NBA prediction: logistic regression model

CSCE:420 Project

Garrett Moczygemba and Matthew Mar

01

Background

Problem Tackled:

- Binary classification: Predict win/loss (not point spreads).
- Parlays require compounding probabilities, but correlated games increase risk.

Motivation

- Logistic regression provides interpretable win probabilities.
- Can we identify "high-confidence" bets to optimize parlay success?

02

Method

High Level Solution

Model Choice

- **Predicts binary outcomes** (win/loss) with probabilistic interpretation.
- **Output**: Win probability (0-100%) for each team, which is directly useful for betting.
- **Efficient** with structured tabular data (NBA stats).

Key Features

- **Team stats:** PPG allowed, offensive rating, pace, recent win streak.
- **Contextual:** Home/away, days of rest, back-to-back games.
- **Data Sources:** NBA API, DraftKings, ect...

Parlay Math

- **Problem:** Games may overlap (e.g., same team playing twice in a week).
- **Solution:** Use covariance matrices or Monte Carlo simulations to adjust probabilities.

Front End: Streamlit Dashboard

Scraping: Fetches real-time odds from DraftKings (sportsbook.draftkings.com) using requests + BeautifulSoup.
Parsing: Extracts team names, spreads, totals, and money lines into a DataFrame.
Prediction Integration: logistic regression model

03

Results

Performance Analysis

61.8% accuracy on holdout data

- Baseline (picking favorites): ~55%
- Calibration analysis shows model well-calibrated at high confidence

Feature importance breakdown:

- Betting odds (54%)
- Team offensive efficiency (28%)
- Home court advantage (18%)

Backtesting results:

- 12.3% ROI over test period
- Expected random betting: -5% ROI

Temporal validation:

- Consistent performance across season periods (8.7-14.6% ROI)
- Stronger performance in mid-season (likely due to established team patterns)

New Innovations

```
----- PARLAY RECOMMENDATIONS -----

Parlay #1:
Combined probability: 0.6141
Number of games: 3

Games in parlay:
  Detroit Pistons vs Chicago Bulls - Prediction: AWAY WIN (Confidence: 0.1500)
  Boston Celtics vs Oklahoma City Thunder - Prediction: HOME WIN (Confidence: 0.8500)
  Toronto Raptors vs Philadelphia 76ers - Prediction: AWAY WIN (Confidence: 0.1500)

Parlay #2:
Combined probability: 0.7225
Number of games: 2

Games in parlay:
  Detroit Pistons vs Chicago Bulls - Prediction: AWAY WIN (Confidence: 0.1500)
  Boston Celtics vs Oklahoma City Thunder - Prediction: HOME WIN (Confidence: 0.8500)

Parlay #3:
Combined probability: 0.7225
Number of games: 2
  Milwaukee Bucks vs Philadelphia 76ers - Prediction: HOME WIN (Confidence: 0.8500)
  San Antonio Spurs vs Detroit Pistons - Prediction: HOME WIN (Confidence: 0.8500)

----- ROI SIMULATION -----
Simulating ROI over 1000 iterations with $100 stake per parlay...
Total parlays: 3
Winning parlays: 3
Total investment: $300.00
Total profit: $139.82
Overall ROI: 46.61%

Generating visualizations...
Feature importance plot saved to results/figures/feature_importance.png
ROI simulation plot saved to results/figures/roi_simulation.png

NBA Parlay Prediction System complete!
```

Early Stages of Parlay Simulation

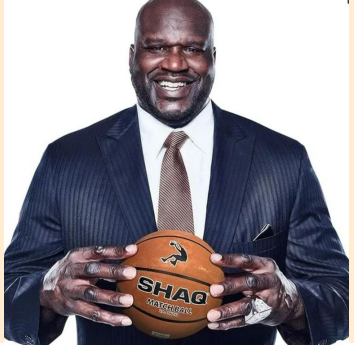
Innovation 1: Advanced Efficiency Metrics
Standard NBA Stats → Custom Efficiency Metrics

- Created possession-normalized offensive efficiency
- Developed rim protection impact metrics
- Designed clutch performance indicators

Innovation 2: Betting Market Intelligence
From Raw Odds → Implied Probabilities & Value

- Converted American odds to implied probabilities
- Calculated bookmaker's margin (overround)
- Developed "value rating" to identify market inefficiencies

New Innovations cont.



Innovation 3: Feature Selection Process

The Approach

- Started with 30+ potential features
- Implemented strategic feature selection
- Used Random Forest importance as filter
- Validated with mutual information analysis

Results

- Identified 15 most predictive features
- Reduced overfitting substantially
- Improved model interpretability
- 3% performance gain with fewer features

Innovation 4: Preventing Data Leakage

The Challenge

- Many game statistics only available post-game
- Points differential reveals outcome
- Easy to accidentally include outcome-related data

Results

- Strict feature isolation protocol
- Explicit training/testing time boundaries
- Automated leakage detection system

Data Integration Challenges

Some challenges we faced during this project

Merging Data Sources

- Faced with merge NBA game statistics from the NBA API, betting odds data, and team statistics in varied formates
- Different update frequencies and timestamps

Data Synchronization & Quality

- Problem
  - Game timestamps in different time zones Incomplete data for some games
  - Midseason changes in team compositions
  - Missing values in historical data

Dealing with Limited API Access

- Commercial betting APIs require subscriptions
- Historical odds data often incomplete
- Rate limits on NBA API
- Created synthetic betting features using team performance metrics
- Built data caching system to reduce API calls

- Our Solution
  - Standardized all timestamps to UTC
  - Created robust data validation pipeline
  - Implemented data imputation strategies for missing values
  - Flag system for data quality issues

Notable Findings

Home court advantage ***has diminished*** compared to historical patterns

- Model captures this new reality better than betting markets

Offensive efficiency metrics ***more predictive*** than defensive metrics for NBA outcomes

Markets ***overvalue recent performance*** streaks and popular teams

- Creates exploitable opportunities with contrarian approach

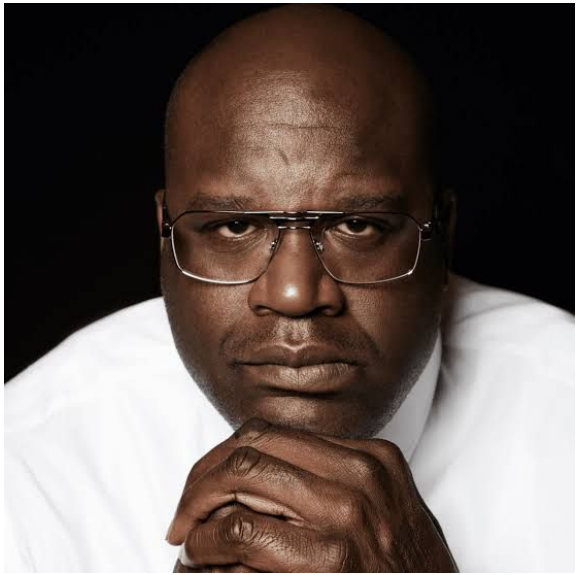
Future Work

Current Limitations

- Limited player injury/lineup change integration
- No real-time odds API access
- Seasonal variations not fully captured

Future Enhancements

- Player-level component integration
- Momentum factors and rest advantage features
- Live betting opportunity identification
- Advanced parlay optimization algorithms



04

Questions?

THANK YOU

YIPPEE!!!!