NBA Game Prediction Analysis: Predicting Pacers Playoff Outcomes (2024–2025 Season)

Github link:https://github.com/vacillating/NBA_Playoff_Prediction_2025

1. Introduction

Inspired by attending Game 6 of the NBA Eastern Conference Finals featuring the Indiana Pacers, I chose to focus my AAE 718 Project 4 on predicting NBA playoff game outcomes. The Pacers stood out for their dynamic gameplay characterized by remarkable comebacks, clutch moments, and strategic coaching decisions. This report details the development and evaluation of machine learning models intended to predict the outcomes of the Pacers' playoff games during the 2024–2025 NBA season.

2. Data Sources

All data utilized in this analysis was obtained via the `nba_api` Python library, enabling dynamic fetching and processing of real-time NBA statistics. Data encompassed regular-season and playoff games, specifically focusing on game outcomes (win/loss), points scored (`PTS`), plus-minus (`PLUS_MINUS`), and matchup details (home/away).

3. Methods

3.1 Data Collection and Processing

Data was systematically fetched through the NBA API, capturing games from both the regular season and playoffs. Each game record included:

- Game outcome (`WIN`)
- Home or away indicator (`IS_HOME`)
- Point differential from previous games (`POINT_DIFF`)
- Rolling averages over five games: points (`ROLLING_PTS`), plus-minus (`ROLLING_PLUS_MINUS`), and win rate (`ROLLING_WIN_RATE`).

3.2 Modeling Approaches

Initially, predictions were conducted using a single-model Random Forest classifier. However, due to limited data and initially misleading accuracy (~70%, similar to naively predicting all games as wins), further improvements were necessary. To enhance the Random Forest model:

- Additional features were incorporated, such as rolling averages and game momentum metrics.
- Parameter tuning (e.g., adjusting number of trees and maximum depth) was conducted to reduce overfitting and improve predictive robustness.

Furthermore, a dual-model strategy was adopted to improve interpretability and compare predictive performance:

- **Random Forest Classifier:** Enhanced feature engineering and hyperparameter optimization for capturing complex nonlinear relationships.
- **Logistic Regression Classifier:** Introduced to provide interpretability, baseline comparison, and robust performance with limited data.

Training was performed exclusively on regular-season data, with playoff games reserved for model evaluation.

4. Results

4.1 Model Accuracy Comparison

Model	Training Accuracy	Testing Accuracy
Random Forest	0.95	0.80
Logistic Regression	0.87	0.80

Both models achieved **80% accuracy** on playoff predictions, clearly outperforming the naive baseline prediction approach (~70% accuracy).

4.2 Playoff Predictions

Both models predicted the Pacers would lose Game 5 (June 13, 2025, at OKC):

Game	Random Forest Prediction	Logistic Regression Prediction
G5	Loss	Loss

5. Discussion

5.1 Implications of Findings

The modest accuracy (80%) underscores the complexity of predicting NBA playoff outcomes, particularly for a team known for unexpected results. Factors such as momentum, individual player performance, and strategic adjustments are difficult to quantify and model precisely.

5.2 Limitations

The primary limitations of this analysis include:

- **Data scarcity:** Only one season was analyzed, limiting generalizability.

- **Feature limitations:** The absence of player-level data (injuries, individual statistics) and strategic data (e.g., coaching adjustments).

- **Sample size:** Limited playoff games significantly restrict model training robustness.

5.3 Future Improvements

To improve predictive accuracy, future analyses should:

- Integrate detailed player-level statistics (injuries, performance consistency).

- Incorporate tactical and strategic variables (coaching style, player rotations).

- Expand to multiple seasons for greater model stability.

- Experiment with advanced models (e.g., neural networks, gradient boosting methods) given larger datasets.

6. Conclusion

This project demonstrated the feasibility and challenges of predicting NBA playoff outcomes using machine learning techniques. The dual-model approach provided a balanced exploration between complexity and interpretability, reinforcing key learnings from the AAE 718 Data Science course. Future enhancements will require richer datasets and more sophisticated feature engineering to capture the nuanced dynamics of professional basketball.