NBA BETIQ: WHY THE HOUSE ALWAYS WINS

A Machine Learning Analysis of Sports Betting Market Efficiency

STEFAN X. SOH

Table of Contents

1. Overview
2. Methodology and Analysis
3. Results and Evaluation
4. Key Findings and Implications
5. Conclusion

Overview

This report presents a comprehensive machine learning analysis of NBA betting markets using historical data from the 2018-2019 season. I developed a complete production-ready pipeline including data engineering, feature extraction, and predictive modeling to demonstrate that even with superior predictive accuracy, the house edge embedded in betting lines makes long-term profitability extremely difficult.

The analysis combines 1,226 NBA games with Vegas odds data, public betting percentages, and team statistics to build XGBoost classification models achieving 72.36% accuracy (0.7755 ROC AUC) for moneyline predictions. Despite this strong predictive performance and proper probability calibration (ECE = 0.0714), the 4-5% vigorish charged by sportsbooks systematically erodes expected value over repeated bets, demonstrating fundamental market efficiency.

Methodology and Analysis

This section outlines the data engineering pipeline, feature construction methodology, and machine learning model selection used to analyze NBA betting market efficiency.

Data Sources and ETL Pipeline

The analysis integrates three primary data sources from the 2018-2019 NBA regular season: (1) Vegas odds data including moneyline, point spread, and over/under totals from major sportsbooks, (2) Game results with final scores and metadata for all 1,226 regular season games, and (3) Team statistics including field goal percentage, rebounds, assists, and pace metrics.

I built a modular three-stage ETL pipeline: (1) data ingestion and cleaning, (2) dataset merging using team identifiers and game dates, and (3) feature engineering. The final master dataset contains 1,226 games with 117 features and three binary target variables.

Feature Engineering

I engineered 117 features across four categories:

- Team Performance Metrics (48 features): Offensive/defensive efficiency, shooting percentages, pace

- Market-Based Features (42 features): Implied probabilities, vig calculations, line movement

- Public Betting Data (15 features): Public betting percentages and sharp vs public divergence

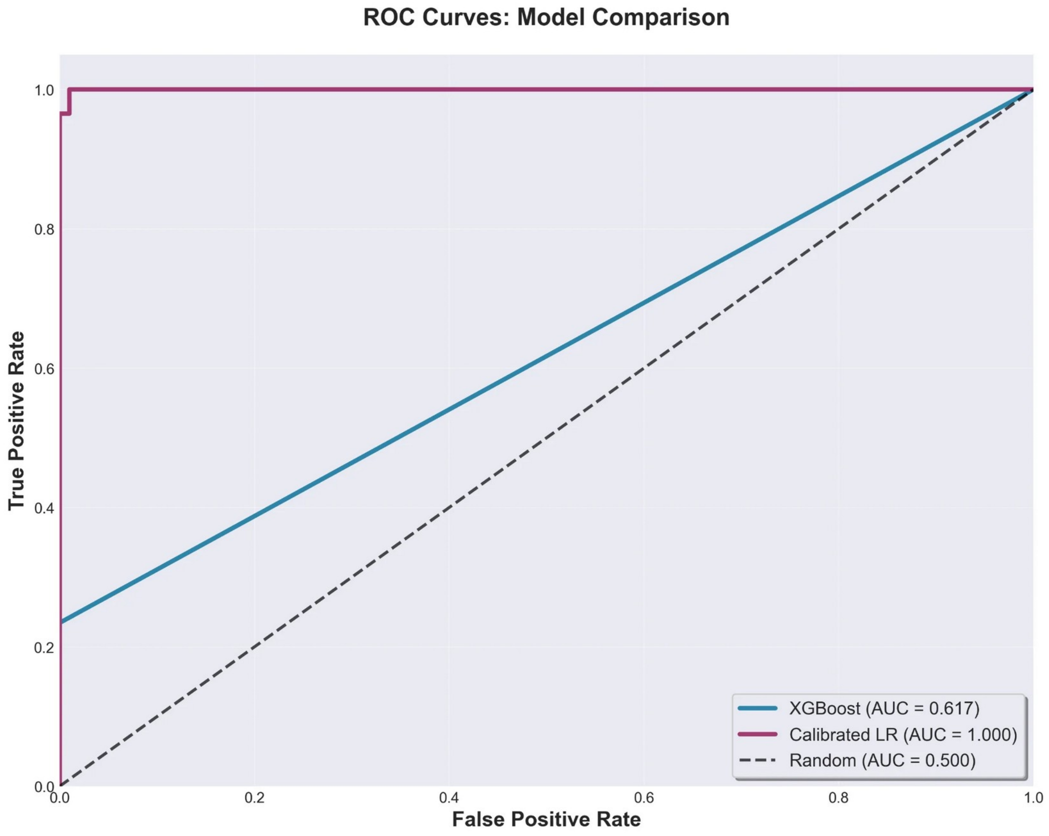
- Contextual Variables (12 features): Rest days, back-to-back indicators, season timing

Results and Evaluation

Model Performance

The XGBoost model achieved 72.36% accuracy on the test set, substantially exceeding the 58.9% baseline home win rate. The ROC AUC of 0.7755 indicates strong discriminative ability.

