strengths, rules, and dynamics shift over decades. A model trained on 1946-1980 games may not transfer well to 2010-2015 games. Our chronological split respects this but doesn't fully address it.

Causality: Elo ratings correlate with wins but don't prove causation. Better teams earn higher ratings precisely because they win more.

4.7 Practical Implications

From a sports analytics perspective:

- **Betting/Analytics Markets**: A 65% predictor beats the random baseline but likely underperforms professional sports bettors, who incorporate real-time information (injuries, lineups) and advanced models.
- **Team Management**: Elo-style ratings could inform strategic decisions, though full team strength assessment requires additional data.
- **Educational Value**: This work demonstrates how simple statistical methods can capture meaningful patterns in complex domains.

4.8 Future Work

Potential extensions of this research:

- **Incorporate modern data**: Update dataset to 2025-2026 season, evaluate whether the model generalizes or not.
- **Add features**: Integrate player-level statistics, team form (recent win/loss streaks) and injury reports.
- **Temporal models**: Use any time-series models to capture momentum effects.
- **Ensemble methods**: Combine Elo with player stats, team rest and schedule difficulty.
- **Real-time prediction**: Deploy model as web app for live game outcome forecasting.