The Indiana Pacers 2025 Play-off Run: Discovering Trends to Predict Future Game Outcomes

Fritz Kussow, Matt Damsma, Timothy Gilmore, and Alonso Carrasco

1 - Abstract

This project aims to organize statistics and features from the 2024-2025 Indiana Pacers' season and find what similarities exist between wins and losses for games played in that time frame. We'll want to pull

basic statistics from each game, including all 82 regular season games and all 23 playoff games. The statistics we'll look at include who played in each game, how long they played in each game, and each player's individual stats (points, rebounds, assists, turnovers, etc.) for each game they played in. Using this data, we likely can organize players into certain 'types.' For instance, if there's a game where a player shoots 80% from three, we could say that the pacers had a sharpshooter for that game, which would likely improve their chance of winning. Additionally, we'll take into account the makeup of whatever team the Pacers played against. For instance, it's possible that the Pacers had a much worse win percentage against teams with a sharpshooting big that could stretch the floor. Hopefully, by focusing on the five highest-minute players and organizing them into certain player archetypes, we can create a decently accurate model to predict the outcomes of pacers games based on the pregame starting five.

2 - Introduction

Professional basketball has entered an era where data driven decision making shapes everything from game strategies to team roster construction. The NBA's statistical tracking system captures hundreds of metrics per game. In return, this creates many datasets that offer great opportunities for analytical exploration. Predicting the outcome of NBA games is not an easy task. This is because every NBA game has great variability where a single shot or defensive stop could change the result of the game. For the Indiana Pacers, the leveraging of the vast amount of data collected by the NBA can prove to be very beneficial, providing a competitive advantage. The core problem this project aims to address is: Can we predict Indiana Pacers game outcomes more accurately by finding patterns in similar historical games?

3 - Methodology

3.1 - Data Acquisition and Preprocessing

The dataset for this project will be constructed from publicly available NBA game data, sourced from sports statistics APIs (e.g., nba_api) or comprehensive basketball data websites (e.g., Basketball-Reference.com). The dataset will encompass all 105 games (82 regular season, 23 playoff) played by the Indiana Pacers during the 2024-2025 season.
For each of these 105 games, we will collect:

- Game-level data: Date, opponent, location (home/away), and final score.
- Player-level data: Detailed box scores for every player who participated in each game (for both the Pacers and their opponent). This includes minutes played (MP) and standard counting stats (points, rebounds, assists, steals, blocks, turnovers, fouls, FGA, FGM, 3PA, 3PM, FTA, FTM).

The raw data will be preprocessed into a structured format where each row represents a single game (N=105). The primary target variable will be a binary outcome: Pacers_Win (1 for a win, 0 for a loss). All other collected data will be used to engineer predictive features. Preprocessing will also involve handling missing values, such as players who were on the roster but did not play (DNP).

3.2 - Feature Engineering: Player Archetypes

A central component of our methodology is to move beyond individual player statistics and model the composition of the lineups. As proposed in the abstract, we will operationalize this by creating "player archetypes."

- Archetype Clustering: We will apply a K-Means clustering algorithm to define these archetypes. The clustering will be performed on a normalized dataset of all players (from both the Pacers and their opponents) who played significant minutes during the season. The features used for clustering will include per-36-minute statistics and advanced metrics (e.g., True Shooting Percentage (TS%), Usage Rate (USG%), Assist Rate, Rebound Rate) to capture a player's style and on-court role.
- Archetype Labeling: After clustering, we will manually inspect the statistical profile of each cluster to assign a descriptive label.
- Game-level Feature Vector: For each of the 105 games, we will create a feature vector representing the lineup composition. This vector will be a count of how many players belonging to each archetype were among the top five minute-getters for that game. This will be done for both the Pacers and their opponent.

The final feature set for each game will include:

- Pacers_Archetype_1_Count
- Pacers_Archetype_2_Count
- ...
- Opponent_Archetype_1_Count
- Opponent_Archetype_2_Count
- ...
- Game_Location (0 for Away, 1 for Home)

This approach directly models the interaction of lineup archetypes, addressing the hypothesis that the type of opponent impacts the Pacers' win probability.

3.3 - Model Development and Evaluation

Given the binary nature of our target variable (Pacers_Win), this project will formulate the problem as a binary classification task. Due to the small dataset size (N=105), we will prioritize models that are less prone to overfitting and offer high interpretability.

- Selected Models: We will implement and compare several classification algorithms:
    - **Logistic Regression**: A baseline model that is highly interpretable, allowing us to quantify how the presence of specific Pacers or opponent archetypes directly impacts the log-odds of winning.
    - **K-Nearest Neighbors (K-NN)**: This model works by finding the 'k' most similar games from the past (the "nearest neighbors") and then predicting the outcome based on how those similar games turned out. It's a good, straightforward approach for a "pattern matching" problem like this.
    - **Naive Bayes**: The Naive Bayes probabilistic classifier calculates the likelihood of a win or loss given the lineup archetypes. Since the dataset is smaller, this classifier could perform quite well.

- Model Validation: A simple train-test split would be unreliable with only 105 data points. Therefore, we will employ 10-fold cross-validation to generate a stronger estimate of each model's performance. The data will be shuffled and split into 10 subsets; the model will be trained 10 times, each time using 9 of the subsets for training and 1 for testing.
- Evaluation Metrics: Model performance will be assessed using a suite of standard classification metrics:
    - Accuracy: The overall percentage of correctly predicted game outcomes.
    - Precision, Recall, and F1-Score: These metrics are especially crucial if the win-loss record is imbalanced (e.g., if the Pacers had a 70-win season, a model that always predicts "Win" would have high accuracy but zero utility).

The deliverable will be a model that is highly accurate and provides actionable insights into which lineup compositions and opponent archetypes led to wins and losses for the 2024-2025 Pacers.