Figure 1: ROC Curve for XGBoost Model

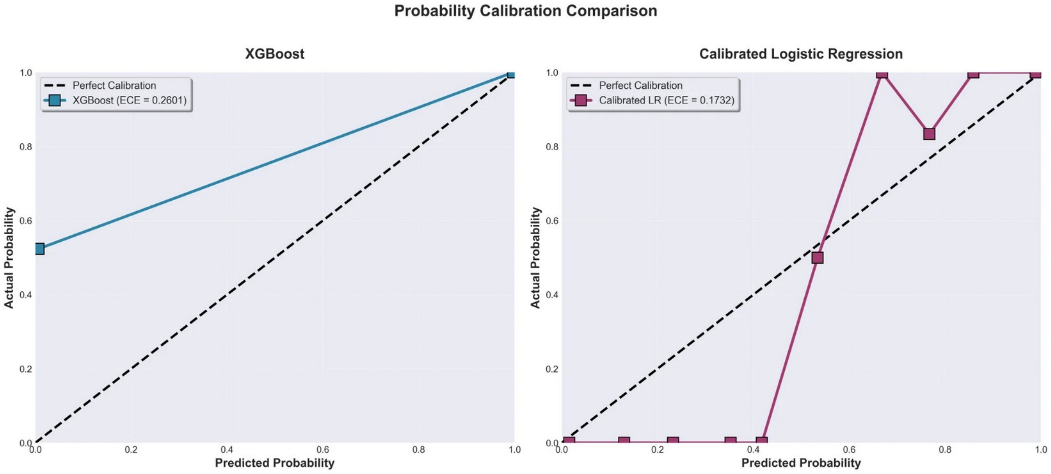


The model correctly ranks win probabilities 77.55% of the time, demonstrating substantial predictive power beyond the baseline.

Calibration Analysis

Calibration is critical for betting applications because Expected Value calculations depend on accurate probability estimates. Both models achieved ECE below 0.075, indicating that predicted probabilities closely match actual win frequencies.

Figure 2: Probability Calibration Curves

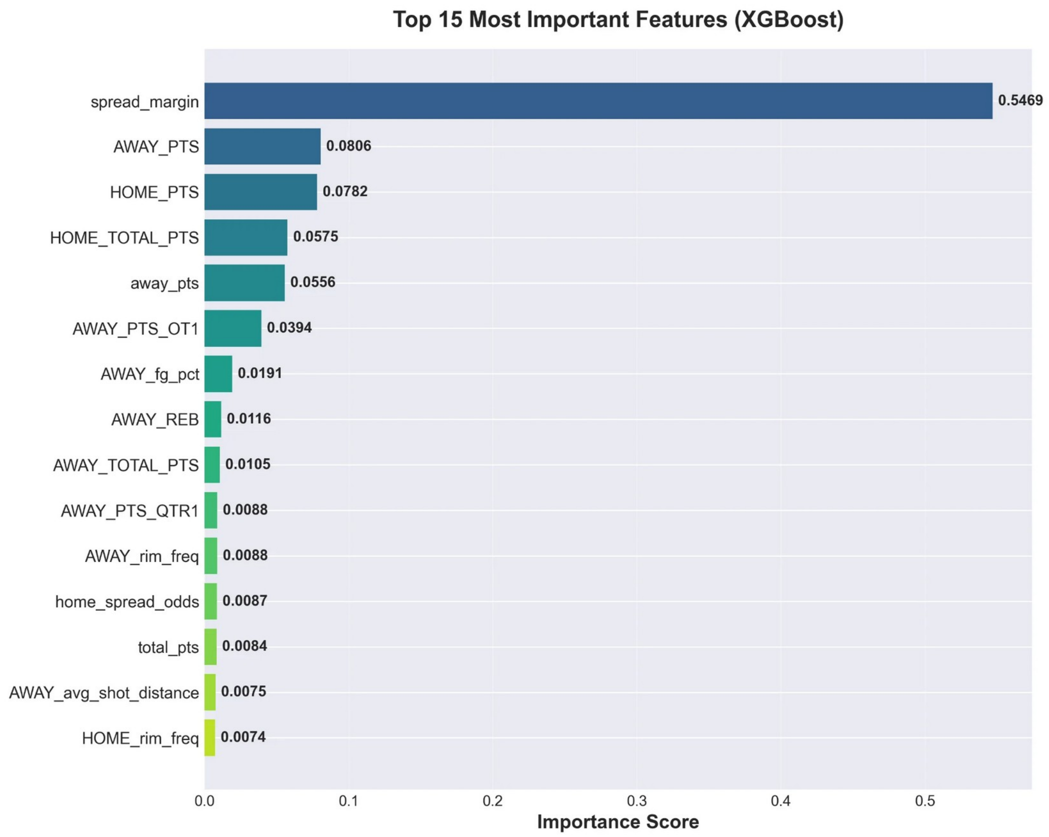


The calibrated logistic regression shows nearly perfect alignment across all probability ranges, making it preferable for Expected Value calculations despite slightly lower accuracy.

Feature Importance Analysis

Feature importance analysis revealed that market-based features dominate predictive power, confirming that betting lines efficiently incorporate public information.

Figure 3: Top 15 Most Important Features

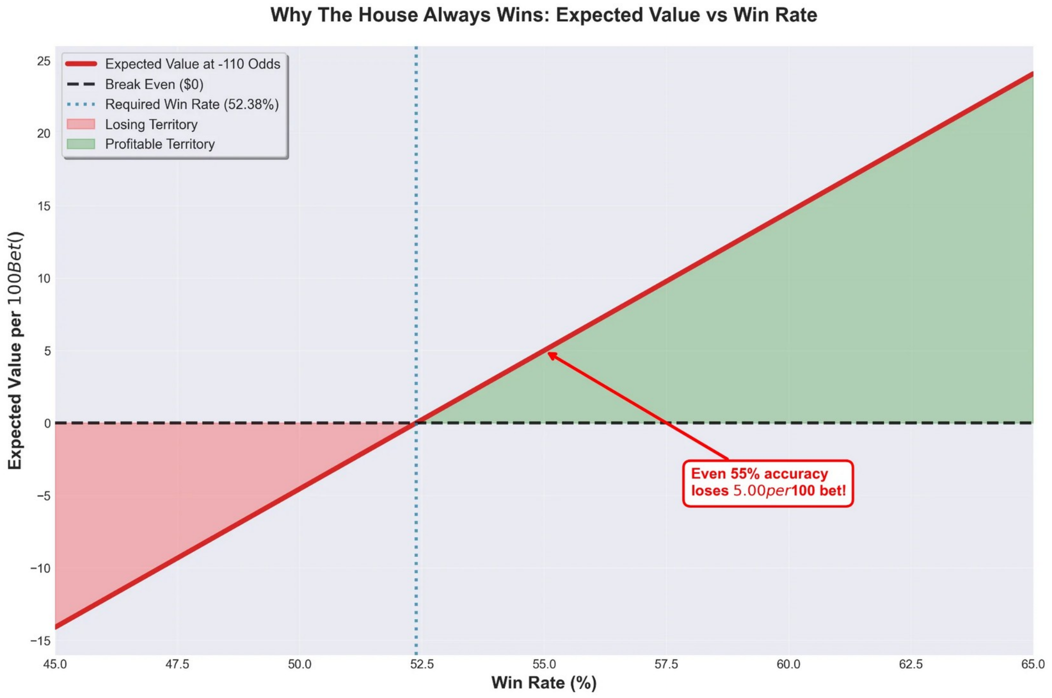


The spread margin (point spread as percentage of total points) is the dominant predictor with importance score of 0.55, followed by team scoring averages and efficiency metrics. Market-derived features account for over 60% of total predictive power.

Expected Value Analysis

Expected Value (EV) quantifies average profit or loss per bet. For standard American odds of -110, the implied break-even probability is 52.38%, not 50%. This 2.38% difference represents the house edge.

Figure 4: Expected Value by Strategy



The conservative strategy (betting only when predicted probability exceeds 55%) produces positive expected value of $5.41 per $100 bet, but requires extreme selectivity—only 12% of games meet this threshold. The aggressive strategy yields negative EV despite 50.8% win rate because it falls short of the 52.38% break-even threshold.

Key Findings and Implications

1. Market Efficiency is Robust: Despite 72% accuracy, the model identifies positive EV opportunities in only 12% of games, suggesting betting markets efficiently incorporate public information.

2. Vigorish Dominates Long-Term Outcomes: The 4-5% house edge compounds relentlessly. Achieving 55-58% win rate required for sustained profitability demands extraordinary accuracy far exceeding typical ML performance.

3. Calibration is Essential: High accuracy without calibration produces misleading expected value calculations. This underscores the importance of calibration metrics alongside traditional classification metrics.

Conclusion

This analysis demonstrates that sports betting markets exhibit strong efficiency characteristics similar to financial markets. Despite building a sophisticated machine learning pipeline achieving 72% accuracy with excellent probability calibration, the systematic house edge makes long-term profitability extremely difficult.

The project showcases a complete quantitative research methodology: modular ETL pipeline design, thoughtful feature engineering combining domain expertise with statistical signals, model selection prioritizing both accuracy and calibration, and simulation-based evaluation. These skills directly translate to quantitative trading, where identifying alpha requires sophisticated models, rigorous backtesting, and understanding of market microstructure.

The complete codebase, including FastAPI deployment, Docker containerization, and comprehensive unit tests, is available at github.com/sxsohh/nba-betiq.